# SCHOLARLY PUBLICATIONS

*A CURRENT AWARENESS BULLETIN*
*OF RESEARCH OUTPUT*

## *@DTU*

(117th Edition)

***SEPTEMBER 2022***

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

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

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

This issue contains the information published during **September, 2022**. 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 Fully-Digital BIST Wrapper Based on Ternary Test Stimuli for the Dynamic Test of a 40 nm CMOS 18-bit Stereo Audio ΣΔ ADC, Manuel J. Barragan, Rshdee Alhakim, Haralampos-G. Stratigopoulos, Matthieu Dubois, Salvador Mir, Hervé Le-Gall, Neha Bhargava and ***8.1.Ankur Bal***, Electrical, DTU

2. A Hybrid Finite Difference Method for Singularly Perturbed Delay Partial Differential Equations with Discontinuous Coeffifficient and Source, ***6.Nitika Sharma*** and ***3.Aditya Kaushik***, Mathematics, DTU

3. A Multivariate Analysis of Agricultural Electronic Trading Adoption, ***6.Sanjay Chaudhary*** and ***3.P K Suri***, DSM, DTU

4. A Novel Graphical Password Scheme to Avoid Shoulder-Surfing Attacks in Android Devices, ***8.Mayank Kaushik***, ***8.Ashish Rawat***, ***3.Varsha Sisaudia*** and ***8.Love Parashar***, IT, DTU

5. A novel type-II intuitionistic fuzzy clustering algorithm for mammograms segmentation, ***3.Sonika Dahiya*** and Anjana Gosain, CSE, DTU

6. A Passive Grounded MOSCAP- Memrisor Emulator, ***6.Rahul Kumar Gupta, 3.Prof. M. S. Chaudhary, 3. Dr. Sachin Taran*** and Dr. Varun Saxena, Electronics, DTU

7. A self-attention hybrid emoji prediction model for code-mixed language: (Hinglish), ***8.Gadde Satya Sai Naga Himabindu***, ***8.Rajat Rao*** & ***3.Divyashikha Sethia***, CSE, DTU

8. A Single-Step Subranging Relaxation Oscillator-Based Open-Loop Sigma-Delta ADC, Sudhanva Vasishta , ***8.K. R. Raghunandan***, Ananth Dodabalapur and T. R. Viswanathan, Electronics, DTU

9. Achieving Low-Noise, High-Bandwidth, Nanoscale-Resolution Position Sensing with GMR Sensors, Tim John Joseph and ***8.1.V. Kartik***, Mechanical, DTU

10. ADAPTATION OF DOMAIN-SPECIFIC TRANSFORMER MODELS WITH TEXT OVERSAMPLING FOR SENTIMENT ANALYSIS OF SOCIAL MEDIA POSTS ON COVID-19 VACCINES, ***8.Anmol Bansal***, ***8.Arjun Choudhry***, ***8.Anubhav Sharma*** and ***3.Seba Susan***, IT, DTU

11. Adsorption of cytosine on prebiotic siliceous clay surface induced with metal dications: Relevance to origin of life, Preeti Bhatt, Chandra Kala Pant, Pramod Pandey, Yogesh Pandey, Satish Chandra Sati & ***3.Mohan Singh Mehata***, Physics, DTU

12. An electrochemical immunosensor for detection of H. pylori secretory protein VacA on g-C3N4/ZnO nanocomposite modified Au electrode, Kirti Saxena, Arun Kumar, Nidhi Chauhan, Manika Khanuja, ***3.Bansi D. Malhotra*** and Utkarsh Jain, Biotechnology, DTU

13. An Extensive Data-Based Assessment of Optimization Techniques for Distributed Generation Allocation: Conventional to Modern, ***6.Vivek Saxena***, ***3.Narendra Kumar*** and ***3.Uma Nangia***, Electrical, DTU

14. An Iterative Method for Image Denoising Based on Ramanujan Sums, ***8.Priyansh Jaiswal***, ***8.Noman Ahmed Ansari***, ***3.Anup Kumar Mandpura*** and ***3.Neeta Pandey***, Electrical and Electronics, DTU

15. An overview on synthesis, properties and applications of polycaprolactone copolymers, blends & composites, ***6.1.Radha Sachan***, ***3.Sudhir G. Warkar*** and ***3.Roli Purwar***, Applied Chemistry, DTU

16. Analysis of land use and land cover changes and their impact on temperature using landsat satellite imageries, ***6.Rubeena Vohra*** and ***3.K. C. Tiwari***, Civil, DTU

17. Analysis of various Pre-processing filtering methods for MRI brain tumour detection in a noisy environment, ***3.Priyanka Jain*** and ***7.Priyasha Parmar***, ECE, DTU

18. Chaotic Mapping Based Advanced Aquila Optimizer With Single Stage Evolutionary Algorithm, ***3.MONIKA VERMA***, ***3.MINI SREEJETH***, ***3.MADHUSUDAN SINGH***, THANIKANTI SUDHAKAR BABU AND HASSAN HAES ALHELOU, Electrical, DTU

**19.** Circuit Complexity in an interacting quenched Quantum Field Theory, Sayantan Choudhury, Rakshit Mandish Gharat, Saptarshi Mandal and ***8.Nilesh Pandey***, Applied Physics, DTU

**20.** Comparative Analysis of Particle Swarm Optimization and Artificial Neural Network Based MPPT with Variable Irradiance and Load, ***8.Spandan Srivastava***, ***8.1.Charu Lata***, ***8.Prateek Lohan*** and ***3.Rinchin W. Mosobi***, Electrical, DTU

**21.** Comparative Study of NCM and NCA Electrode Material for Capacity-Fade Using 1-D Modeling, ***8.Ebani Gupta***, ***8.Vikas Verma***, ***8.Gaurav Khulway*** and ***3.Amrish K. Panwar***, Applied Physics, DTU

**22.** Comparison of Different Coil Structures for Wireless Power Transfer Using Magnetic Resonant Coupling, ***8.Pranjal Bajpai***, ***8.Pranav Saihgal***, ***8.Pulkit Kapoor*** and ***3.Saurabh Mishra***, Electrical, DTU

**23.** Comparison of Neural Networks and XGBoost Algorithm for Music Genre Classification, ***8.Ritik Gusain***, ***8.Sagar Sonker***, ***7.Sachin Kumar Rai***, ***3.Ankita Arora*** and ***3.S.T. Nagarajan***, Electrical, DTU

**24.** Constructing Sentiment Sentence Embedding Model Using Transfer Learning, ***3.Rajesh Kumar Yadav***, ***8.Shivam Singhal***, ***8.Shashank Chugh*** and ***8.Shivam Jaiswal***, CSE, DTU

**25.** Contrast Aware Image Dehazing using Generative Adversarial Network, ***3.Anil Singh Parihar***, ***3.Kavinder Singh***, ***8.Aryan Ganotra***, ***8.Aviral Yadav*** and ***8.Devashish***, CSE, DTU

**26.** DEEP CONVOLUTIONAL ARCHITECTURES FOR EXTRAPOLATIVE FORECASTS IN TIME-DEPENDENT FLOW PROBLEMS, ***8.Pratyush Bhatt***, ***8.Yash Kumar*** and Azzeddine Soulaïmani, Mechanical, DTU

**27.** Design and Development of a Solar Electric Delivery Pod, ***8.Abhishek Arora*** and ***3.Janardan Kesari***, Mechanical, DTU

**28.** Detection and Analysis of Flooding Attacks in Wireless Sensor Networks, ***6.Gebrekiros Gebreyesus Gebremariam***, ***3.Jeepanda Panda*** and ***3.S. Indu***, ECE, DTU

**29.** ECG Based Early Heart Attack Prediction Using Neural Networks, ***8.Ashish Kumar***, Ketan Rathor, Snehit Vaddi, Devanshi Patel, Preethi Vanjarapu and Manichandra Maddi, IT, DTU

**30.** ECG Heartbeat Classification of Myocardial Infarction and Arrhythmia using CNN, ***8.Shrey Agarwal***, ***8.Sarthak Khattar***, ***8.Saiyam Soni*** and ***3.Dr. Anukul Pandey***, Electronics, DTU

**31.** Ensemble Deep Convolution Neural Network for Sars-Cov-V2 Detection, ***8.Subrat Sarangi***, ***8.Uddeshya Khanna*** and ***3.Rohit Kumar***, Applied Mathematics, DTU

**32.** Ensemble of Hybrid CNN Models for Brain Tumor Classification, ***3.Indu Singh***, ***8.Rahul Kumar Yadav***, ***8.Rahul Virahaniya*** and ***8.Noor Alam***, CSE, DTU

**33.** Epicatechin analogues may hinder human parainfluenza virus infection by inhibition of hemagglutinin neuraminidase protein and prevention of cellular entry, ***8.1.Sidharth Bhasin***, ***8.Megh Nadar*** and ***3.Yasha Hasija***, Biotechnology, DTU

**34.** Estimation of Air Flow Around the Tall Buildings of Different Shapes, ***6.Rahul Kumar Meena***, ***3.Ritu Raj*** and ***3.S Anbukumar***, Civil, DTU

**35.** Facial Micro-Expression Recognition (FMER) using Model Compression, ***6.Nikhil Singh*** and ***3.Rajiv Kapoor***, Electronics, DTU

**36.** Fake News Detection in Different Applications: State-of-the-art and Future Directions, ***6.Aman Jolly*** and ***3.Shailender Kumar***, CSE, DTU

**37.** Fifty Years of Research in Deposit Insurance: A Bibliometric Analysis and Review, Varda Sardana and ***6.Shubham Singhania***, USME, DTU

**38.** First-Principles Study of Enhanced Absorption in Van der Waals Heterostructure of MoS2/ Cd0.90Zn0.10Te0.93Se0.07 in the Visible Region, ***3.Anurag Chauhan*** and ***8.Prateek Tomar***, CSE and Electronics, DTU

**39.** Floating Solar Power Plant Installation and Indian Overview, ***8.Vaibhav***, ***8.Vardhan Suri*** and ***3.Dr. J.P. Kesari***, Software and Mechanical, DTU

**40.** Heavy metals concentrations and risk assessment in the sediment of Ganga River between Kanpur and Prayagraj, U.P., India, ***6.MOHIT AGGARWAL***, ***3.S ANBUKUMAR*** and ***3.T VIJAYA KUMAR***, Civil, DTU

**41.** HISNet: a Human Image Segmentation Network aiding bokeh effect generation, ***8.Shaurya Gupta*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

**42.** Identifying and Mitigating the Barriers for Vehicle-to-Grid Adoption in India, ***8.Aditya***, ***8.Rachit Bansal***, ***8.Akshit Goel***, ***3.N.K. Jain*** and ***3.Uma Nangia***, Electrical, DTU

**43.** Identifying Influential Nodes for Smart Enterprises using Community structure with Integrated Feature Ranking, ***3.Sanjay Kumar***, Akshi Kumar and B. S. Panda, CSE, DTU

**44.** Impact Analysis of Institutional Credit on Agricultural Output in India, Ghanshyam Pandey, Jagriti Jaiswal and ***3.Seema Kumari***, Humanities, DTU

**45.** Impact of temperature on negative capacitance based Junctionless Accumulation Mode Ferroelectric FET(JAM-FE-FET), ***6.Snehlata Yadav***, ***3.Sonam Rewari*** and ***3.Rajeshwari Pandey***, Electronics, DTU

**46.** Improved Edge Detection Approach to Tackle Edge Thickness and Better Edge Connectivity, ***8.Yatharth Saxena***, ***8.Nirdesh Mishra***, ***8.Mohd Sameer*** and ***3.Pankaj Dahiya***, Electronics, DTU

**47.** Latest Advancements in Electric Vehicles, ***8.Arhant Tewari***, ***8.Aniket Jha***, ***8.Hammad***, ***3.Dr. JP Kesari***, Mechanical and Electronics, DTU

**48.** Linearity analysis of T-gate HEMT with graded back-barrier for wireless applications, ***6.Megha Sharma***, ***6.Bhavya Kumar*** and ***3.Rishu Chaujar***, Applied Physics, DTU

**49.** Nexus of knowledge management and organizational performance: a cross-country study of China and Pakistan higher educational institutes, ***6.1.Asha Thomas***, M. Saleem Ullah Khan Sumbal, Mehwish Malik and Umar Farooq Sahibzada, Management, DTU

**50.** Nonlinear Interaction of 3D Kinetic Alfv'en Wave with Null Point and Turbulence Generation in the Solar Corona, Garima Patel, ***6.Neha Pathak***, R. Uma and R.P. Sharma, Applied Physics, DTU

**51.** Novel Ensemble of Deep Learning Neural Network and Support Vector Machine for Landslide Susceptibility Mapping in Tehri Region, Garhwal Himalaya, Sunil Saha, Anik Saha, Tusar Kanti Hembram, Barnali Kundu and ***3.Raju Sarkar***, Civil, DTU

**52.** N-type diamane: An effective emitter layer in crystalline silicon heterojunction solar cell, ***6.Naima***, ***3.Pawan K. Tyagi*** and ***3.Vinod Singh***, Applied Physics, DTU

**53.** Object Detection using Image Dehazing : A Journey Of Visual Improvement, ***8.Ritik Tanwar***, ***8.Shubham***, ***8.Shubham Verma*** and ***3.Manoj Kumar***, DTU

**54.** Optimised Algorithm for Step Count Estimation Using Sensor Data from Smartphones and Wearables, ***8.Apratim Shrivastav***, ***8.Utkarsh Kuchhal***, ***8.Samarth Singhal***, ***8.Tanmay Jain***, ***3.Divyashikha Sethia***, Vidushi Chaudhary, Rajiv Nigam and Ashish Kumar Namdeo, CSE, DTU

**55.** Overview of hydrogen production from biogas reforming: Technological advancement, ***6.Ravindra Kumar***, ***3.Anil Kumar*** and ***3.Amit Pal***, Mechanical and Environment, DTU

**56.** Performance Evaluation of Four Meta-Heuristic MPPT Algorithms for a DC-DC Converter based PV System under Dynamic Weather Conditions (DWC), ***3.Vinay Kumar***, ***8.Yash Dagar*** and ***3.Prakash Chittora***, Electrical, DTU

**57.** Prediction of COVID 19 Using Chest X-Ray Images through CNN optimised using Genetic Algorithm, ***3.Dr. Aruna Bhat***, ***8.Utkarsh Bhardwaj***, ***8.Muskaan Singla*** and ***8.Krati Garg***, CSE, DTU

**58.** Prediction of Cryptocurrency prices using Transformers and Long Short term Neural Networks, ***8.1.Ayush Tanwar*** and ***3.Vinod Kumar***, CSE, DTU

**59.** Quad MIMO Antenna for UWB Applications, ***8.R. Prashant***, ***8.Raghav Goel***, ***8.Raghav Sharma*** and ***3.Deepika Sipal***, Electronics, DTU

**60.** Quantum dots: The cutting-edge nanotheranostics in brain cancer management, Pratik Chakraborty, Sabya Sachi Das, Abhijit Dey, Apala Chakraborty, Chiranjib Bhattacharyya, Ramesh Kandimalla, Biswajit Mukherjee, Abilash Valsala Gopalakrishnan, Sandeep Kumar Singh, Shubham Kant, Parma Nand, Shreesh Ojha, ***3.Pravir Kumar***, Niraj Kumar Jha, Saurabh Kumar Jha and Saikat Dewanjee, Biotechnology, DTU

**61.** Recent Advances in Understanding SARS-CoV-2 Infection and Updates on Potential Diagnostic and Therapeutics for COVID-19, Damini Verma, Amit K. Yadav, ***6.Navneet Chaudhary***, Maumita D. Mukherjee, Pramod Kumar, Anil Kumar, Pratima R. Solanki, Biotechnology, DTU

**62.** Recent Trends in the Optimization of Renewable Distrib-uted Generation: A Review, ***6.Vivek Saxena***, ***3.Narendra Kumar*** and Uma Nangia, Electrical, DTU

**63.** Simulation and Optimization of Ultrasonic Transducer, ***8.Arkya Bagchi***, ***8.Akash Chowdhary***, ***8.Abhishek Takki*** and ***3.Raminder Kaur***, Applied Chemistry, DTU

**64.** Solar Power Driven Vehicles, ***8.1.Prajakta Tiwari***, ***8.Anjali***, ***8.Zeenat Aara*** and ***3.Dr. J.P. Kesari***, Mechanical, DTU

**65.** Speed Control Analysis of Brushless DC Motor Using PI, PID and Fuzzy-PI Controllers, ***8.Jayesh Tripathi***, ***8.Krishna Sharma*** and ***3.J.N. Rai***, Electrical, DTU

**66.** SURFACE ROUGHNESS ANALYSIS OF MACHINED HARD METAL WORK-PIECE, Anil Kumar, Dr. Neeraj Kumar and ***3.Dr. M S Niranjan***, Mechanical, DTU

**67.** Targeting Reduction in Carbon Emission of Power Plants by Regulating Supply of Biomass, ***8.Pratham Sudhakar***, ***8.Aman Yadav*** and ***3.Dr. J.P. Kesari***, Mechanical, DTU

**68.** Temperature Analysis on Short Channel Effects of Modified NCFET : A Simulation Study, ***7.Rashi Mann***, ***6.Yash Pathak*** and ***3.Rishu Chaujar***, Physics, DTU

**69.** Wear behaviour on a Pin-on-Disc setup for thermal spray of coating using ABAQUS: study and computational analysis, R. S. Walia, Ravi Pratap Singh, ***3.Ravi Butola*** and Mohit Tyagi, Mechanical, DTU

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

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

# A Fully-Digital BIST Wrapper Based on Ternary Test Stimuli for the Dynamic Test of a 40 nm CMOS 18-bit Stereo Audio ΣΔ ADC

Manuel J. Barragan, Rshdee Alhakim, Haralampos-G. Stratigopoulos, Matthieu Dubois, Salvador Mir, Hervé Le-Gall, Neha Bhargava, Ankur Bal

▶ **To cite this version:**

Manuel J. Barragan, Rshdee Alhakim, Haralampos-G. Stratigopoulos, Matthieu Dubois, Salvador Mir, et al.. A Fully-Digital BIST Wrapper Based on Ternary Test Stimuli for the Dynamic Test of a 40 nm CMOS 18-bit Stereo Audio ΣΔ ADC. IEEE Transactions on Circuits and Systems I: Regular Papers, IEEE, 2016, 10.1109/TCSI.2016.2602387 . hal-01447789

**HAL Id: hal-01447789**

**https://hal.archives-ouvertes.fr/hal-01447789**

Submitted on 2 Sep 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TCSI.2016.2602387

# A Fully-Digital BIST Wrapper Based on Ternary Test Stimuli for the Dynamic Test of a 40nm CMOS 18-bit Stereo Audio $\Sigma\Delta$ ADC

Manuel J. Barragan, *Member, IEEE,* Rshdee Alhakim, *Member, IEEE,*
Haralampos-G. Stratigopoulos, *Member, IEEE,* Matthieu Dubois, Salvador Mir, *Member, IEEE,* Hervé Le Gall,
Neha Bhargava, and Ankur Bal

***Abstract*—This paper proposes a fully-digital BIST architecture for the dynamic test of $\Sigma\Delta$ ADCs. The proposed BIST relies on generating a ternary stream that encodes a high-linearity analog sinusoidal and injecting it directly at the input of the $\Sigma\Delta$ modulator. Compared to the well-known bitstream, the use of three logic levels in the ternary stream reduces the quantization noise and, thereby, results in a test with a higher dynamic range that covers the full scale of the ADC. The output response is analyzed on-chip using a simplified version of the sine-wave fitting algorithm to compute the SNDR. A standard SPI bus provides digital external access to the embedded test instruments. The proposed BIST wrapper has been integrated into a 40 nm CMOS 18-bit stereo audio $\Sigma\Delta$ ADC IP core provided by ST Microelectronics. It incurs an overall area overhead of 7.1% and the total test time is 28ms per channel. Experimental results on fabricated chips demonstrate an excellent correlation between the BIST and the standard functional specification test.**

## I. Introduction

CURRENT market trends in the IC industry towards the design of highly integrated mixed-signal Systems-on-Chip (SoCs) and heterogeneous Systems-in-Package (SiPs) are supported by the reuse of parametrized Intellectual Property cores (IPs) –either in-house developed or provided by different IP vendors– that implement different analog, mixed-signal, and digital functions within the system. Testing the digital IP cores in these systems is a relatively straightforward task. A wide variety of broadly proven Design-for-Test (DfT) strategies for accessing these cores and evaluating their correct functionality are routinely embedded in every digital IP. These DfT strategies are built on the concepts of standardized scan chain buses and structural tests. On the other hand, testing mixed-signal IP cores in these systems still relies on functional specification testing. Functional specification testing uses specialized automated test equipment (ATE) for applying appropriate test stimuli, retrieving the output responses, and processing the test results. However, in a complex SoC/SiP –where the external access to the internal IPs is difficult, or directly impossible– these operations may turn challenging to perform at a reasonable cost [1]. Indeed, it has been estimated that the cost of testing mixed-signal IP cores has already become the dominant contribution to the overall manufacturing cost [2].

A promising solution to the aforementioned issues is the development of mixed-signal Built-In Self-Test (BIST) techniques. BIST techniques aim at moving some of the ATE functionality into each IP, in such a way that generation and manipulation of test signals become internal to the IP, thus eliminating the need of external analog test access and relaxing the ATE requirements. In an ideal scenario, each analog and mixed-signal IP core should have its own dedicated embedded test instruments adapted to its custom-made test programs and interfaced via a system-level bus to a low-cost digital ATE for controlling test execution and retrieving test results. This way, BIST may significantly reduce the overall production cost of these systems by taking into consideration the test approach right at the IP design stage. Moreover, the benefits of deploying mixed-signal BIST solutions are not limited to production test. BIST opens the door to further developments for performance and reliability enhancements along the complete life-cycle of an integrated circuit, including calibration, self-healing, on-line test, adaptive operation, fault diagnosis and failure analysis, etc.

In this paper, we present the design and experimental validation of a fully-digital BIST wrapper for the dynamic test of a 40nm CMOS 18-bit stereo audio $\Sigma\Delta$ ADC IP core designed by ST Microelectronics. Testing the dynamic performance of an ADC in terms of Signal-to-Noise-and-Distortion ratio (SNDR) requires exciting the ADC Under Test (ADCUT) with a high-resolution sinusoidal stimulus –at least two effective bits higher than the resolution of the ADCUT. Moreover, the demand for high-quality audio devices is driving the improvement in the performance of $\Sigma\Delta$ ADCs toward higher resolutions and dynamic ranges. Therefore, generating the required test stimulus efficiently on-chip towards a fully integrated BIST implementation is a very challenging task. The proposed BIST wrapper achieves this objective by relying on the generation of a digital ternary test stimulus that encodes a high-resolution sine-wave. The ternary test stimulus is directly injected into the input of the modulator of the $\Sigma\Delta$ ADC

M. J. Barragan, R. Alhakim, and S. Mir are with CNRS, Université Grenoble–Alpes, TIMA, F-38000, Grenoble, France (e-mail: manuel.barragan@imag.fr).

H. Stratigopoulos is with Sorbonne Universités, UPMC Univ. Paris 06, CNRS, LIP6, 4 place Jussieu, 75005, Paris, France.

M. Dubois is with Pyxalis, CentrÕAlp, 170 rue de chatagnon, 38346 Moirans, France.

H. Le Gall is with ST Microelectronics, 12 Rue Jules Horowitz, 38000, Grenoble, France.

N. Bhargava and A. Bal are with ST Microelectronics, Plot No. 1, Knowledge Park III, 201308, Greater Noida, India.

Copyright (c) 2016 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.

requiring minimal design modifications. The BIST wrapper also includes a test response analyzer for computing the SNDR on-chip and providing a 1-bit go/no-go test response, as well as a control unit for managing the test execution. The proposed BIST wrapper incurs an overall area overhead of 7.1% and the total test time is 28 ms per channel. We discuss in detail the co-design of the BIST wrapper and the $\Sigma\Delta$ ADC and we present experimental results based on fabricated chips.

The rest of the paper is organized as follows. In Section II, we give an overview of the state-of-the-art for $\Sigma\Delta$ ADC BIST. In Section III, we present the architecture of the proposed BIST wrapper. In Section IV, we present the experimental validation of the proposed BIST approach. Finally, Section V concludes the paper summarizing the main contributions.

## II. Previous Work

In recent years there has been an intense research effort towards developing BIST strategies for $\Sigma\Delta$ ADCs. In [3], an Oscillation-Based BIST (OBIST) strategy for $\Sigma\Delta$ modulators is proposed that consists of reconfiguring the modulator under test into an oscillator. Both the frequency and amplitude of the oscillation are monitored and processed to distinguish faulty from fault-free circuits. This test strategy provides a structural verification of the modulator under test targeting defects. However, the sensitivity to process variations and the need of extensive circuit reconfiguration can deteriorate the quality of this approach.

In [4], [5], it is proposed to estimate the magnitude of the leakage in the integrators of the modulator which is directly related to the in-band quantization noise power at the modulatorÕs output, thus allowing to estimate the dynamic performance of the modulator. However, as demonstrated in [6], additional tests targeting different modulator parameters (i.e. settling time, amplifier non-linearity, etc.) are required for a complete validation of the modulatorÕs performance.

Another idea that has been thoroughly explored in the literature is the use of digital $\Sigma\Delta$ bitstreams. The starting point is to use an ideal $\Sigma\Delta$ modulator in software to convert a high-resolution analog sine-wave to a digital bitstream. The bitstream is truncated to get an optimized sequence that is periodically reproduced using a circular shift-register [7]. Bitstreams can also be generated using on-chip digital oscillators, but their design is not straightforward [8]. The bitstream can be converted to a high-resolution analog sine-wave by passing it through a 1-bit digital-to-analog converter (DAC) followed by a low-pass filter to remove quantization noise. Although this solution is applicable for a built-off test where the 1-bit DAC and the low-pass filter are placed on the load board, it is not applicable for BIST since for high-resolution applications the low-pass filter may be far more complex than the modulator under test [9], [10].

Interestingly, in the case of Switched-Capacitor (SC) $\Sigma\Delta$ ADCs, the bitstream can be fed directly into the modulator by adding simple circuitry at its input [11]–[13]. A 1-bit DAC – typically a replica of the feedback DAC in the modulator input stage– is used to convert the bitstream to the analog reference voltages of the modulator. However, the direct injection of the bitstream to the input of the ADC will overload the modulator, unless the reference voltages of the 1-bit DAC are adequately attenuated [12]. This attenuation also scales down the amplitude of the encoded sine-wave, which, in effect, will be far below the Full Scale (FS) of the ADCUT. This is a key limitation of this test technique, since for an accurate dynamic test the amplitude of the test stimulus has to be as close as possible to the FS such that the harmonic distortion introduced by the ADCUT is amply manifested [14].

A solution to this problem has been proposed in [15], [16], where the input switched-capacitor network of the $\Sigma\Delta$ modulator is reconfigured in test mode such that it accepts multiple bitstreams delayed from each other. This technique has been proved to attenuate the noise power of the test stimulus, allowing this way to test the modulator at amplitudes closer to the FS. However, this comes at the cost of significantly increasing the design complexity of the input stage of the modulator.

An elegant approach to overcome this issue was proposed in [17]. A sinusoidal test stimulus is encoded in a ternary stream that is composed of three logic levels $\{-1,\, 0,\, +1\}$ and injected directly at the input of the modulator of the $\Sigma\Delta$ ADC under test. The ternary stimulus is generated in the digital domain by adding a $\Sigma\Delta$ encoded binary bitstream with a delayed version of itself. From a spectral point of view, the Discrete Fourier Transform (DFT) of a length-$N$ bitstream $\{b_0, \cdots, b_{N-1}\}$ is given by

$$B(k) = \sum_{n=0}^{N-1} b_n \cdot e^{-j\frac{2\pi \cdot n}{N}k}, \quad k = 0, \cdots, N-1. \tag{1}$$

The $N$ periodic extension leads to a spectrum made by $N$ lines located at $f_k = kf_s/N$, where $f_s$ is the sampling frequency. The DFT of the circular-shifted bitstream delayed by one sample $\{b_1, \cdots, b_{N-1}, b_0\}$ is obtained by applying the time shift theorem to (1)

$$B_\delta(k) = B(k) \cdot e^{-j\frac{2\pi}{N}k}, \quad k = 0, \cdots, N-1. \tag{2}$$

The DFT of the ternary stream is then obtained by averaging the DFT of the original bitstream and the DFT of the circular-shifted version of it and by using the linearity property of the DFT

$$T(k) = \frac{B(k)}{2}\left(1 + e^{-j\frac{2\pi}{N}k}\right), \quad k = 0, \cdots, N-1. \tag{3}$$

The Power Spectral Density (PSD) of the bitstream and the ternary stream are given by

$$\begin{aligned} S_B(k) &= \|B(k)\|^2 \\ S_T(k) &= S_B(k) \cdot cos^2\left(\frac{\pi k}{N}\right) \\ & \qquad k = 0, \cdots, N-1, \end{aligned} \tag{4}$$

respectively. Therefore the spectra of the ternary stream presents a lowpass filter behavior that attenuates the shaped quantization noise at high frequencies. In other words, the ternary stream overloads less the modulator compared to the bitstream and, thereby, offers the possibility of testing the $\Sigma\Delta$ ADC closer to full scale. In fact, the proposed ternary

TABLE I
SUMMARY OF PERFORMANCE AND DESIGN PARAMETERS OF THE ADCUT

| | |
|---|---|
| Input range | 1.4 V |
| Input bandwidth | 24 kHz |
| Sampling frequency (MASH input) | 6.144 MHz |
| Sampling frequency (ADC output) | 48 kHz |
| Nominal resolution | 18 bits |
| SNDR (typical) | 94 dB |
| Analog power supply | 3.3 V |
| Digital power supply | 1.1 V |
| Power consumption | 41.8 mW |

encoder is equivalent to a first-order lowpass FIR filtering of the binary stream to attenuate the shaped quantization noise. In principle, higher-order FIR filtering could be used to reduce the quantization noise of the test stimulus even further, but doing so would also require a complex high-linearity multi-bit DAC to inject the digitally filtered test stimulus at the input of the modulator.

On the other hand, the injection of the ternary stimulus at the input of the modulator does not require a significant reconfiguration of the input stage. In particular, the ternary stream is converted to a three-level analog signal and injected directly at the input stage of the modulator. In fact, the implementation is, in essence, equivalent to the case of applying a binary bitstream. The proposed approach presents a minimal overhead and allows measuring SNDR for amplitudes close to FS.

The theoretical principles behind the concepts of ternary test stimulus generation and injection were described in [17] and were validated using behavioral model simulations. The feasibility and suitability of the ternary test stimulus for high-volume production test was further validated in [18], [19] by means of parametric test metrics estimation techniques based also on behavioral model simulations.

In this paper, we present for the first time a hardware implementation of the BIST approach based on ternary test stimuli.

## III. SELF-TESTABLE ΣΔ ADC ARCHITECTURE

### A. System overview

Fig. 1 shows a conceptual block diagram of the complete ΣΔ ADC IP core incorporating the BIST wrapper for measuring on-chip the SNDR. In normal operation mode, the analog input signals for Left and Right channels are first filtered by an active-RC second-order low-pass anti-aliasing filter that also provides a single-ended-to-fully-differential signal conversion. A 2:1 MASH SC modulator converts these differential signals to a digital bitstream and a fourth-order digital decimation filter eliminates the shaped high-frequency noise to provide the 18-bit digital output. The IP features a single output port with 18 bits in parallel, where the Left and Right channel outputs are time-interleaved according to the rising and falling edges of a control signal (labelled FSYNC_IN in Fig. 1). Relevant performance and design parameters of the ADC IP are summarized in Table I.

In BIST mode, an optimized digital bitstream stored in a circular shift-register is periodically repeated and combined with a delayed version of itself to generate a digital ternary test stimulus encoding a high-resolution sinusoidal test stimulus within the bandwidth and input dynamic range of the ADCUT. The ternary test stimulus is injected directly at the input of the 2:1 MASH modulator of one of the ADC channels, which has been previously selected and disconnected from the input signal path. The resulting output bitstream is routed to the Left channel of the decimation filter to produce an 18-bit test response signal, labelled $S_{out}$ in Fig. 1. At the same time, the original optimized digital bitstream is fed to a fully-digital implementation of the 2:1 MASH modulator. The resulting output bitstream is fed to the Right channel of the decimation filter to produce a high-precision 21-bit digital sine-wave, labelled $S_{ref}$ in Fig. 1, that serves as the reference signal. The digital 2:1 MASH assures the synchronization between Right and Left channel output signals, since both the modulator in the channel under test and the digital 2:1 MASH introduce a $z^{-3}$ signal delay. Then, the test output response from the Left channel and the reference signal from the Right channel are extracted from the time-interleaved output of the ADCUT using the synchronization signal FSYNC_IN and fed to the test response analyzer block. This block computes the SNDR of the channel under test from the response and reference signals using a simplified version of the sine-wave fitting algorithm. The obtained SNDR is finally compared to a pre-loaded specification limit and a go/no-go digital signal is generated. The complete test is managed by a BIST control block which also includes a standard Serial-Peripheral Interface (SPI) digital bus for external programming and communication.

The following subsections detail the design and functionality of the main building blocks of the proposed BIST wrapper.

### B. Ternary stimulus generator

Fig. 2 shows a schematic view of the on-chip digital ternary test stimulus generator and its interface signals with the ADCUT, the BIST control, and the digital 2:1 MASH modulator. The core of the ternary test stimulus generator circuitry is a circular shift-register composed of $n = 1126$ flip-flops. During test mode, this circular shift-register stores and reproduces a bitstream that encodes one period of a high-precision sinusoidal test stimulus. The circular shift-register is clocked at the modulator sampling frequency of $f_s = 6.144$ MHz and produces a bitstream with a fundamental tone at $f_{test} = f_s/n = 5.46$ kHz. This frequency is fixed by design and it has been chosen such that the main tone of the test stimulus and the first three harmonic components (i.e. up to the fourth harmonic component) lie within the audio bandwidth of the ADCUT. The bitstream is obtained by simulating in software an ideal ΣΔ modulator excited by a pure sine-wave.

The ternary test stimulus is obtained by adding the bitstream stored in the circular shift-register with a version of the same bitstream delayed by one clock cycle. As demonstrated in [17] and discussed in Section II, this operation has a lowpass filtering effect on the shaped quantization noise at high frequencies and achieves a significant reduction of the total quantization noise power with respect to the original

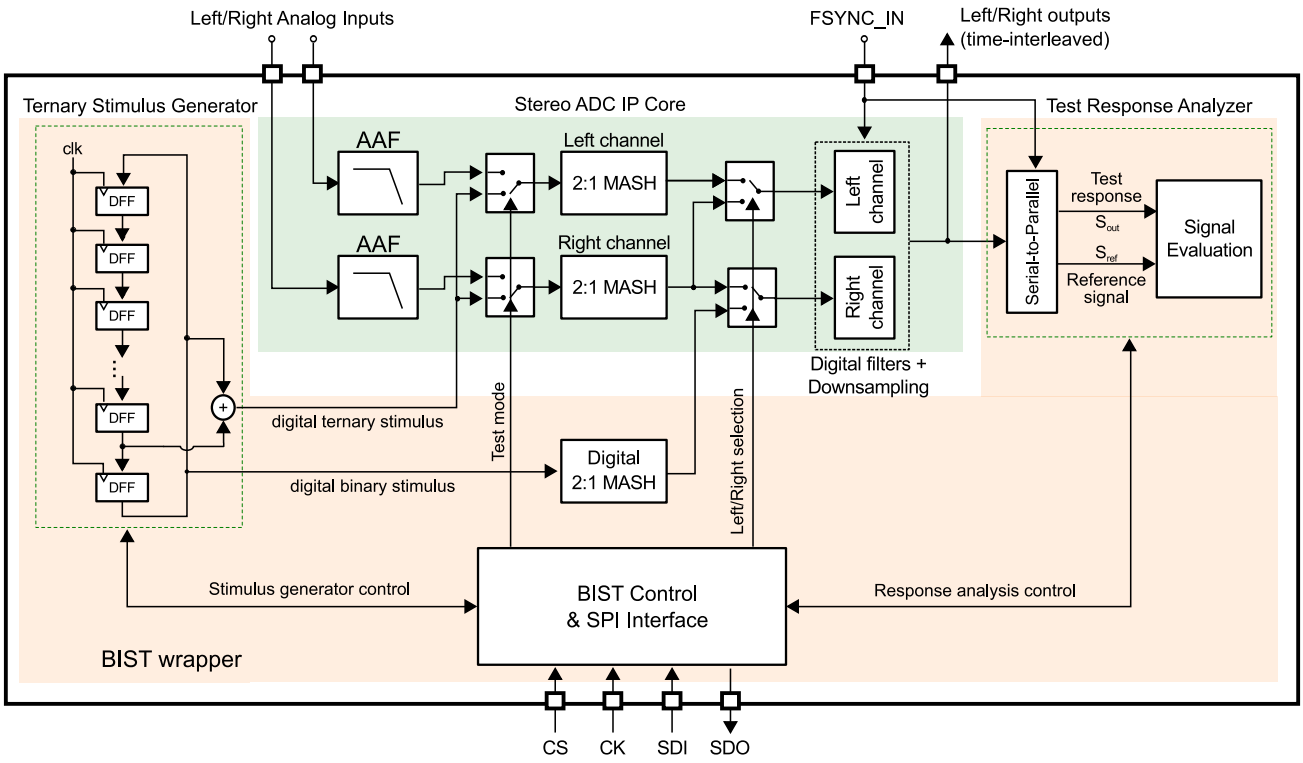



Fig. 1. Architecture of the self-testable ΣΔ ADC.

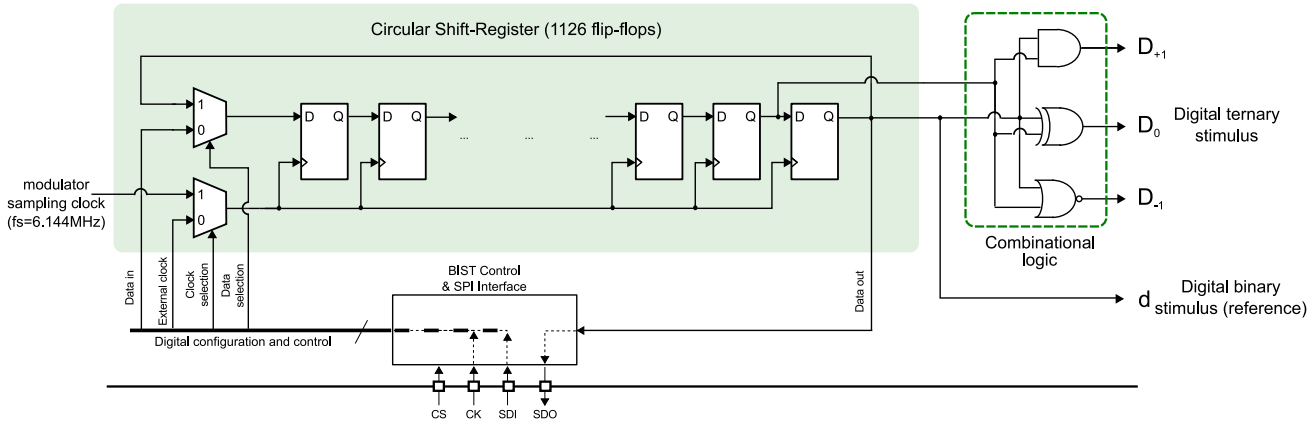



Fig. 2. Micro-architecture of the ternary test stimulus generator block.

bitstream. The combination of the two bitstreams only requires the addition of three logic gates to generate the three digital signals $D_{-1}$, $D_0$, and $D_{+1}$ encoding the three levels of the ternary test stimulus.

It is widely accepted in the industry that the test stimulus has to be at least 2 effective bits more linear than the specification of the ADCUT for accurate and reliable results. Otherwise, the non-linearity of the ADCUT is masked by the non-linearity of the test stimulus [20]. For this purpose, an optimization is performed with the aim to select the bitstream that maximizes the SNDR of the ternary test stimulus and renders the ternary test stimulus appropriate for an accurate dynamic testing of the ADCUT. The interested reader is referred to [17] for a detailed description of the optimization strategy.

Fig. 3 plots together the nominal SNDR of the 2:1 MASH modulator in the ADCUT as a function of its input amplitude and the SNDR of the optimized ternary test stimulus as a function of the amplitude of the encoded sine-wave, both measured at $f_{test} = 5.46$ kHz. The amplitude is varied across the complete FS of the modulator. As it can be seen, the optimization loop manages to find ternary test stimuli that have a SNDR at least 2.5 equivalent bits higher than the SNDR of the modulator, thus respecting the requirement that the test stimulus must have a resolution at least two effective bits higher than the resolution of the ADCUT.

It is important to note that the proposed BIST wrapper has been designed for production testing, where a single fixed-frequency in-band test stimulus is applied according to the standard [20]. Therefore, the proposed BIST is not meant for

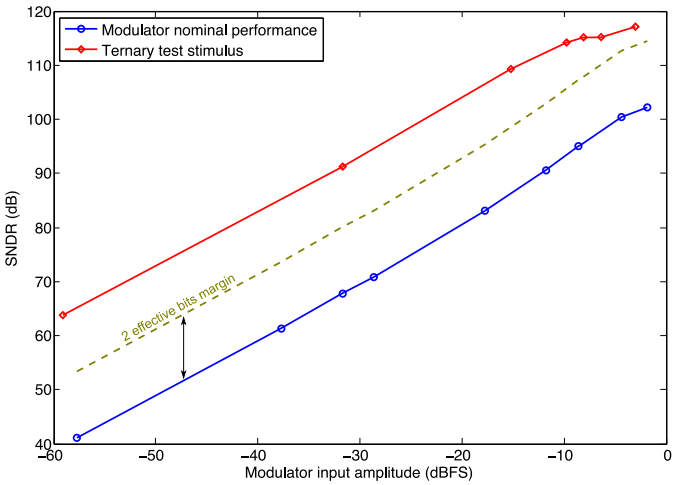



Fig. 3. SNDR of the modulator and the optimized ternary test stimulus as a function of the input amplitude.

full electrical characterization across the complete bandwidth of the ADCUT. Furthermore, pushing the frequency of the test stimulus towards the bandwidth limit is not a straightforward task. From a hardware point of view, the frequency of the stimulus can be easily increased by reducing the number of flip-flops, $n$, in the circular shift-register or, equivalently, by encoding multiple signal periods into the register. The optimization algorithm can also be easily redesigned, in order to get appropriate test stimuli with a high enough SNDR at high frequencies. However, obtaining high-linearity stimuli close to the bandwidth limit may require employing a high-order $\Sigma\Delta$ modulation scheme within the optimization loop. The challenge in this scenario is that the decimation filter in the ADCUT itself may need to be modified so as to cope with the shaped quantization noise.

The ternary stimulus generator is managed by the BIST control block and is interfaced to the external pins (or to other IPs in a SoC/SiP) through the SPI bus. The optimized bitstream is loaded into the circular shift-register through the Serial Data Input (SDI) port. Serial data are synchronized to an external clock fed to the BIST control through the CK port. The content of the circular shift-register can also be read out for additional verification and diagnosis. Output data are sent out through the Serial Data Output (SDO) port and are synchronized to the external clock. During test execution, the BIST control unit enables the test mode, closes the feedback loop of the circular shift-register, and clocks the circular shift-register at the sampling clock frequency of the modulator (i. e. clock phase $\phi_1$ in Fig. 4).

### C. Ternary test stimulus injection

The ternary test stimulus is directly injected into the input of the modulator using a three-level DAC controlled by signals $D_{-1}$, $D_0$, and $D_{+1}$. Fig. 4 shows the original schematic of the input switching stage of the 2:1 MASH modulator and the diagram of its clock phases. The 2:1 MASH modulators in the ADC IP core work with two non-overlapped clock phases, denoted by $\phi_1$ and $\phi_2$ in Fig. 4. The input sampling operation is divided into two sub-phases within clock phase $\phi_1$. During clock sub-phase $\phi_{1a}$, the input capacitors are reset by shorting their nodes to the common-mode voltage. During clock sub-phase $\phi_{1b}$, the input capacitors are charged to the input voltages. The digital output bitstreams at the output of the modulator, denoted by *outXYZ*, are fed back to the input of the modulator in clock phase $\phi_2$, which is also divided into two steps. In a first step, the feedback signals *out1YZ* are activated, connecting the input node to reference voltages *AGND* or *VHX*. Note that index *Y* makes reference to the signal path –positive or negative– to which the switch is connected, while index *Z* makes reference to the type of transistor. Reference voltages *AGND* and *VHX* offer a fast signal settling at the cost of a low precision. They are used to pre-load the input capacitor to a value close to the $\Sigma\Delta$ ADC references. In a second step, feedback signals *out2YZ* are activated, connecting the input node to the ADC reference voltages *VL* or *VH*. These reference voltages offer a high precision reference value, but signal settling is slower. The division of the modulator feedback operation in these two steps allows relaxing the design of the integrated bandgap reference generator.

Fig. 5 shows the modification of the input switching stage for the purpose of injecting the ternary test stimulus. During test mode, control signal $T$ is set to "1" by the BIST control block, turning off the input sampling switches, denoted by *S6* and *S6'* in Fig. 5, and, thereby, isolating the modulator under test from the input signal path. At the same time, extra switches *STX* are connected to the input node of the sampling capacitors. This set of switches implements a simple and robust three-level differential DAC that reconstructs the analog sine-wave encoded in the ternary test stimulus $\{D_{-1}, D_0, D_{+1}\}$ and injects it directly into the input sampling capacitors of the modulator under test. As opposed to [12], it is not required to scale down the reference voltages *VH* and *VL*, thus no modifications of the bandgap reference generator and no extra external analog references are needed. States "+1" and "−1" in the ternary test stimulus are directly converted to the differential analog levels $VH - VL$ and $VL - VH$, respectively, while the "0" state is implemented by generating a null differential voltage at the input nodes of the sampling capacitors. This is achieved by connecting the input capacitors to the common-mode voltage $V_{cm}$ through switches *ST1* and *ST1'* to discharge them. Alternatively, this null differential voltage could have been produced by activating switches *S1* and *S1'* during clock phase $\phi_{1b}$, but this would have required an extensive modification of the internal clock phase generator to avoid overlapping between clock phases $\phi_{1a}$ and $\phi_{1b}$ during normal operation. We opted for adding independent switches *ST1* and *ST1'* since it is a simpler option.

The overall overhead for injecting the ternary test stimulus is only the six switches *ST1*, *ST1'*, *ST2*, *ST2'*, *ST3*, and *ST3'* which is equivalent to only four CMOS switches. From a design point of view, switches *ST1* and *ST1'* are an exact replica of switches *S1* and *S1'*, whereas switches *ST2*, *ST3*, *ST2'*, and *ST3'* are an exact replica of switches *S4*, *S5*, *S4'*, and *S5'*, respectively.

An aspect that should be carefully considered in the design of the three-level DAC is the impact of the DAC non-linearity

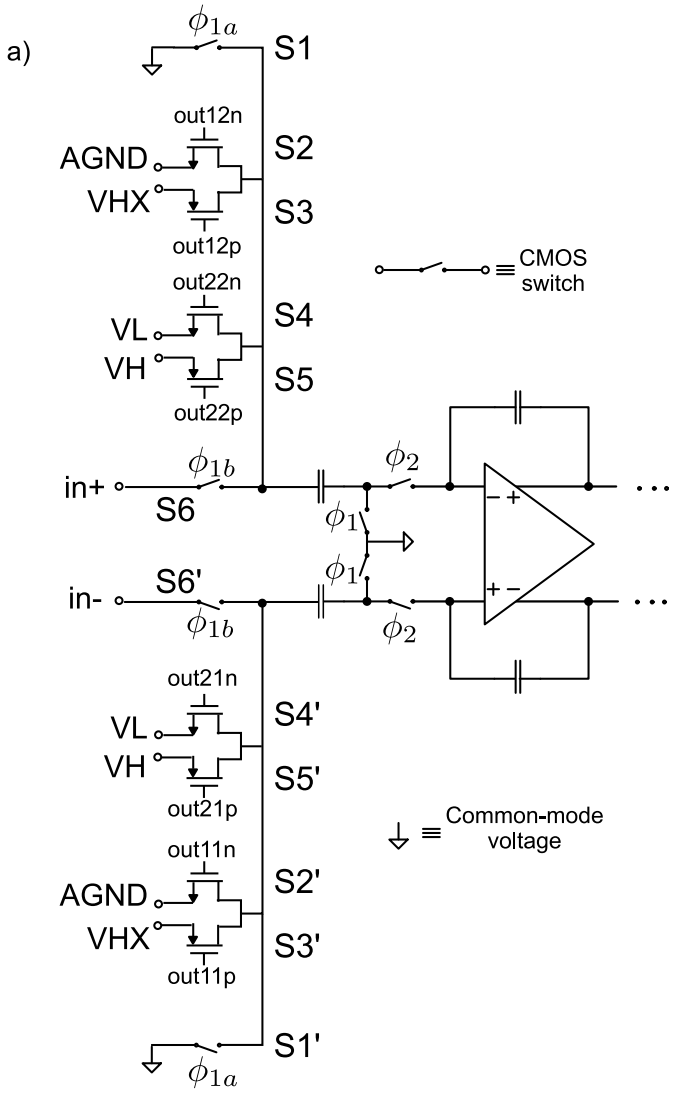



Fig. 4. a) Fully-differential input stage of the 2:1 MASH modulators in the ΣΔ ADC IP core; b) Diagram of modulator's clock phases.

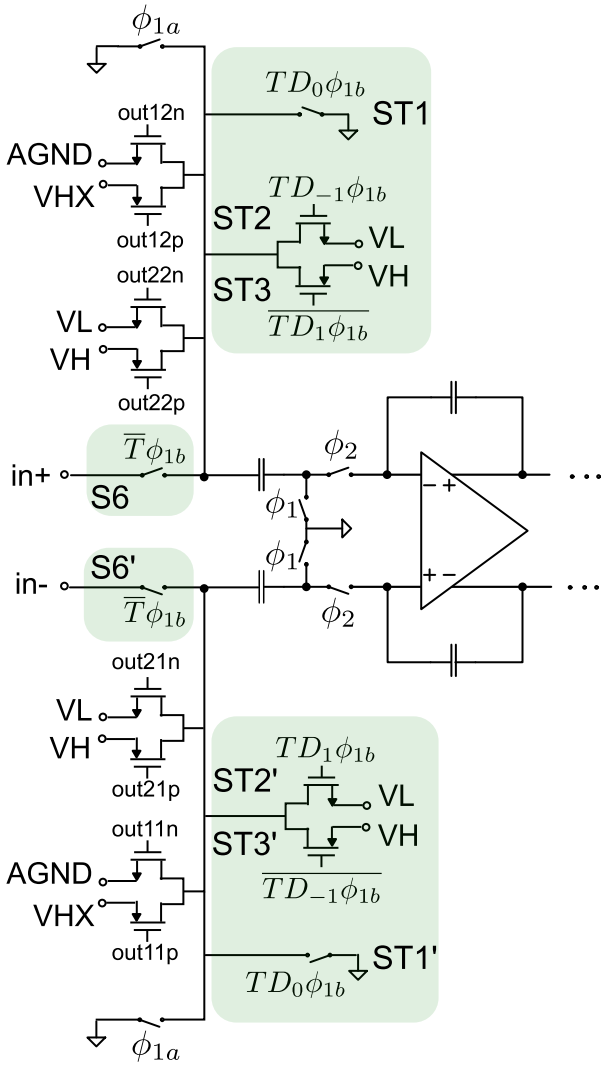



Fig. 5. Modification of the fully-differential input stage of the 2:1 MASH for the purpose of injecting the ternary test stimulus. The modifications have been highlighted.

on the spectral purity of the ternary test stimulus. An ideal three-level DAC should have the linear input-output characteristic depicted in Fig. 6(a). However, second-order non-idealities, such as mismatch and unbalanced charge injection between positive and negative signal paths, may lead to DAC non-linearity. In order to analyze the impact of these non-idealities on the proposed three-level DAC, we consider a simple model, as shown in Fig. 6(b). The effect of the non-indealities on the DAC levels has been modeled as an equivalent deviation on the middle-point output voltage, labelled as $V_{off}$ in Fig. 6(b). In general, $V_{off}$ may also depend on the input digital sequence of the DAC. However, the fully-differential DAC implementation, together with the reset of the sampling capacitor during phase $\phi_{1a}$ before injecting each sample of the test stimulus, result in $V_{off}$ that is approximately constant. Using this simple model for the proposed three-level DAC design, Fig. 7 shows the SNDR at the output of the DAC when the magnitude of $V_{off}$ is swept from 1 nV to 1 mV. In this analysis, the DAC is excited by a digital ternary test stimulus encoding a sinusoid with an amplitude $A = -3$ dBFS. The SNDR of this test stimulus is SNDR = 117 dB and is indicated in Fig. 7 for visual reference. As it can be observed, the linearity in terms of SNDR of the signal at the DAC output drops quickly for $V_{off} > 10$ $\mu$V.

To this end, we performed a transistor-level Monte Carlo simulation of the actual three-level DAC that is employed in

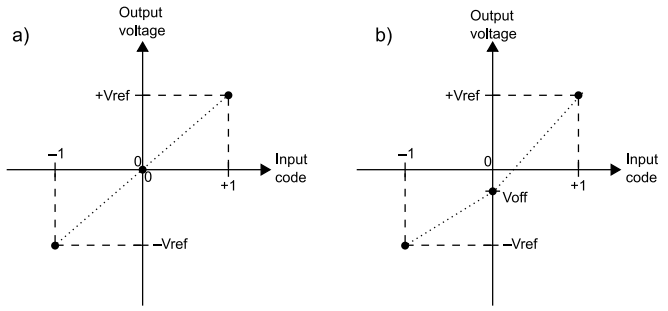



Fig. 6. a) Input-output characteristic of an ideal three-level DAC; b) Non-linear model for a three-level DAC. Non-linear effects are combined into an equivalent offset in the middle-point level of the DAC.

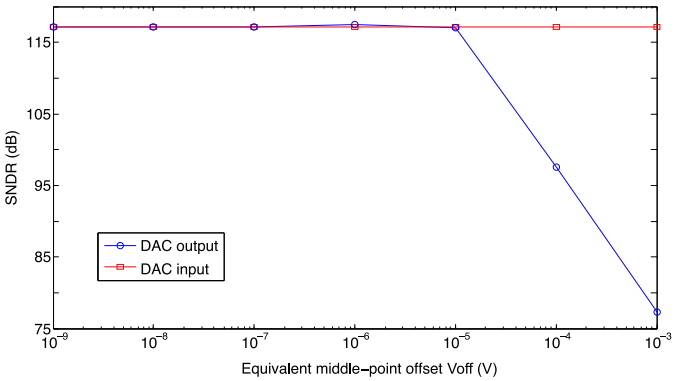



Fig. 7. SNDR at the output of the three-level DAC as a function of the equivalent middle-point offset. The DAC is excited by a digital ternary stimulus encoding a sinusoidal signal with $A = -3$ dBFS and SNDR $= 117$ dB

the BIST wrapper considering both global process variations and mismatch. The Monte Carlo analysis shows a worst equivalent deviation of the middle-point voltage of the DAC $V_{off} < 10$ nV. Referring back to the results in Fig. 7, we conclude that the non-linearity of the proposed DAC has no effect on the BIST technique.

### D. Test response analyzer

The test response evaluation is performed by a version of the sine-wave fitting algorithm [20], [21]. Many efficient implementations of the sine-wave fitting algorithm have been proposed in the literature [12], [22], [23]. In this work, we implement a simplified version of the sine-wave fitting algorithm that was first introduced in [12]. The classical sine-wave fitting algorithm aims at fitting in a least-square sense a pure sinusoidal reference signal to the output response of the ADCUT. The goal is to determine the DC level, amplitude, frequency, and phase of the fundamental tone in the test response signal. Noise and distortion components can then be easily computed as the residues of the fitted curve. The interested reader is referred to [20], [21] for a detailed analysis of the sine-wave fitting technique. The proposed simplification of the sine-wave fitting algorithm is based on the fact that the test response and reference signals have the same frequency and they are completely synchronized, which saves us from having to compute the phase and frequency of the test response signal. Next, we describe the algorithm for computing the SNDR and making a direct go/no-go test decision on-chip. The algorithm is conceptually depicted in Fig. 8.

The algorithm first computes the DC component of the test response signal as

$$DC = \frac{1}{N}\sum_{i=1}^{N} S_{out}(i), \tag{5}$$

where $S_{out}(i)$ are the samples of the test response signal at the output channel under test and $N$ is the total number of samples considered in the evaluation. In our implementation, the number of samples is fixed to $N = 563$ which corresponds to 64 signal periods coherently sampled. Coherent sampling is not required by the classical sine-wave fitting algorithm, but it further simplifies the computations and, thereby, it reduces significantly the resources required for the hardware implementation. It should be stressed that coherent sampling is easily achieved by construction, since, as explained in Section III-B, the period of the test stimulus is an integer multiple of the modulator sampling clock period.

In the next step, the algorithm computes the point-to-point correlation of test response and reference signals as

$$Correl = \sum_{i=1}^{N} S_{out}(i) S_{ref}(i), \tag{6}$$

where $S_{ref}(i)$ are the samples of the reference signal. The amplitude of the test response signal is related to the computed correlation as

$$A = \frac{2}{N A_{ref}} Correl, \tag{7}$$

where $A_{ref}$ is the amplitude of the reference signal encoded in the ternary test stimulus, which is known *a priori*.

Once the amplitude has been computed, the algorithm adjusts the reference signal to the amplitude $A$ and DC values as

$$S_{ref,adj}(i) = \frac{A}{A_{ref}} S_{ref}(i) + DC. \tag{8}$$

Finally, the algorithm computes the noise and distortion power in the test response signal, denoted by $P_{error}$, as

$$P_{error} = \frac{1}{N}\sum_{i=1}^{N} \left(S_{out}(i) - S_{ref,adj}(i)\right)^2. \tag{9}$$

The SNDR of the test response signal can be computed as

$$SNDR = 10 \log \frac{A^2/2}{P_{error}}. \tag{10}$$

However, to avoid an excessive overhead, the test response analyzer does not compute directly the SNDR value. Instead, it performs a comparison with a preloaded threshold to provide a go/no-go test decision. From (10)

$$\text{go/no-go} = \begin{cases} 1, & \text{if } \frac{A^2}{2} \geqslant P_{error} 10^{SNDR_{spec}/10} \\ 0, & \text{if } \frac{A^2}{2} < P_{error} 10^{SNDR_{spec}/10} \end{cases} \tag{11}$$

where $SNDR_{\text{spec}}$ is the SNDR specification of the ADCUT. The threshold value $Threshold = 10^{SNDR_{\text{spec}}/10}$ is a known

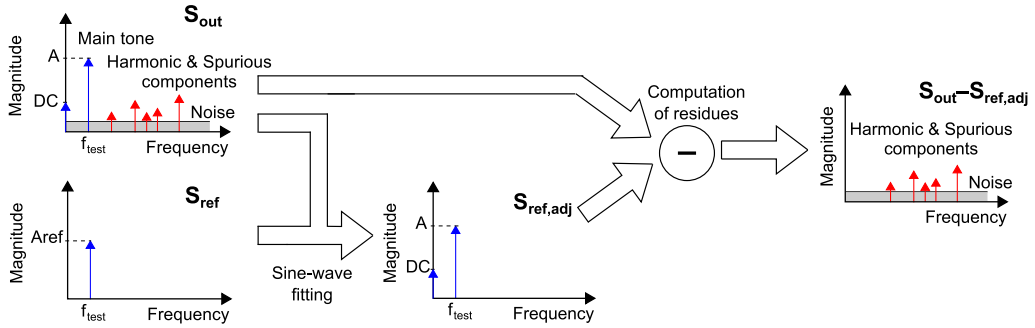



Fig. 8. Graphical representation of the proposed algorithm for test response analysis.

constant that is loaded into the test response analyzer prior to the evaluation.

The proposed test response analyzer has been implemented with an embedded Arithmetic Logic Unit (ALU) and a Finite-State Machine (FSM) controller, schematically depicted in Fig. 9. The ALU has three main blocks –addition, multiplication, and division– together with two blocks for complement operations and one block for comparison operations. The blocks for the complement operations are required to enable subtractions, while the block for comparison is required to compare $P_{error}$ with the preloaded threshold value. The proposed implementation uses Booth's multiplication algorithm to build the multiplier, while the division is actually performed by multiplying by the inverse of the denominator. The FSM controller is programmed to evaluate equations (5)-(11) sequentially, using only one adder, one multiplier, and one divider. In order to further reduce the area overhead of the test response analyzer, $S_{out}$ and $S_{ref}$ samples are not stored into a memory, but sequentially read from the outputs of the ADCUT and fed to the evaluation algorithm. As a consequence, the evaluation algorithm is executed in two sequential steps: in the first step the algorithm evaluate equations (5)-(7) over $N$ ADCUT output samples, while in the second step $N$ additional samples are read to evaluate equations (8)-(11).

The functionality of the test response analyzer is also controlled by the BIST control block through the SPI bus. The reference values $A_{ref}$ and *Threshold* are externally loaded through the SDI port, synchronized to the external clock fed by the CK port. The BIST control block manages the execution of the evaluation algorithm and outputs the go/no-go test decision through the SDO port at the end of the evaluation.

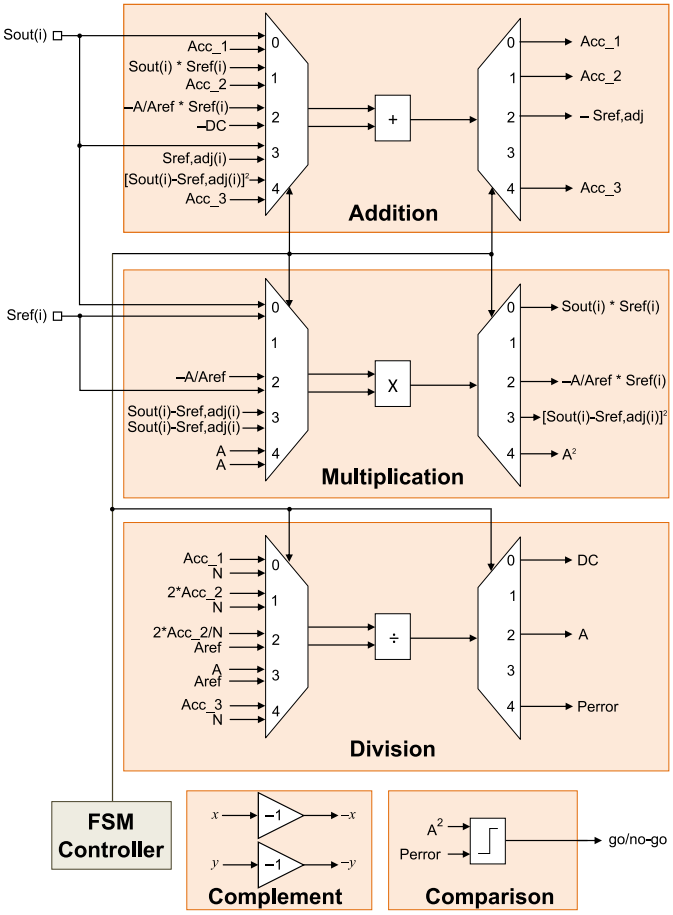



Fig. 9. Micro-architecture of the test response analyzer.

## *E. BIST control and SPI interface*

The BIST control block manages the test program by supervising the execution of the test and, as it has been described previously, it also interfaces the BIST wrapper with the external ATE through a standard SPI bus. The BIST control block consists of four main units, as shown in Fig. 10, namely a SPI Slave port Unit, a Fetch and Execute Unit, a Configuration Unit, and a Reset Unit.

The SPI Slave port Unit has been designed to provide a robust and simple serial interface between the BIST wrapper and the ATE. This standard serial connection allows the ATE to configure the test parameters (i.e. load input stimulus, write configuration registers, etc.) and read out the test results through the SPI standard ports CS, CK, SDO, and SDI. The SPI Slave port Unit is connected to the Fetch and Execute Unit using a parallel interface which is clocked at the master system clock of the IP, which is denoted by SYS_CK in Fig. 10. The SPI Slave port Unit works on two asynchronous clock domains, namely the external clock CK and the master system clock SYS_CK. It has been designed to guarantee a correct data transfer over these two asynchronous clock domains without any data loss or undesired meta-stability behavior.

The Reset Unit initializes all registers in the BIST wrapper. This global reset instruction can be activated directly via

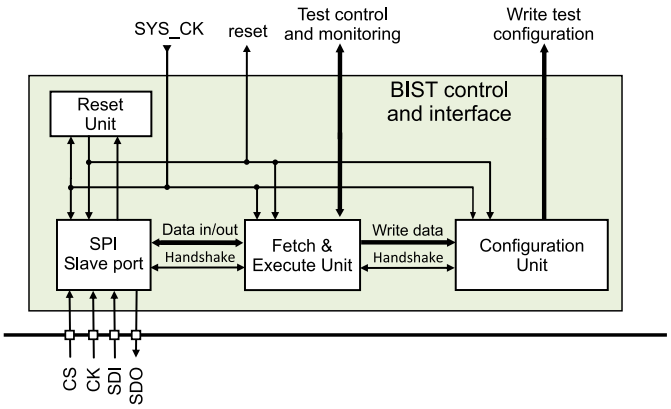



Fig. 10. Micro-architecture of the BIST control and SPI interface block.

the SPI Slave port Unit sending the appropriate command sequence.

The Fetch and Execute Unit has been designed to receive 8-bit command messages from the SPI Slave port Unit, decode them, and execute the corresponding action. It transmits all external configuration data coming from the SPI Slave port Unit to the Configuration Unit. Then, the Configuration Unit loads these data to the appropriate register in the BIST wrapper. The Fetch and Execute Unit also receives the test result data and sends them to the SPI Slave port Unit to be read out through the SPI interface. In addition, it acts as the global test processor of the BIST wrapper and controls the execution of the self-test procedure.

A complete test sequence starts by loading a test configuration through the SPI bus (e.g. bitstream in the circular shift-register and reference values $A_{ref}$ and *Threshold*) and defining the channel under test (i. e. Left or Right). Then, upon receiving the "execute test" command, the BIST control block interrupts the normal operation of the ADCUT by enabling the test mode and the ternary test stimulus is injected into the input of the selected channel. The first 20 periods of the test response are discarded to avoid capturing transient effects, then the following 128 periods are processed by the test response analyzer which compares the SNDR of the test response signal to the preloaded specification value, as detailed previously. Once the self-test has been completed, the BIST control restores normal operation and stores the test results, which can be thereafter read out via the SPI bus upon receiving the appropriate commands. The whole self-test procedure takes 28 ms per ADC channel.

## IV. Experimental Results

We have redesigned a 40nm CMOS 18-bit stereo audio ΣΔ ADC IP provided by ST Microelectronics to include the complete BIST wrapper. In total, 56 prototype chips were fabricated. Fig. 11 shows a microphotograph of one of the fabricated chips. The BIST wrapper incurs an area overhead that is only 7.1% of the area of the ΣΔ ADC. This area overhead is dominated by the test stimulus generator, test response analyzer, and BIST control blocks, which are fully digital blocks. The area of the modulators has not been

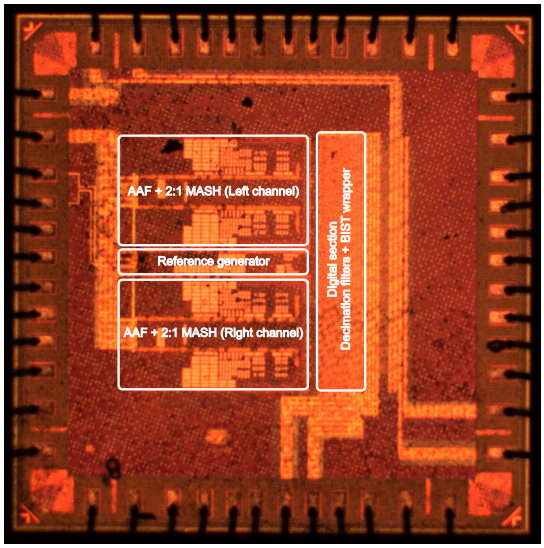



Fig. 11. Microphotograph and floorplanning of the fabricated ΣΔ ADC IP with BIST.

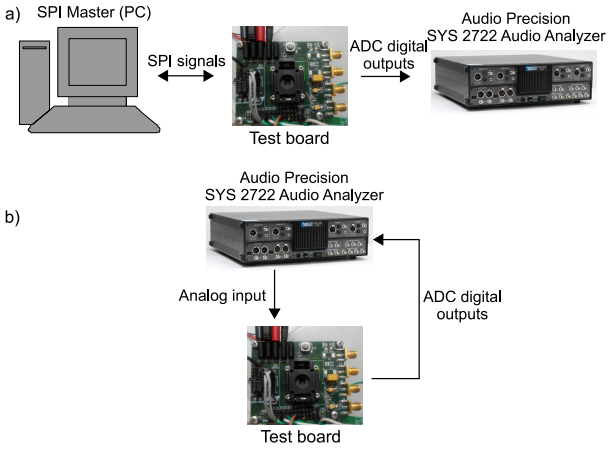



Fig. 12. a) Test setup for injection of digital ternary test stimulus in BIST mode; b) Functional specification test setup.

modified since the extra switches for injecting the ternary test stimulus have been inserted within the existing layout of the modulators. Furthermore, no power consumption overhead is observed during test mode since power consumption is dominated by the analog section of the ΣΔ ADC which is not modified.

In order to validate the feasibility of the proposed BIST strategy and the functionality of the test instruments within the BIST wrapper, we set up a batch of experiments to compare the BIST to the standard functional specification test approach. The BIST is based on a digital ternary test stimulus that encodes a high-resolution analog sine-wave, whereas the standard functional specification test approach is based on applying directly a high-resolution analog sine-wave at the input of the ADCUT.

Our first test setup, conceptually depicted in Fig. 12, aims at showing the equivalence between the proposed digital ternary test stimulus and the standard functional specification test. Firstly, we use the configuration of Fig. 12(a) where

we employ a conventional workstation as SPI master device for configuring and executing the BIST program and an Audio Precision Audio Analyzer for computing the SNDR. The functional specification test uses the test configuration of Fig. 12(b) where an Audio Precision Audio Analyzer is employed both for generating the high-resolution analog sine-wave test stimulus and for computing the SNDR. The procedure is repeated considering seven different input amplitudes, i.e. $-59.1$ dBFS, $-31.7$ dBFS, $-14.6$ dBFS, $-10.6$ dBFS, $-8.6$ dBFS, $-6.4$ dBFS, and $-3.0$ dBFS, and for the 2 channels of all 56 fabricated chips (i.e. 112 ADCUTs in total).

The SNDR values obtained using the digital ternary test stimulus and the standard functional specification test, as well as their comparison, are shown in Fig. 13. In the x-axis, we show the amplitude of the signal at the input of the modulator, while in the y-axis we show the measured SNDR values obtained by processing directly the ADCUT output, i.e. without employing the test response analyzer block. Fig. 13(a) shows the SNDR measured using the functional specification test, Fig. 13(b) shows the SNDR measured using the digital ternary test stimulus, and Fig. 13(c) shows a superposition of both sets of curves in Fig. 13(a) and Fig. 13(b) for a direct visual comparison.

By comparing Fig. 13(a) and Fig. 13(b), for the standard functional specification test set up, we observe that the amplitude of the signal at the input of the modulator is consistently attenuated. This attenuation is clearly visible by looking at the position of the data points on the x-axis. As it can be observed, the circle markers in Fig. 13(a) are consistently 2.3 dBFS below the diamond markers in Fig. 13(b) when the markers are projected to the x-axis (i.e. modulator input amplitude) for all seven test stimuli considered in our experiment. This constant attenuation of the input amplitude is due to the anti-aliasing filter that is placed before the modulator. It is not apparent when using the digital ternary test stimulus generator because the anti-aliasing filter is disconnected from the signal path and the ternary test stimulus is fed directly into the modulator. In Fig. 13(b), the higher spread of the SNDR measurements that appears for input amplitudes close to the FS can be easily explained by the fact that these input amplitudes are in the vicinity of the stability limit of the modulator. In contrast, in Fig. 13(a), the SNDR measurements obtained by the functional specification test do not show a similar spread because the input amplitude is attenuated by the anti-aliasing filter and the signal amplitude at the input of the modulator is less than the stability limit. The knee that is apparent in the SNDR measurements in Fig. 13(a) is due to the additional distortion introduced by the anti-aliasing filter that essentially limits the SNDR of the ADC in normal operation. From Fig. 13(c) it can be seen that the SNDR measurements obtained using the digital ternary test stimulus track very well, along the complete input amplitude range, the SNDR measurements obtained with the standard functional specification test.

Fig. 14 shows the histogram of the SNDR estimation error, defined as the difference between the SNDR obtained by the standard functional specification test and by using the digital ternary test stimulus. The histogram is produced by considering all fabricated chips and by measuring the SNDR

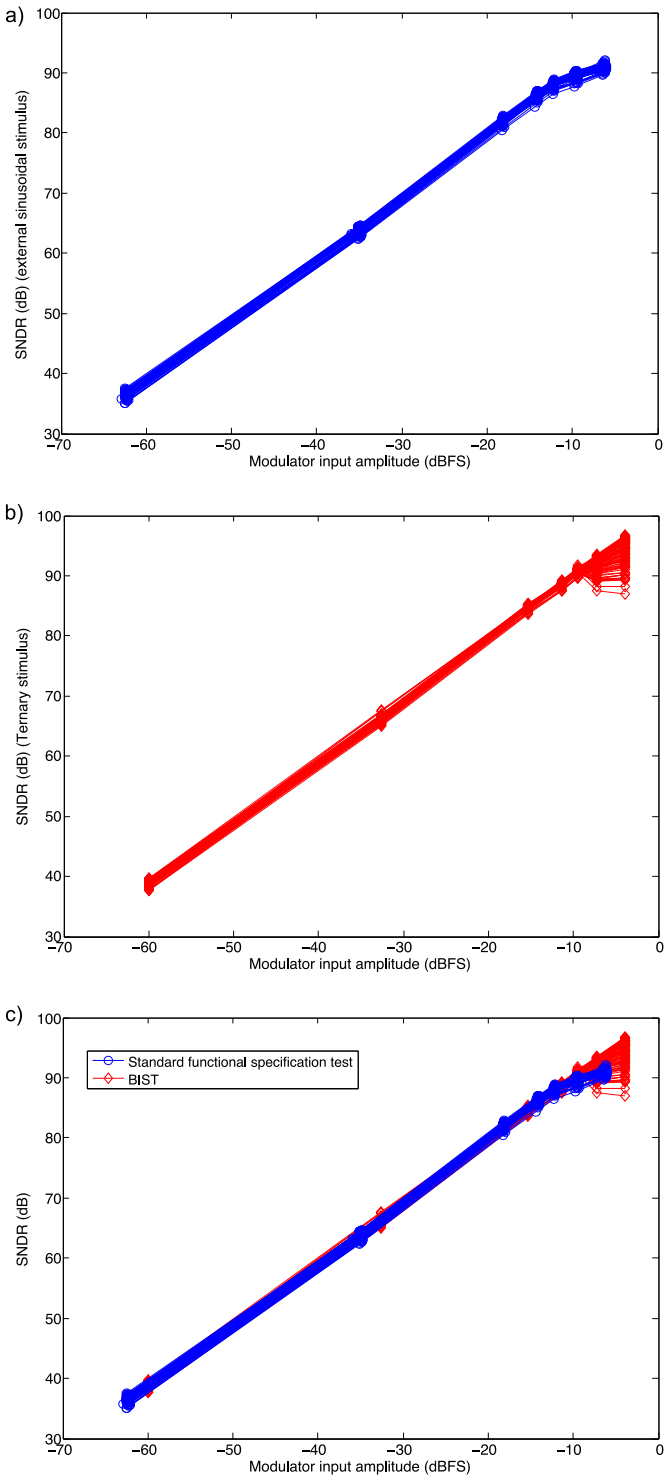



Fig. 13. SNDR measured by processing directly the ADCUT output versus the signal amplitude at the input of the modulator: (a) standard functional specification test and (b) digital ternary test stimulus. Each curve corresponds to the SNDR of one channel of a fabricated chip (i. e. in total $2 \times 56 = 112$ curves). The curves in (a) and (b) are superimposed in (c) for the purpose of comparison.

in both ways at different input amplitudes, i.e. $-59.1$ dBFS, $-31.7$ dBFS, $-14.6$ dBFS, $-10.6$ dBFS, and $-8.6$ dBFS. Higher amplitudes close to the FS are excluded since, as explained above, they give rise to other phenomena (e.g.

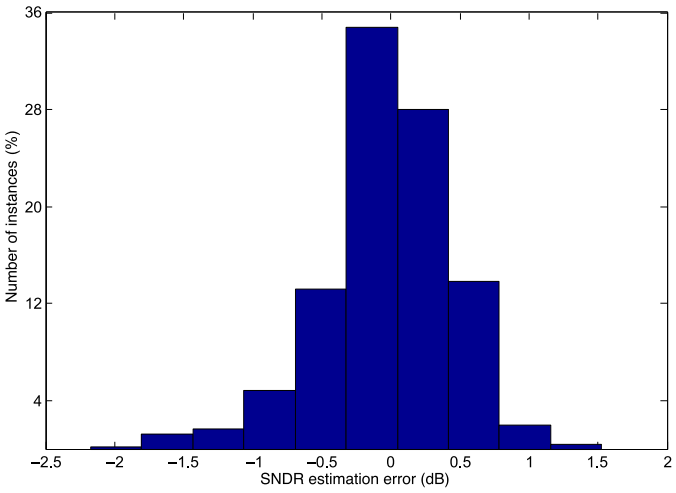



Fig. 14. Histogram of SNDR estimation error.

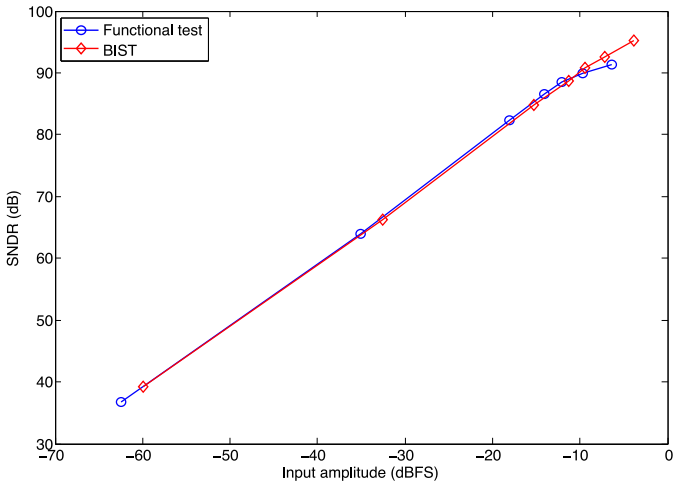



Fig. 15. SNDR measurement using the BIST (the test response analyzer is implemented in software) and the standard functional specification test as a function of the signal amplitude at the input of the modulator.

instability of modulators, distortion introduced by the anti-aliasing filter) which are not related to the BIST precision. A good indicator of the average precision of the SNDR measurement using the digital ternary test stimulus is the standard deviation $\sigma$ of the SNDR error distribution. From Fig. 14, we calculate $3\sigma = 1.4$ dB, which is essentially equivalent to the precision of the standard functional specification test performed on dedicated ATE. Therefore, the results demonstrate the equivalence between the proposed digital ternary test stimulus and the functional specification test that employs a high-precision analog sine-wave test stimulus.

Regarding the test response analyzer, upon first silicon inspection, we found a routing error of the digital input signal FSYNC_IN that had passed undetected during the pre-silicon verification process. This error affects only the functionality of the test response analyzer block. Nevertheless, since this is a fully-digital block and the output signals of the ADCUT are accessible from primary pins, we were able to replicate the functionality of this block in software to validate the proposed signal processing algorithm. We used the test setup in Fig. 12(a) with the exception that the digital output of the ADC is now routed back to the workstation and fed to a software implementation of the test response analyzer block. The SNDR computed by the test response analyzer block is compared to the SNDR measurement obtained by the functional specification test across different input amplitudes for one random chip, as shown in Fig. 15. As it can be seen, the two measurements track each other very well across the complete input amplitude range. As before, for input amplitudes close to FS, the differences are explained by the anti-aliasing filter that is disconnected during the BIST execution.

## V. Conclusions

We have presented a 40nm CMOS 18-bit stereo audio $\Sigma\Delta$ ADC that features a fully-digital BIST wrapper for measuring the SNDR on-chip. The proposed BIST is based on injecting a digital ternary test stimulus that encodes a high-precision analog sine-wave directly into the input of the modulator. The test response analysis is also carried out on-chip based on a simplified sine-wave fitting algorithm. The BIST includes a dedicated control block based on a standard SPI bus for switching to test mode, monitoring the test execution, and reading out the test results. The estimated total area overhead of the BIST is 7.1% and the total test time per channel is 28 ms. Experimental results on 56 fabricated chips prove the feasibility and effectiveness of the proposed BIST approach. Results show an excellent match across the complete dynamic range between SNDR measurements obtained using BIST and SNDR measurements obtained using standard functional specification test on precision equipment.

Future work will be aimed at extending the functionality of the BIST circuitry for enhancing the performance and reliability of the $\Sigma\Delta$ ADC. One clear research path in this line is the calibration of the $\Sigma\Delta$ ADC by taking advantage of the reference digital channel in the BIST wrapper to provide a "golden reference" signal.

## References

[1] M. J. Barragan, G. Huertas, A. Rueda, and J. L. Huertas, "(Some) Open Problems to Incorporate BIST in Complex Heterogeneous Integrated Systems," in *Electronic Design, Test and Application, 2010. DELTA '10. Fifth IEEE International Symposium on*, 2010, pp. 8–13.

[2] F. Poehl, F. Demmerle, J. Alt, and H. Obermeir, "Production test challenges for highly integrated mobile phone SOCs – A case study," in *Test Symposium (ETS), 2010 15th IEEE European*, 2010, pp. 17–22.

[3] G. Huertas, D. Vázquez, E. Peralias, A. Rueda, and J. L. Huertas, "Oscillation-based test in oversampled $\Sigma\Delta$ modulators," *Microelectronics journal*, vol. 33, no. 10, pp. 799–806 2002.

[4] J.-L. Huang and K.-T. Cheng, "Testing and characterization of the one-bit first-order delta-sigma modulator for on-chip analog signal analysis," in *Test Conference, 2000. Proceedings. International*, 2000, pp. 1021–1030.

[5] G. Leger and A. Rueda, "Digital test for the extraction of integrator leakage in first- and second-order $\Sigma\Delta$ modulators," *Circuits, Devices and Systems, IEE Proceedings*, vol. 151, no. 4, pp. 349–358, 2004.

[6] ——, "Low-Cost Digital Detection of Parametric Faults in Cascaded $\Sigma\Delta$ Modulators," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 56, no. 7, pp. 1326–1338, 2009.

[7] B. Dufort and G. W. Roberts, "On-chip analog signal generation for mixed-signal built-in self-test," *Solid-State Circuits, IEEE Journal of*, vol. 34, no. 3, pp. 318–330, 1999.

[8] H.-C. Hong, S.-C. Liang, and H.-C. Song, "A Cost Effective BIST Second-Order ΣΔ Modulator," in *Design and Diagnostics of Electronic Circuits and Systems, 2008. DDECS 2008. 11th IEEE Workshop on*, 2008, pp. 1–6.

[9] H. Mattes, S. Sattler, and C. Dworski, "Controlled sine wave fitting for ADC test," in *Test Conference, 2004. Proceedings. ITC 2004. International*, 2004, pp. 963–971.

[10] S. Mir, L. Rolindez, C. Domigues, and L. Rufer, "An implementation of memory-based on-chip analogue test signal generation," in *Design Automation Conference, 2003. Proceedings of the ASP-DAC 2003. Asia and South Pacific*, 2003, pp. 663–668.

[11] C.-K. Ong, K.-T. Cheng, and L. C. Wang, "A new sigma-delta modulator architecture for testing using digital stimulus," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 51, no. 1, pp. 206–213, 2004.

[12] L. Rolindez, S. Mir, J. L. Carbonero, D. Goguet, and N. Chouba, "A stereo audio ΣΔ ADC architecture with embedded SNDR self-test," in *Test Conference, 2007. ITC 2007. IEEE International*, 2007, pp. 1–10.

[13] H.-C. Hong, "A Design-for-Digital-Testability Circuit Structure for ΣΔ Modulators," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 15, no. 12, pp. 1341–1350, 2007.

[14] R. Schreier and G. C. Temes, *Understanding delta-sigma data converters*. IEEE press Piscataway, NJ, 2005, vol. 74.

[15] H.-C. Hong and S.-C. Liang, "A Decorrelating Design-for-Digital-Testability Scheme for ΣΔ Modulators," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 56, no. 1, pp. 60–73, 2009.

[16] S.-C. Liang and H.-C. Hong, "A Digitally Testable ΣΔ Modulator Using the Decorrelating Design-for-Digital-Testability," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 3, pp. 503–507, 2011.

[17] M. Dubois, H. G. Stratigopoulos, and S. Mir, "Ternary stimulus for fully digital dynamic testing of SC ΣΔ ADCs," in *Mixed-Signals, Sensors and Systems Test Workshop (IMS3TW), 2012 18th International*, 2012, pp. 5–10.

[18] M. Dubois, H. G. Stratigopoulos, S. Mir, and M. J. Barragan, "Evaluation of digital ternary stimuli for dynamic test of ΣΔ ADCs," in *Very Large Scale Integration (VLSI-SoC), 2014 22nd International Conference on*, 2014, pp. 1–6.

[19] H. G. Stratigopoulos, M. J. Barragan, S. Mir, H. Le Gall, N. Bhargava, and A. Bal, "Evaluation of low-cost mixed-signal test techniques for circuits with long simulation times," in *Test Conference (ITC), 2015 IEEE International*, 2015, pp. 1–7.

[20] "IEEE standard for terminology and test methods for analog-to-digital converters," *IEEE Std 1241-2010 (Revision of IEEE Std 1241-2000)*, pp. 1–139, 2011.

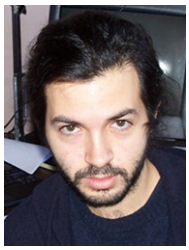
[21] "IEEE Standard for Digitizing Waveform Recorders," *IEEE Std 1057-1994*, 1994.

[22] M. De Bock and X. Xing and L. Weyten and G. Gielen and P. Rombouts, "Calibration of DAC Mismatch Errors in ΣΔ ADCs Based on a Sine-Wave Measurement," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 60, no. 9, pp. 567–571, 2013.

[23] P. Handel, "Properties of the IEEE-STD-1057 four-parameter sine wave fit algorithm," *IEEE Transactions on Instrumentation and Measurement*, vol. 49, no. 6, pp. 1189–1193, 2000.

**Manuel J. Barragan** (M'14) received a M.Sc. degree in Physics in 2003 and a PhD in Microelectronics in 2009, both from the University of Seville, Spain. He is currently a researcher with the French National Research Council (CNRS) at TIMA Laboratory, France. His research is focused on the topics of test and design for testability of analog, mixed-signal, and RF systems. He served as the General Chair of the 2014 Statistical Test Methods Workshop, as the Program Chair of the 2015 IEEE International Workshop on Test and Validation of High Speed Analog Circuits, as the Program Chair of the 2016 IEEE International Mixed-Signal Testing Workshop, and as Track Chair of the Test and Reliability track of the 2014 IEEE International Conference on Electronics Circuits and Systems. He currently serves in the Technical Program Committee of the Design, Automation, and Test in Europe (DATE) Conference and as a Review Editor for Wiley International Journal of Circuit Theory and Applications. His PhD research won a Silver Leaf Award at the IEEE PRIME conference in 2009 and in 2011 his work was selected for inclusion in the 20th Anniversary Compendium of Most Influential Papers from the IEEE Asian Test Symposium. He received the Best Special Session Award in the 2015 IEEE VLSI Test Symposium.

**Rshdee Alhakim** received the Diploma in electronic engineering from the University of Damascus, Syria, in 2004, the M.Sc. degree in communications engineering technology from the University of Brest, France, in 2008, and the Ph.D. degree in telecommunications engineering and signal processing from the University of Grenoble, France, in 2013. Currently he holds a post-doc position at TIMA Laboratory (France). His research interests are digital signal processing and image processing.



**Haralampos-G. Stratigopoulos** (S'02-M'07) received the Diploma in electrical and computer engineering from the National Technical University of Athens, Greece, in 2001 and the Ph.D. in electrical engineering from Yale University, USA, in 2006. He is currently a Researcher with the French National Center for Scientific Research (CNRS) at Laboratoire dÕInformatique de Paris 6 (LIP6), Paris, France. His research interests are in the areas of design, test, and computer-aided design tools for analog, mixed-signal, and RF circuits and systems. He served as the Program Chair of the 2010 IEEE International Workshop on Test and Validation of High Speed Analog Circuits and the 2011 IEEE International Mixed-Signals, Sensors, and Systems Test Workshop and as the General Chair of the 2015 IEEE International Mixed-Signals Test Workshop. He has served on the Technical Program Committees of Design, Automation, and Test in Europe Conference (DATE), IEEE International Conference on Computer-Aided Design (ICCAD), IEEE VLSI Test Symposium (VTS), IEEE European Test Symposium (ETS), IEEE International Test Conference (ITC) and several others international conferences. He is an Associate Editor of Springer Journal of Electronic Testing: Theory & Applications, IEEE Design & Test Magazine, and IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. He received the Best Paper Award in the 2009, 2012, and 2015 IEEE European Test Symposium.

**Matthieu Dubois** Matthieu Dubois received the Ph.D. degree in microelectronics from the University of Grenoble Alpes, Grenoble, France, in 2011. From 2006 to 2010 he was a PhD student at TIMA laboratory, Grenoble, France, working in the Reliable Mixed-signal Systems (RMS) group. His research field concerned mainly built-in test techniques for ADCs. He is currently working on the development of ADC architectures for imaging applications in Pyxalis, Grenoble, France.

**Salvador Mir** (M'99) has an Industrial Engineering (Electrical, 1987) degree from the Polytechnic University of Catalonia, Barcelona, Spain, and M.Sc. (1989) and Ph.D. (1993) degrees in Computer Science from the University of Manchester, UK. He is a Research Director of CNRS (Centre National de la Recherche Scientifique) at TIMA Laboratory in Grenoble, France. He is currently director of TIMA and member of the RMS (Reliable Mixed-signal Systems) Group. His research interests are in the field of mixed-signal/RF/MEMS test.

**Hervé Le-Gall** started his career on silicon test working with space products in 1982. For many years he has been in charge of managing field experiments in test rooms worldwide. Today he uses this knowledge to develop novel DfT solutions and test methodologies. He managed the High-Speed Interface Test and Engineering teams in ST's Digital Group. Hervé participated in TOETS and ELESIS European projects. His research interests include advanced testing methodologies for ADCs, jitter measurement, and high-speed links. He has published several patents on these topics.

**Neha Bhargava** received Bachelor in Electronics and Telecommunication Engineering from National Institute of Technology Bhopal, India, in 2002 and M.E. in Microelectronics from Indian Institute of Science, Bangalore, India, in 2004. She has since been working with STMicroelectronics, Greater Noida, India on data converters and signal processing solutions.

**Ankur Bal** received the B.E. degree in electrical engineering from Delhi College of Engineering, University of Delhi, Delhi, India, in 1999. He joined STMicroelectronics in 1999 as a Design Engineer. Since December 1999, he has been working at ST Microelectronics on various aspects of VLSI design and Digital Signal Processing. His current interests include high precision sigma delta data converters, high speed Nyquist data converters and Clock generators.

Journal of Marine Science and Technology

Volume 30 | Issue 3 Article 4

# A Hybrid Finite Difference Method for Singularly Perturbed Delay Partial Differential Equations with Discontinuous Coefficient and Source

Nitika Sharma
*Department of Applied Mathematics, Delhi Technological University, Delhi -110042 India*, nitika_phd2k18@dtu.ac.in

Aditya Kaushik
*Department of Applied Mathematics, Delhi Technological University, Delhi -110042 India*, akaushik@dtu.ac.in

Follow this and additional works at: https://jmstt.ntou.edu.tw/journal

Part of the Fresh Water Studies Commons, Marine Biology Commons, Ocean Engineering Commons, Oceanography Commons, and the Other Oceanography and Atmospheric Sciences and Meteorology Commons

## Recommended Citation

Sharma, Nitika and Kaushik, Aditya (2022) "A Hybrid Finite Difference Method for Singularly Perturbed Delay Partial Differential Equations with Discontinuous Coefficient and Source," *Journal of Marine Science and Technology*: Vol. 30: Iss. 3, Article 4.
DOI: 10.51400/2709-6998.2578
Available at: https://jmstt.ntou.edu.tw/journal/vol30/iss3/4

This Research Article is brought to you for free and open access by Journal of Marine Science and Technology. It has been accepted for inclusion in Journal of Marine Science and Technology by an authorized editor of Journal of Marine Science and Technology.

RESEARCH ARTICLE

# A Hybrid Finite Difference Method for Singularly Perturbed Delay Partial Differential Equations with Discontinuous Convection Coefficient and Source Term

Nitika Sharma [a], Aditya Kaushik [a,b,*]

[a] Department of Applied Mathematics, Delhi Technological University, Delhi 110042 India
[b] University Institute of Engineering and Technology, Panjab University, Chandigarh 160014 India

**Abstract**

**The article presents a hybrid finite difference scheme to solve a singularly perturbed parabolic functional differential equation with discontinuous coefficient and source. The simultaneous presence of deviating argument with a discontinuous source and coefficient makes the problem stiff. The solution of the problem exhibits turning point behaviour across discontinuity as $\varepsilon$ tends to zero. The hybrid scheme presented is a composition of a central difference scheme and a midpoint upwind scheme on a specially generated mesh. At the same time, an implicit finite difference method is used to discretize the time variable. Consistency, stability, and convergence of the presented numerical approach have been investigated. The presented method converges uniformly independent of the perturbation parameter. Numerical results have been presented for two test examples that verify the effectiveness of the scheme.**

*Keywords:* Singular perturbation, Functional differential equation, Finite difference method, Interior layer

## 1. Introduction

Differential equations model a wide range of phenomena in almost every branch of sciences and engineering. Often mathematical models assume specific behaviour or phenomenon which depends on the present and the past states of a system [17,42]. In other words, previous events have a direct impact on future outcomes. Modelling such systems leads to functional differential equations that are more realistic and frequently appear in a wide range of applications. Classical examples cover oscillatory, excitable as well as chaotic behavior [1] in physiological control systems [29], nonlinear optics [19], population dynamics [16] and neuroscience [31]. Particularly in nonlinear optics, the finite-time communication delays are typically much larger than the device's internal time-scales [39], and therefore, give rise to rich dynamical phenomena [27]. Specific examples of delay systems with (multiple) large delays include semiconductor lasers with two optical feedback loops of different lengths [30], ring-cavity lasers with optical feedback [12], and others [10,37].

When we associate a mathematical model with physical phenomena, we often capture the essentials and ignore the minor components involving small parameters. A mathematical model with small parameters is called a perturbed model, while an unperturbed or reduced model is a degenerate simplified model. If a small parameter is multiplied with the highest-order derivative term, then the problem is called singularly perturbed from a mathematical perspective [4]. In that case, the

Received 6 January 2022; revised 2 March 2022; accepted 19 April 2022.
Available online 12 August 2022

* Corresponding author.
E-mail address: akaushik@dtu.ac.in (A. Kaushik).

https://doi.org/10.514/2709-6998.2578
2709-6998/© 2022 National Taiwan Ocean University.

corresponding degenerate problem cannot satisfy the prescribed boundary/initial conditions. Moreover, the stiffness, attributed to the simultaneous presence of a deviating argument and discontinuous coefficient, give rise to multiple time scales as $\varepsilon$ approaches the limiting value zero. There exist narrow regions across the turning points where the solution varies exponentially and approaches a discontinuous limit [28,36]. In these regions, the solution reveals sharp interior layers, and it is not easy to find an asymptotic expansion in terms of $\varepsilon$ or to find a consistent numerical approximation.

The numerical methods on a uniform mesh fail to approximate singular perturbation functional differential equations. They require an unacceptably large number of mesh points to sustain the approximation because the mesh width depends on the perturbation parameter. This limitation of the conventional numerical methods has encouraged researchers to develop robust or adaptive numerical techniques that perform well enough independent of the perturbation parameter. Many researchers have made efforts to develop parameter uniform methods for partial functional differential equations; for an overview, see [5,20] and references therein. In [2], the authors studied a Dirichlet boundary value problem for the partial functional differential equation. In this paper, a numerical method comprising a standard finite difference operator (centered in space, implicit in time) is presented. The method presented is robust in the sense that its numerical solution converges to the exact solution uniformly. The higher-order methods are of particular interest and reasonably understood for problems without a deviating argument [8,23]. Nevertheless, attempts of having a higher order of uniform convergence for parabolic functional differential equations are lacking. In [25], a higher-order uniformly convergent hybrid method is proposed for parabolic problems with a delay. A scheme of HODIE type [24] is used in spatial direction over a generalized Shishkin mesh, whereas the implicit Euler scheme discretizes the time variable. Richardson extrapolation improves the order of convergence in time-variable. It is proved that the computed solution is uniformly convergent of $\mathscr{O}((L/N)^4 + (\Delta t)^2)$. In [22], a stable finite difference scheme is presented for singularly perturbed differential-difference equations with delay and advanced term. In [13], a standard finite difference method on an equidistributed layer adapted mesh is presented. The method approximates a singularly perturbed parabolic functional differential equations of convection-diffusion type using an upwind finite difference method on a piecewise uniform Shishkin mesh. In [38], a singularly perturbed differential–difference equation with small shifts is considered. The terms containing the delay and advance parameters are approximated by using the Taylor's series expansion. A quadratic B-spline collocation method is used in the space direction on exponentially graded mesh while the Crank–Nicolson finite difference method is used in the time direction on uniform mesh.

For a time-delayed convection-diffusion problem, a hybrid scheme is reported in [9]. The method presented is a composition of an upwind scheme and a central difference scheme and has almost second-order accuracy in space which is optimal compared to [14]. While a hybrid difference scheme for parabolic differential equation with delay and advance terms is studied in [15,34]. Besides, uniformly convergent methods based on fitted mesh and fitted operator approach are presented in [6]. The fitted operator methods used have coefficients of exponential type and are quite capable of reproducing the multiscale character of the exact solution of the problem [36]. On the other hand, the fitted mesh method [32,41] adopted concentrates on the appropriate choice of the grid points in narrow layer regions. Moreover, a Robin type boundary value problem for parabolic functional differential equation is brought to attention in [3]. A numerical method comprising a standard finite difference scheme on a rectangular piecewise uniform fitted mesh is presented. It is shown that the errors are bounded in the maximum norm by $C(N_x^{-2}\ln^2 N_x + N_t^{-1})$. In [21], the authors considered a class of singularly perturbed parabolic problem with delay and discontinuous data. An implicit numerical scheme is used to find the numerical solution. The method is uniformly convergent of $\mathscr{O}(N^{-1}\ln^2 N + \Delta t)$. However, the analysis of higher-order methods for singularly perturbed parabolic partial functional differential equations with discontinuous data and degenerating convective terms has seen little development and lacks due attention. Consequently, this paper aims to extend the idea of parameter uniform difference schemes to solve a singularly perturbed parabolic functional differential equation with discontinuity in coefficients with turning points and presents a higher-order accurate approximation.

## 2. The continuous problem

Consider the following initial-boundary-value problem on the domain $S^- \cup S^+ = (0,1) \times (0,T] \cup (1,2) \times (0,T]$:

$$\left.\begin{aligned} &L_\varepsilon u = \varepsilon u_{xx}(x,t) + a(x)u_x(x,t) - b(x)u(x,t) - u_t(x,t) \\ &\quad = f(x,t) + c(x)u(x-1,t) \text{ in } S^- \cup S^+, \\ &u(x,t) = p_0(x) \text{ on } [0,2] \times \{t=0\}, \\ &u(x,t) = p_1(x,t) \text{ in } [-1,0] \times [0,T], \\ &u(x,t) = p_2(t) \text{ on } \{x=2\} \times [0,T], \end{aligned}\right\} \tag{2.1}$$

where $\varepsilon \ll 1$ is a small positive parameter, $b$ and $c$ are sufficiently smooth functions such that $c(x) > 0$, $b(x) > 0$ for all $x \in [0, 2]$. Moreover, we assume that

$$\left.\begin{aligned} &a(x) = \begin{cases} a_1(x) \text{ if } 0 \le x \le 1 \\ a_2(x) \text{ if } 1 < x \le 2, \end{cases} f(x,t) = \begin{cases} f_1(x,t) \text{ if } (x,t) \in \overline{S}^- \\ f_2(x,t) \text{ if } (x,t) \in \overline{S}^+, \end{cases} \\ &-\gamma_1^* < a_1(x) < -\gamma_1 < 0, \quad \gamma_2^* > a_2(x) > \gamma_2 > 0, \\ &|[a]| \le C, \quad |[f]| \le C, \end{aligned}\right\} \tag{2.2}$$

where $\gamma = \min\{\gamma_1, \gamma_2\}$ and $\gamma^* = \max\{\gamma_1^*, \gamma_2^*\}$. The solution of (2.1) satisfies $[u] = 0$ and $[u_x] = 0$ at $x = 1$. Here, $[u]$ denotes the jump of $u$ defined at the point of discontinuity $x = 1$ as $[u](1, t) = u(1^+, t) - u(1^-, t)$ where $u(1^\pm, t) = \lim_{x \to 1 \pm 0} u(x,t)$.

The functions $p_0$, $p_1$ and $p_2$ are Hölder continuous and the compatibility conditions hold at the corners $(0,0)$, $(2,0)$ and at the transition point $(1,0)$. On the domain $S^-$, the delay term $u(x-1,t) = p_1(x-1,t)$. Under these assumptions, the solution of (2.1) exists and unique $u \in C^{1+\lambda}(D = (0,2) \times (0,T]) \cap C^{2+\lambda}(S^- \cup S^+)$ [26,35].

The presence of a discontinuity and a delay makes the problem challenging. The solution $u(x,t)$ of (2.1) displays a strong interior layer in the neighbourhood of the point $x = 1$. Moreover, it is easy to follow that the differential operator $L_\varepsilon$ satisfies the following maximum principle on $\overline{D}$.

**Lemma 2.1** Suppose $P \in C^0(\overline{D}) \cap C^2(S^- \cup S^+)$ satisfies $P(x,t) \le 0$ for all $(x,t) \in \Gamma := \overline{D} \setminus D$ and $L_\varepsilon P(x,t) \ge 0$ for all $(x,t) \in S^- \cup S^+$. Then $P(x,t) \le 0$ for all $(x,t) \in \overline{D}$.

*Proof.* Choose $(x^k, t^k) \in D$ such that $P(x^k, t^k) = \max_{(x,t) \in \overline{D}} P(x,t)$. Consequently,

$$P_x(x^k,t^k) = 0, \quad P_t(x^k,t^k) = 0 \quad \text{and} \quad P_{xx}(x^k,t^k) < 0.$$

Suppose $P(x^k, t^k) > 0$ and it follows that $(x^k, t^k) \notin \Gamma$. If $(x^k, t^k) \in S^- \cup S^+$, note that $L_\varepsilon P(x^k, t^k) < 0$. A contradiction to the assumption and consequently the required result follows.

Establishing the boundedness of solution is an important application of the maximum principle. As an immediate application, we obtain

**Lemma 2.2** Let u be the solution of (2.1). Then

$$\| u \|_{\infty,\overline{D}} \le \| u \|_{\infty,\Gamma} + \frac{1}{\gamma} \| f \|_{\infty,\overline{D}}, \quad \gamma = \min\{\gamma_1, \gamma_2\}. \tag{2.3}$$

*Proof.* Consider

$$\psi_\pm = \begin{cases} -\| u \|_{\infty,\Gamma} - \dfrac{x}{\gamma} \| f \|_{\infty,\overline{D}} \pm u & \text{if } x \le 1 \\ -\| u \|_{\infty,\Gamma} - \dfrac{(2-x)\| f \|_{\infty,\overline{D}}}{\gamma} \pm u & \text{if } x \ge 1. \end{cases}$$

For $(x,t) \in S^-$, it follows that

$$\begin{aligned} L_\varepsilon \psi_\pm(x,t) &= \pm L_\varepsilon u - a_1(x) \frac{\| f \|}{\gamma} + b(x) \| u \| \\ &\quad + b(x) \frac{x \| f \|}{\gamma} \ge 0 \end{aligned}$$

since $a_1(x) \le 0$ and $b(x) > 0$. Similarly, for $(x,t) \in S^+$, it is easy to verify that $L_\varepsilon \psi_\pm(x,t) \ge 0$. The required result (2.3) now follows from Lemma 2.1.

In general, one can assume homogeneous boundary conditions $p_0 = p_1 = p_2 = 0$ by subtracting from $u$ some suitable smooth function that satisfies the original boundary conditions [36]. To find sharper bounds on the solution and its derivatives, we decompose the solution $u$ into smooth and singular components. We write $u = v + w$. The smooth component $v$ is solution of

$$\left.\begin{aligned} &L_\varepsilon v(x,t) = f(x,t) + c(x)v(x-1,t) \text{ in } S^- \cup S^+, \\ &v(x,t) = u(x,t) \text{ in } [-1,0] \times [0,T], \\ &v(1^-,t) = j_1(t), \ v(1^+,t) = j_2(t), \ t \in (0,T], \\ &v(x,t) = u(x,t) \text{ on } \{x=2\} \times (0,T], \\ &v(x,t) = u(x,t) \text{ on } [0,2] \times \{t=0\}, \end{aligned}\right\} \tag{2.4}$$

where the functions $j_1(t)$ and $j_2(t)$ will be computed using the Theorem 2.4. The singular component $w$ is solution of

$$\left.\begin{aligned} &L_\varepsilon w(x,t) = c(x)w(x-1,t) \text{ in } S^- \cup S^+, \\ &w(x,t) = 0 \text{ in } [-1,0] \times [0,T], \\ &w(x,t) = 0 \text{ on } \{x=2\} \times (0,T], \\ &w(x,t) = 0 \text{ on } [0,2] \times \{t=0\}, \\ &[w]\left(1,t\right) = -[v]\left(1,t\right), \\ &\left[\frac{\partial w}{\partial x}\right]\left(1,t\right) = -\left[\frac{\partial v}{\partial x}\right]\left(1,t\right), t \in \left(0,T\right] \end{aligned}\right\} \tag{2.5}$$

Following are the direct results from [18] that will be used in Theorem 2.4.

**Theorem 2.3** Let $a, b, c \in C^2[0,2]$, $f \in C^{2+\lambda}(\overline{D})$ and $a(x) \geq \gamma > 0$ for $x \in [0,2]$. Also suppose $p_0, p_1$ and $p_2$ are identically zero so that the compatibility conditions hold in $\Gamma_c = (-1,0) \cup (0,0) \cup (2,0) \cup (1,0)$ and

$$\frac{\partial^{k+m} F}{\partial x^k \partial t^m} = 0, \quad 0 \leq k + 2m \leq 2,$$

where $F(x,t) = f(x,t) + c(x)u(x-1,t)$. Then $u \in C^{4+\lambda}(\overline{D})$ and

$$\left\| \frac{\partial^{k+m} u}{\partial x^k \partial t^m} \right\|_{\infty, D} \leq C\varepsilon^{-k}, \quad 0 \leq k + 2m \leq 4. \tag{2.6}$$

**Theorem 2.4** There exist smooth functions $j_1(t)$, $j_2(t)$ such that the smooth and singular component defined in (2.4) and (2.5), respectively, satisfy the following bounds for $0 \leq k \leq 3$, $m \geq 0$ and $0 \leq k + 2m \leq 4$

$$\left\| \frac{\partial^{k+m} v}{\partial x^k \partial t^m} \right\|_{\infty, S^- \cup S^+} \leq C, \quad \left\| \frac{\partial^4 v}{\partial x^4} \right\|_{\infty, S^- \cup S^+} \leq C\varepsilon^{-1}$$

and

$$\left| \frac{\partial^{k+m} w}{\partial x^k \partial t^m}(x,t) \right| \leq \begin{cases} C\left(\varepsilon^{-k} \exp(-(1-x)\gamma_1/\varepsilon)\right) & \text{for } (x,t) \in S^- \\ C\left(\varepsilon^{-k} \exp(-(x-1)\gamma_2/\varepsilon)\right) & \text{for } (x,t) \in S^+, \end{cases}$$

$$\left| \frac{\partial^4 w}{\partial x^4}(x,t) \right| \leq \begin{cases} C\left(\varepsilon^{-4} \exp(-(1-x)\gamma_1/\varepsilon)\right) & \text{for } (x,t) \in S^- \\ C\left(\varepsilon^{-4} \exp(-(x-1)\gamma_2/\varepsilon)\right) & \text{for } (x,t) \in S^+. \end{cases}$$

*Proof.* We conduct our analysis separately in the subregions $S^-$ and $S^+$ to obtain stronger bounds on the solution and its derivatives. We start with the subregion $S^+$ and define $\overline{D}^*$ such that $\overline{S}^+ \subset \overline{D}^*$. Let $D^* = \Omega^* \times (0,T]$, $\Omega^* = [-2,2]$ and define a function $v^*$ on $D^*$ as

$$v^* = v_0^* + \varepsilon v_1^* + \varepsilon^2 v_2^* + \varepsilon^3 v_3^*,$$

where the functions $v_0^*$, $v_1^*$ and $v_2^*$ are solutions of

$$\left.\begin{aligned} & a^* \frac{\partial v_0^*}{\partial x} - b^* v_0^* - \frac{\partial v_0^*}{\partial t} = f^* + c^* v_0^*(x-1,t) \quad \text{in} \quad D^*, \\ & v_0^*(x,t) = p_1^*(x,t) \quad \text{in} \quad [-3,0] \times [0,T], \\ & v_0^*(x,t) = p_0^*(x) \quad \text{on} \quad [-2,2] \times \{t=0\}, \\ & a^* \frac{\partial v_i^*}{\partial x} - b^* v_i^* - \frac{\partial v_i^*}{\partial t} = -\frac{\partial^2 v_{i-1}^*}{\partial x^2} + c^* v_i^*(x-1,t) \quad \text{in} \quad D^*, \\ & v_i^*(x,t) = 0 \quad \text{in} \quad [-3,0] \times (0,T], v_i^*(x,0) = 0 \quad \text{on} \quad [-2,2] \times \{t=0\}, \quad i = 1,2 \end{aligned}\right\} \tag{2.7}$$

and $v_3^*$ is solution of

$$\left.\begin{aligned} & \varepsilon \frac{\partial^2 v_3^*}{\partial x^2} + a^* \frac{\partial v_3^*}{\partial x} - b^* v_3^* - \frac{\partial v_3^*}{\partial t} = c^* v_3^*(x-1,t) - \frac{\partial^2 v_2^*}{\partial x^2} \text{ in } D^*, \\ & v_3^*(x,t) = 0 \text{ in } [-3,0] \times [0,T], \\ & v_3^*(x,0) = 0 \text{ on } [-2,2] \times \{t=0\}, \\ & v_3^*(x,t) = 0 \text{ on } \{x=2\} \times (0,T] \end{aligned}\right\} \tag{2.8}$$

Here, the coefficients $a^*$, $b^*$ and $c^*$ as well as the condition $p_0^*$ are respective smooth extensions of $a$, $b$, $c$ and $p_0$ from the domain $[1,2]$ to the domain $[-2,2]$. The functions $f^*$ and $p_1^*$ are the smooth extensions of $f$ and $p_1$ from the domain $\overline{S}^+$ to the domain $\overline{D}^*$. In a neighbourhood of the point $(-2,0)$, the functions $p_0^*$, $p_1^*$ and $f^*$ are built such that $p_0^* = p_1^* = f^* = 0$. Assuming that $a^*$, $b^*$, $c^*$ and $f^*$ are sufficiently smooth on $\overline{D}^*$. We set all the initial-boundary data associated with (2.7) equal to zero. Define $F^* = f^* + c^* v^*(x-1,t)$ and impose the following compatibility condition on the set $\Gamma_c$:

$$\frac{\partial^{k+m} F^*}{\partial x^k \partial t^m} = 0 \quad \text{for} \quad 0 \leq k + m \leq 7. \tag{2.9}$$

Follows from the results of [7] for the first order differential equations defined in (2.7), we have $v_i^* \in C^{9-2i+\lambda}(\overline{D}^*) \cap C^{8-2i}(\overline{D}^*)$, $i = 0,1,2$ which implies $\frac{\partial^2 v_2^*}{\partial x^2} \in C^{2+\lambda}(\overline{D}^*)$ and therefore $v_3^* \in C^{4+\lambda}(\overline{D}^*)$. Then it follows that

$$\left\| \frac{\partial^{k+m} v_i^*}{\partial x^k \partial t^m} \right\|_{\infty, D^*} \leq C, \quad i = 0,1,2,$$

$$\left\| \frac{\partial^{k+m} v_3^*}{\partial x^k \partial t^m} \right\|_{\infty, D^*} \leq C\varepsilon^{-k} \quad \text{for} \quad 0 \leq k + 2m \leq 4.$$

The smooth component $v$ is now defined as a restriction of $v^*$ on the domain $\overline{S}^+$. Define $v^*(1,t) = j_2(t) = v(1,t)$. Thus $v$ satisfies

$$L_\varepsilon v = f + c(x)v(x-1,t) \quad \text{in} \quad S^+,$$

$$v(x,t) = v^*(x,t) \quad \text{on} \quad \{x=1\}\times(0,T],$$

$$v(x,t) = u(x,t) \quad \text{on} \quad [1,2]\times\{t=0\}.$$

Since $v^* = v_0^* + \varepsilon v_1^* + \varepsilon^2 v_2^* + \varepsilon^3 v_3^*$, thus

$$\left\|\frac{\partial^{k+m}v^*}{\partial x^k\partial t^m}\right\|_{\infty,S^+} \le \left\|\frac{\partial^{k+m}v_0^*}{\partial x^k\partial t^m}\right\|_{\infty,S^+} + \varepsilon\left\|\frac{\partial^{k+m}v_1^*}{\partial x^k\partial t^m}\right\|_{\infty,S^+} + \varepsilon^2\left\|\frac{\partial^{k+m}v_2^*}{\partial x^k\partial t^m}\right\|_{\infty,S^+} + \varepsilon^3\left\|\frac{\partial^{k+m}v_3^*}{\partial x^k\partial t^m}\right\|_{\infty,S^+}$$
$$\le C\left(1+\varepsilon^{3-k}\right) \quad \text{for} \quad 0\le k+2m\le 4.$$

As $v$ is a restriction of $v^*$ on the domain $\overline{S}^+$, we have

$$\left\|\frac{\partial^{k+m}v}{\partial x^k\partial t^m}\right\|_{\infty,S^+} \le C\left(1+\varepsilon^{3-k}\right) \quad \text{for} \quad 0\le k+2m\le 4.$$

Thus for $0 \le k \le 3$ and $0 \le k+2m \le 4$, the smooth component $v$ satisfies

$$\left\|\frac{\partial^{k+m}v}{\partial x^k\partial t^m}\right\|_{\infty,S^+} \le C \quad \text{and} \quad \left\|\frac{\partial^4 v}{\partial x^4}\right\|_{\infty,S^+} \le C\varepsilon^{-1}.$$

Similarly, the desired bounds for smooth component $v$ in the subregion $S^-$ can be obtained. Define the barrier functions on $S^-\cup S^+$

$$\Phi^\pm(x,t) = \begin{cases} (\pm v - C)\exp\left(\dfrac{-(1-x)\gamma_1}{\epsilon}\right) \pm w, & (x,t)\in S^- \\ (\pm v - C)\exp\left(\dfrac{-(x-1)\gamma_2}{\epsilon}\right) \pm w, & (x,t)\in S^+, \end{cases} \tag{2.10}$$

where $C$ is a constant. For $(x,t)\in S^-$, using assumption $(a+\gamma_1)<0$ to obtain

$$L_\varepsilon\phi^\pm(x,t) = L_\varepsilon\left((\pm v - C)\exp\left(\frac{-(1-x)\gamma_1}{\varepsilon}\right)\right) \pm L_\varepsilon w(x,t) \ge 0. \tag{2.11}$$

Similarly, for $(x,t)\in S^+$ we obtain

$$L_\varepsilon\phi^\pm(x,t) = \frac{\gamma_2}{\varepsilon}\exp\left(\frac{-(x-1)\gamma_2}{\varepsilon}\right)\left((\pm v - C)(-a+\gamma_2) - 2\varepsilon v_x + \frac{Cb\varepsilon}{\gamma_2}\right)$$
$$\pm\frac{\gamma_2}{\varepsilon}\exp\left(\frac{-(x-1)\gamma_2}{\varepsilon}\right)\left(\frac{f + cv(x-1,t)}{\gamma_2}\right) \ge 0. \tag{2.12}$$

Moreover, $\phi^\pm\in C^0(\overline{D})$ and $\phi^\pm(x,t)\le 0$ for $(x,t)\in\Gamma$. Consequently, the required bounds on $w$ follows from Lemma 2.1. The required estimates for the derivatives of $w$ follow from [33].

## 3. Difference scheme

The solution of the problem exhibits strong interior layer at $x = 1$. Therefore, we descretize the domain by constructing a rectangular mesh $\overline{D}^{N,M} = \overline{S}_x^N \times T_t^M$ in such a way that it will condense around the point $x = 1$. We write

$$[0,2] = [0,1-\sigma]\cup[1-\sigma,1]\cup[1,1+\sigma]\cup[1+\sigma,2],$$

where $\sigma = \min\{\frac{1}{2}, \sigma_0\varepsilon\ln N\}$ and $\sigma_0$ is a constant that will be chosen later on. We place $\frac{N}{4}$ mesh points in each of the subintervals. Consequently, we obtain

$$S_x^N = \left\{x_i : i = 1,2\ldots,\frac{N}{2}-1\right\}\cup\left\{x_i : i = \frac{N}{2}+1,\ldots,N-1\right\}$$

and

$$h_i = x_i - x_{i-1}, \quad i = 1,2,\ldots,N,$$
$$\widehat{h}_i = h_i + h_{i+1}, \; i = 1,2,\ldots,N-1,$$

We define the uniform mesh for the domain $[0,T]$, as follows

$$T_t^M = \{t_k = k\Delta t, k = 0, \ \ldots \ , M, \Delta t = T/M\}.$$

$$h_i = \begin{cases} H = \dfrac{4(1-\sigma)}{N}, & i = 1,\ldots,N/4, 3N/4+1,\ldots,N \\ h = \dfrac{4\sigma}{N}, & i = N/4+1,\ldots,N/2, N/2+1,\ldots,3N/4. \end{cases}$$

To discretize the differential operator in (2.1), we first define the finite difference operators on the mesh $\overline{D}^{N,M}$ as

$$D_x^+ v_i^k = \frac{v_{i+1}^k - v_i^k}{h_{i+1}}, \quad D_x^- v_i^k = \frac{v_i^k - v_{i-1}^k}{h_i},$$
$$D_x^0 v_i^k = \frac{v_{i+1}^k - v_{i-1}^k}{h_{i+1}+h_i}, \; D_t^- v_i^k = \frac{v_i^k - v_i^{k-1}}{\Delta t},$$

and $\delta_x^2 v_i^k = \frac{2(D_x^+ v_i^k - D_x^- v_i^k)}{h_{i+1}+h_i}$. Also define $v_{i\mp\frac{1}{2}}^k = \frac{(v_{i\mp1}^k + v_i^k)}{2}$, $v_{i\mp\frac{1}{2}} = \frac{(v_{i\mp1}+v_i)}{2}$.

We use the classical central difference scheme in the intervals $(1-\sigma,1),(1,1+\sigma)$ and the midpoint upwind scheme in the rest of the intervals. At the point of discontinuity, second-order one-sided difference approximations are used to keep the continuity of the spatial derivative. We use the backward-Euler method for discretizing the time derivative. The discrete problem thus reads:

$$\begin{cases} U_{i,0} = p_0(x_i) & \text{for } i=0,\ldots,N \\ L_{mu}^{N,M,(L)} U_{i,k+1} = f_{\frac{i-1}{2},k+1} + c_{\frac{i-1}{2}} p_1 & \text{for } i=1,\ldots,\dfrac{N}{4} \\ L_{cen}^{N,M} U_{i,k+1} = f_{i,k+1} + c_i p_1 & \text{for } i=\dfrac{N}{4}+1,\ldots,\dfrac{N}{2}-1 \\ L_{cen}^{N,M} U_{i,k+1} = f_{i,k+1} + c_i U_{i-N/2,k+1} & \text{for } i=\dfrac{N}{2}+1,\ldots,\dfrac{3N}{4}-1 \\ L_{mu}^{N,M,(R)} U_{i,k+1} = f_{\frac{i+1}{2},k+1} + c_{\frac{i+1}{2}} U_{i-N/2,k+1} & \text{for } i=\dfrac{3N}{4},\ldots,N-1 \\ D_x^F U_{i,k+1} - D_x^B U_{i,k+1} = 0 & \text{for } i=\dfrac{N}{2} \\ \text{for } k=0,\ldots,M-1, \end{cases} \tag{3.1}$$

where

$$\begin{cases} \mathcal{L}_{mu}^{N,M,(L)} U_{i,k+1} &= \epsilon\delta_x^2 U_{i,k+1} + a_{\frac{i-1}{2}} D_x^- U_{i,k+1} - b_{\frac{i-1}{2}} U_{\frac{i-1}{2},k+1} - D_t^- U_{\frac{i-1}{2},k+1} \\ L_{cen}^{N,M} U_{i,k+1} &= \epsilon\delta_x^2 U_{i,k+1} + a_i D_x^0 U_{i,k+1} - b_i U_{i,k+1} - D_t^- U_{i,k+1} \\ \mathcal{L}_{mu}^{N,M,(R)} U_{i,k+1} &= \epsilon\delta_x^2 U_{i,k+1} + a_{\frac{i+1}{2}} D_x^+ U_{i,k+1} - b_{\frac{i+1}{2}} U_{\frac{i+1}{2},k+1} - D_t^- U_{\frac{i+1}{2},k+1} \end{cases} \tag{3.2}$$

and

$$\begin{cases} D_x^F U_{\frac{N}{2},k} = \left( -U_{\frac{N}{2}+2,k+1} + 4U_{\frac{N}{2}+1,k+1} - 3U_{\frac{N}{2},k+1} \right) \Big/ 2h \\ D_x^B U_{\frac{N}{2},k} = \left( U_{\frac{N}{2}-2,k+1} - 4U_{\frac{N}{2}-1,k+1} + 3U_{\frac{N}{2},k+1} \right) \Big/ 2h. \end{cases} \tag{3.3}$$

After simplifying the terms in (3.1), we obtain

$$\begin{cases} U_i^0 = p_0(x_i) & \text{for} \quad i = 0, \ldots, N \\ L_\varepsilon^{N,M} U_{i,k+1} = \tilde{f}_{i,k+1} & \text{for} \quad i = 1, \ldots, N-1 \\ U_{i,k+1} = p_1(x_i, t_{k+1}) & \text{for} \quad i = -N/2, \ldots, 0 \\ U_{N,k+1} = p_2(t_{k+1}) & \text{for} \quad k = 0, \ldots, M-1, \end{cases} \tag{3.4}$$

where

$$L_\epsilon^{N,M} U_{i,k+1} = \begin{cases} \left[ r_i^- U_{i-1,k+1} + r_i^0 U_{i,k+1} + r_i^+ U_{i+1,k+1} \right] + \left[ p_i^- U_{i-1,k} + p_i^0 U_{i,k} + p_i^+ U_{i+1,+k+1} \right] \\ \text{for} \quad i = 1, \ldots, N/2-1, N/2+1, \ldots, N-1 \\ q_i^{-,2} U_{i-2,k+1} + q_i^{-,1} U_{i-1,k+1} + q_i^0 U_{i,k+1} + q_i^{+,1} U_{i+1,k+1} + q_i^{+,2} U_{i+2,k+1} \\ \text{for} \quad i = N/2 \end{cases} \tag{3.5}$$

and

$$\tilde{f}_{i,k+1} = \begin{cases} m_i^- f_{i-1,k+1} + m_i^0 f_{i,k+1} + m_i^+ f_{i+1,k+1} + s_i^+ p_1(x_i, t_{k+1}) + s_i^- U_{i-N/2,k+1} \\ \text{for} \quad i = 1, \ldots, N/2-1, N/2+1, \ldots, N-1 \\ 0 \quad \text{for} \quad i = N/2. \end{cases} \tag{3.6}$$

Here, for $i \in \{1 : N/4\}$

$$\begin{cases} r_i^- = \left(\dfrac{2\epsilon}{\widehat{h}_i h_i} - \dfrac{a_{i-1/2}}{h_i} - \dfrac{b_{i-1/2}}{2} - \dfrac{1}{2\Delta t}\right), \\ r_i^0 = \left(\dfrac{-2\epsilon}{h_i h_{i+1}} + \dfrac{a_{i-1/2}}{h_i} - \dfrac{b_{i-1/2}}{2} - \dfrac{1}{2\Delta t}\right), \\ r_i^+ = \dfrac{2\epsilon}{\widehat{h}_i h_{i+1}}, \quad p_i^- = \dfrac{1}{2\Delta t}, \quad p_i^0 = \dfrac{1}{2\Delta t}, \\ p_i^+ = 0, \quad m_i^- = \dfrac{1}{2}, \quad m_i^0 = \dfrac{1}{2}, \quad m_i^+ = 0, \\ s_i^+ = c_{i-1/2}, \quad s_i^- = 0, \end{cases} \tag{3.7}$$

for $i \in \{N/4 + 1 : 3N/4 - 1\}/\{\frac{N}{2}\}$

$$\begin{cases} r_i^- = \left(\dfrac{2\varepsilon}{\widehat{h}_i h_i} - \dfrac{a_i}{\widehat{h}_i}\right), \quad r_i^0 = \left(\dfrac{-2\epsilon}{h_i h_{i+1}} - b_i - \dfrac{1}{\Delta t}\right), \\ r_i^+ = \left(\dfrac{2\epsilon}{\widehat{h}_i h_{i+1}} + \dfrac{a_i}{\widehat{h}_i}\right), \quad p_i^- = 0, \quad p_i^0 = \dfrac{1}{\Delta t}, \\ p_i^+ = 0, \quad m_i^- = 0, \quad m_i^0 = 1, \quad m_i^+ = 0, \end{cases} \tag{3.8}$$

for $i \in \{N/4 + 1 : N/2 - 1\}$

$$s_i^+ = c_i, \quad s_i^- = 0,$$

for $i \in \{N/2 + 1 : 3N/4 - 1\}$

$$s_i^+ = 0, \quad s_i^- = c_i,$$

for $i \in \{3N/4 : N - 1\}$

$$\begin{cases} r_i^- = \dfrac{2\varepsilon}{\widehat{h}_i h_i}, \quad r_i^0 = \left(\dfrac{-2\varepsilon}{h_i h_{i+1}} - \dfrac{a_{i+1/2}}{h_{i+1}} - \dfrac{b_{i+1/2}}{2} - \dfrac{1}{2\Delta t}\right), \\ r_i^+ = \left(\dfrac{2\varepsilon}{\widehat{h}_i h_{i+1}} + \dfrac{a_{i+1/2}}{\widehat{h}_{i+1}} - \dfrac{b_{i+1/2}}{2} - \dfrac{1}{2\Delta t}\right), \\ p_i^- = 0, \; p_i^0 = \dfrac{1}{2\Delta t}, \; p_i^+ = \dfrac{1}{2\Delta t}, \\ m_i^- = 0, \quad m_i^0 = \dfrac{1}{2}, \quad m_i^+ = \dfrac{1}{2}, \quad s^+ = 0, \quad s^- = c_{i+1/2} \end{cases} \tag{3.9}$$

and lastly,

$$q_{N/2}^{-,2} = \frac{-1}{2h}, \quad q_{N/2}^{-,1} = \frac{2}{h}, \quad q_{N/2}^0 = -\frac{3}{h}, \quad q_{N/2}^{+,1} = \frac{2}{h}, \quad q_{N/2}^{+,2} = \frac{-1}{2h}. \tag{3.10}$$

## 4. Error estimates

The difference operator $L_\varepsilon^{N,M}$ in (3.5) fails to fulfill the conditions of discrete maximum principle (as for this we need $q_i \geq 0$ to prove $A$ to be an $M$-matrix defined in Equation (4.9)). As a consequence, we have to modify the equation (3.4), for $i = N/2$

$$\begin{aligned} &q_{N/2}^{-,2} U_{N/2-2,k+1} + q_{N/2}^{-,1} U_{N/2-1,k+1} + q_{N/2}^0 U_{N/2,k+1} \\ &\quad + q_{N/2}^{+,1} U_{N/2+1,k+1} + q_{N/2}^{+,2} U_{N/2+2,k+1} = 0. \end{aligned} \tag{4.1}$$

From (3.4), for $i = N/2 - 1$, we have

$$\begin{aligned} &\left(\frac{2\varepsilon - h a_{N/2-1}}{2h^2}\right) U_{N/2-2,k+1} = f_{N/2-1,k+1} \\ &+ c_{N/2-1} p_1(x_{-1}, t_{k+1}) \\ &- r_{N/2-1}^0 U_{N/2-1,k+1} + r_{N/2-1}^+ U_{N/2,k+1} \\ &+ \frac{1}{\Delta t} U_{N/2-1,k} \end{aligned} \tag{4.2}$$

and, for $i = N/2 + 1$

$$\begin{aligned} &\left(\frac{2\varepsilon + h a_{N/2+1}}{2h^2}\right) U_{N/2+2,k+1} = f_{N/2+1,k+1} + c_{N/2+1} U(x_1, t_{k+1}) \\ &- r_{N/2+1}^0 U_{N/2+1,k+1} + r_{N/2+1}^+ U_{N/2,k+1} \\ &+ \frac{1}{\Delta t} U_{N/2+1,k}. \end{aligned} \tag{4.3}$$

Now putting the values of $U_{N/2-2,k+1}$ and $U_{N/2+2,k+1}$ from Equations (4.2) and (4.3) into (4.1), we get

$$\begin{aligned} &q_{N/2}^{-,2} \left(\frac{2h^2}{2\varepsilon - h a_{N/2-1}}\right) \Big( f_{N/2-1,k+1} + c_{N/2-1} p_1 \\ &\left(x_{-1}, t_{k+1}\right) - r_{N/2-1}^0 U_{N/2-1,k+1} \\ &- r_{N/2-1}^+ U_{N/2,k+1} - \frac{1}{\Delta t} U_{N/2-1,k} \Big) + q_{N/2}^{-,1} U_{N/2-1,k+1} \\ &+ q_{N/2}^0 U_{N/2,k+1} + q_{N/2}^{+,1} U_{N/2+1,k+1} \\ &+ q_{N/2}^{+,2} \left(\frac{2h^2}{2\varepsilon + h a_{N/2+1}}\right) \Big( f_{N/2+1,k+1} + c_{N/2+1} U\Big(x_1, t_{k+1}\Big) \\ &- r_{N/2+1}^0 U_{N/2+1,k+1} \end{aligned}$$

$$-r^{+}_{N/2+1}U_{N/2,k+1}-\frac{1}{\Delta t}U_{N/2+1,k}\Big)=0. \tag{4.4}$$

After rearranging the terms in (4.4), the discrete problem thus reads

$$\begin{cases} U_{i,0}=p_0(x_i) & \text{for} \quad i=0,\ldots,N \\ L^{N,M}_{\tau}U_{i,k+1}=\tilde{f}_{\tau,i,k+1} & \text{for} \quad i=1,\ldots,N-1 \\ U_{i,k+1}=p_1(x_i,t_{k+1}) & \text{for} \quad i=-N/2,\ldots,0 \\ U_{N,k+1}=p_2(t_{k+1}) & \text{for} \quad k=0,\ldots,M-1, \end{cases} \tag{4.5}$$

where

$$L^{N,M}_{\tau}U_{i,k+1}=\begin{cases} \left(r^{-}_i U_{i-1,k+1}+r^{0}_i U_{i,k+1}+r^{+}_i U_{i+1,k+1}\right) & \\ +\left(p^{-}_i U^{k}_{i-1}+p^{0}_i U^{k}_{i}+p^{+}_i U_{i+1,k}\right) & if\, i=N/2 \\ L^{N,M}_{\epsilon}U_{i,k+1} & if\, i\neq N/2 \end{cases} \tag{4.6}$$

and

$$\tilde{f}_{\tau,i,k+1}=\begin{cases} \left[m^{-}_i f_{i-1,k+1}+m^{0}_i f_{i,k+1}+m^{+}_i f_{i+1,k+1}\right] & \\ +l_1 p_1(x_{-1},t_{k+1})+l_2 U(x_1,t_{k+1}) & \text{if } i=N/2 \\ \tilde{f}_{i,k+1} & \text{if } i\neq N/2, \end{cases}$$

where, for $i = N/2$

$$\begin{cases} r^{-}_i=\dfrac{1}{2h}\left(4-\dfrac{2\left(2\epsilon+h^2 b_{i-1}+\dfrac{h^2}{\Delta t}\right)}{2\epsilon-ha_{i-1}}\right), \\ r^{0}_i=\dfrac{1}{2h}\left(-6+\dfrac{2\epsilon+ha_{i-1}}{2\epsilon-ha_{i-1}}+\dfrac{2\epsilon-ha_{i+1}}{2\epsilon+ha_{i+1}}\right), \\ r^{+}_i=\dfrac{1}{2h}\left(4-\dfrac{2\left(2\epsilon+h^2 b_{i+1}+\dfrac{h^2}{\Delta t}\right)}{2\epsilon+ha_{i+1}}\right), \\ p^{-}_i=\dfrac{h}{(2\epsilon-ha_{i-1})\Delta t}, \quad p^{0}_i=0, \quad p^{+}_i=\dfrac{h}{(2\epsilon+ha_{i+1})\Delta t}, \\ m^{-}_i=\dfrac{h}{(2\epsilon-ha_{i-1})}, \quad m^{0}_i=0, \quad m^{+}_i=\dfrac{h}{(2\epsilon+ha_{i+1})}, \\ l_1=\dfrac{-hc_{i-1}}{2\epsilon-ha_{i-1}}, \quad l_2=\dfrac{-hc_{i+1}}{2\epsilon+ha_{i+1}} \end{cases}$$

and for $i\neq N/2$, these coefficients are defined in (3.7−3.9). Let $D^{N,M}=\overline{D}^{N,M}\cap D$ and $\Gamma=\overline{D}^{N,M}\setminus D^{N,M}$.

**Lemma 4.1** Let $N\geq N_0$ be such that

$$\frac{N_0}{\ln N_0}\geq 2\sigma_0\gamma^{*} \quad \text{and} \tag{4.7}$$

$$\left(\|b\|_{\infty}+\Delta t^{-1}\right)\leq\frac{\gamma N_0}{2}. \tag{4.8}$$

Let $Y$ be the mesh function such that $Y\leq 0$ on $\Gamma^{N,M}$ and $L^{N,M}_{\tau}Y\geq 0$ in $D^{N,M}$. Then $Y\leq 0$ in $\overline{D}^{N,M}$.

*Proof.* Write $L^{N,M}_{\tau}$ as

$$-L^{N,M}_{\tau}Y_{i,k+1}=\left[A_{i,i-1}Y_{i-1,k+1}+A_{i,i}Y_{i,k+1}+A_{i,i+1}Y_{i+1,k+1}\right]$$
$$-\left[B_{i,i-1}Y_{i-1,k}+B_{i,i}Y_{i,k}+B_{i,i+1}Y_{i+1,k}\right], \tag{4.9}$$

where $A:=(A_{i,j})$ and $B:=(B_{i,j})$ are written as

$$A_{i,i-1}=-r^{-}_i, \quad A_{i,i}=-r^{0}_i, \quad A_{i,i+1}=-r^{+}_i,$$

$$B_{i,i-1}=p_i^-,\quad B_{i,i}=p_i^0,\quad B_{i,i+1}=p_i^+.$$

Clearly $B\geq 0$, since $p_i^-,p_i^0,p_i^+\geq 0$ and using (4.7) and (4.8) to show $A$ is an $M$-matrix [11]. Remaining part is proved by induction. Assume

$$Y_{i,k}\leq 0,\quad k=0,\ldots,N-1.$$

Then, we can rewrite (4.9) as

$$AY_{i,k+1}=BY_{i,k}-L^{N,M}Y_{i,k+1}.$$

Since $B\geq 0$, $Y_{i,k}\leq 0$, $A^{-1}\geq 0$ and $L_\tau^{N,M}Y_{i,k+1}\geq 0$ by hypothesis. Then, $Y_{i,k+1}\leq 0$ in $D^{N,M}$.

We obtain the following estimate as an immediate consequence of Lemma 4.1.

**Lemma 4.2** Let $U$ be the solution of (4.5) and the conditions (4.7) and (4.8) hold true. Then

$$\|U\|_{\infty,\overline{D}^{N,M}}\leq \|U\|_{\infty,I^{N,M}}+\frac{1}{\gamma}\|\tilde{f}_\tau\|_{\infty,\overline{D}^{N,M}}$$

*Proof.* Let

$$\phi_i^{\pm,k}=-\|U\|_{\infty,I^{N,M}}-\begin{cases}x_i\dfrac{\|\tilde{f}_\tau\|_\infty}{\gamma}\mp U_{i,k} & for\ 1\leq i\leq N/2\\ (2-x_i)\dfrac{\|\tilde{f}_\tau\|_\infty}{\gamma}\mp U_{i,k} & for\ N/2<i\leq N\end{cases}.$$

Then $\phi_N^{\pm,k+1}\leq 0$, $\phi_i^{\pm,k+1}\leq 0$, $-N/2\leq i\leq 0$ and $\phi_i^{\pm,0}\leq 0$, $0\leq i\leq N$. For $i\neq N/2$, $L_\tau^{N,M}\phi_i^{\pm,k+1}\geq 0$. Further

$$L_\tau^{N,M}\phi_{N/2}^{\pm,k+1}=L_\tau^{N,M}\left\{-\|U\|_{\infty,I^{N,M}}-\frac{1}{\gamma}\|\tilde{f}_\tau\|_{\infty,\overline{D}^{N,M}}\mp U_{N/2,k+1}\right\}\geq (D_x^F-D_x^B)\phi_{N/2}^{\pm,k+1}$$

and

$$(D_x^F-D_x^B)\phi_{N/2}^{\pm,k+1}=\frac{1}{2h}\left(-\phi_{N/2+2}+4\phi_{N/2+1}-6\phi_{N/2}-\phi_{N/2-2}+4\phi_{N/2-1}\right)\geq 0.$$

Consequently, the required result follows from Lemma 4.1.

Next, we decompose the solution into smooth and singular component. We write $U_{i,k+1}=V_{i,k+1}+W_{i,k+1}$. Here the mesh functions $V_L$ and $V_R$ satisfy

$$\left.\begin{aligned}&L_\tau^{N,M}V_{L,i,k+1}=\tilde{f}_{\tau,i,k+1}\ \text{ for }\ i=1,\ldots,N/2-1,\\&V_{L,i,k+1}=v(x_i,t_{k+1})\ \text{ for }\ i=-N/2,\ldots,0,\\&V_{L,N/2,k+1}=v(1^-,t_{k+1}),\quad k\geq 0,\\&V_{L,i,0}=v(x_i,0)\ \text{ for }\ i=1,\ldots,N/2\end{aligned}\right\}$$

and

$$\left.\begin{aligned}&L_\tau^{N,M}V_{R,i,k+1}=\tilde{f}_{\tau,i,k+1}\quad \text{for}\quad i=N/2+1,\ldots,N-1,\\&V_{R,N/2,k+1}=v(1^+,t_{k+1}),\quad V_{R,N,k+1}=v(2,t_{k+1}),\quad k\geq 0,\\&V_{R,i,0}=v(x_i,0)\quad \text{for}\quad i=N/2,\ldots,N.\end{aligned}\right\}\tag{4.11}$$

The mesh function $W_L$ and $W_R$ satisfy

$$\left.\begin{aligned}&L_\tau^{N,M}W_{L,i,k+1}=0\quad \text{for}\quad i=1,\ldots,N/2-1,\\&L_\tau^{N,M}W_{R,i,k+1}=0\quad \text{for}\quad i=N/2+1,\ldots,N-1,\\&W_{L,i,k+1}=0\quad \text{for}\quad i=-N/2,\ldots,0,\\&W_{R,N,k+1}=0,\quad k\geq 0,\\&W_{L,i,0}=0\quad \text{for}\quad i=0,\ldots,N/2,\\&W_{R,i,0}=0\quad \text{for}\quad i=N/2,\ldots,N,\\&W_{R,N/2,k+1}+V_{R,N/2,k+1}=W_{L,N/2,k+1}+V_{L,N/2,k+1},\\&D_x^FW_{R,N/2,k+1}+D_x^FV_{R,N/2,k+1}=D_x^BW_{L,N/2,k+1}\\&+D_x^BV_{L,N/2,k+1},k\geq 0.\end{aligned}\right\}\tag{4.12}$$

Numerical solution $U$ satisfies

$$U_{i,k+1}=\begin{cases}V_{L,i,k+1}+W_{L,i,k+1}, & 0\leq i\leq N/2-1\\ V_{L,i,k+1}+W_{L,i,k+1}=V_{R,i,k+1}+W_{R,i,k+1}, & i=N/2\\ V_{R,i,k+1}+W_{R,i,k+1}, & N/2+1\leq i\leq N-1.\end{cases}\tag{4.13}$$

**Lemma 4.3** Let $V_L$ and $V_R$ are the solutions of (4.10), (4.11) and $v$ is the solution of (2.4). Then under the assumptions (4.7) and (4.8), we have

$$\begin{cases} |V_{L,i,k+1} - v(x_i, t_{k+1})| \le C\left(N^{-2} + \Delta t\right)x_i & \text{for } i = 1, \ldots, N/2 - 1 \\ |V_{R,i,k+1} - v(x_i, t_{k+1})| \le C\left(N^{-2} + \Delta t\right)(2 - x_i) & \text{for } i = N/2 + 1, \ldots, N - 1. \end{cases}$$

*Proof.* Consider $\Psi_{L,i}^k = -C(N^{-2} + \Delta t)x_i, \quad i \in \{0, \ldots, N/2\}$. Then

$$L_\tau^{N,M}(V_{L,i,k+1} - v(x_i, t_{k+1})) = \left(L_\varepsilon - L_\tau^{N,M}\right)v(x_i, t_{k+1}).$$

We observe that for $i \in \{1, \ldots, N/2 - 1\}$

$$\left|L_\tau^{N,M}(V_{L,i,k+1} - v(x_i, t_{k+1}))\right|$$

$$\le \begin{cases} C\left[(\epsilon + h_i)(h_i + h_{i+1})\left\|\dfrac{\partial^3 v}{\partial x^3}\right\|_\infty + h_i^2\left(\left\|\dfrac{\partial^2 v}{\partial x^2}\right\|_\infty + \left\|\dfrac{\partial v}{\partial x}\right\|_\infty\right) + \Delta t\left\|\dfrac{\partial^2 v}{\partial t^2}\right\|_\infty\right], \ 1 \le i \le N/4 \\ C\left[h^2\left(\epsilon\left\|\dfrac{\partial^4 v}{\partial x^4}\right\|_\infty + \left\|\dfrac{\partial^3 v}{\partial x^3}\right\|_\infty\right) + \Delta t\left\|\dfrac{\partial^2 v}{\partial t^2}\right\|_\infty\right], \quad N/4 + 1 \le i \le N/2 - 1. \end{cases}$$

Using Theorem 2.4, conditions $h_i \le CN^{-1}$ and $\varepsilon \le N^{-1}$ to obtain

$$\left|L_\tau^{N,M}(V_{L,i,k+1} - v(x_i, t_{k+1}))\right| \le C\left(N^{-2} + \Delta t\right) \le L_\tau^{N,M}\psi_{L,i}^{k+1}.$$

Further the required result follows from Lemma 4.1

$$|V_{L,i,k+1} - v(x_i, t_n)| \le C\left(N^{-2} + \Delta t\right)x_i, \quad 1 \le i \le N/2 - 1.$$

Consider

$$\psi_{R,i} = -C\left(N^{-2} + \Delta t\right)(2 - x_i), \quad i \in \{N/2, \ldots, N\}.$$

Similarly, for $i \in \{N/2 + 1, \ldots, N - 1\}$ we obtain

$$\left|L_\varepsilon^{N,M}(V_{R,i,k+1} - v(x_i, t_{k+1}))\right|$$

$$\le \begin{cases} C\left[h^2\left(\varepsilon\left\|\dfrac{\partial^4 v}{\partial x^4}\right\|_\infty + \left\|\dfrac{\partial^3 v}{\partial x^3}\right\|_\infty\right) + \Delta t\left\|\dfrac{\partial^2 v}{\partial t^2}\right\|_\infty\right], \quad N/2 + 1 \le i \le 3N/4 - 1 \\ C\left[(\varepsilon + h_{i+1})(h_i + h_{i+1}) \left\| \dfrac{\partial^3 v}{\partial x^3}\right\|_\infty + h_{i+1}^2\left(\left\|\dfrac{\partial^2 v}{\partial x^2}\right\|_\infty + \left\|\dfrac{\partial v}{\partial x}\right\|_\infty\right) + \Delta t\left\|\dfrac{\partial^2 v}{\partial t^2}\right\|_\infty\right], \\ 3N/4 \le i \le N - 1. \end{cases}$$

Arguing similarly for $(V_R - v)$, we obtain the following estimate for $i \in \{N/2 + 1, \ldots, N - 1\}$

$$|V_{R,i,k+1} - v(x_i, t_{k+1})| \le C\left(N^{-2} + \Delta t\right)(2 - x_i).$$

Now, we define the following two mesh functions on $S_x^N = \{x_i\}_0^N$:

$$S_i = \prod_{j=1}^{i}\left(1 + \frac{\alpha h_j}{\epsilon}\right), 1 \le i \le N/2, \quad \text{and}$$

$$Q_i = \prod_{j=1}^{N-i}\left(1 + \frac{\alpha h_j}{\epsilon}\right), \quad N/2 \le i \le N - 1$$

so that $S_0 = 1$ and $Q_N = 1$, where $\alpha$ is a positive constant.

**Lemma 4.4** Let $\alpha \le \gamma/2$, then the functions $S_i$ and $Q_i$ satisfy

$$-L_\tau^{N,M}S_i \ge \begin{cases} \dfrac{C}{\varepsilon + \alpha H}S_i & \text{if } \ i = 1, \ldots, N\Big/4 \\ \dfrac{C}{\varepsilon + \alpha h}S_i & \text{if } \ i = N\Big/4 + 1, \ldots, N\Big/2 - 1 \end{cases}$$

and

$$-L_\tau^{N,M}Q_i \geq \begin{cases} \dfrac{C}{\varepsilon+\alpha h}Q_i & \text{if } i=N\Big/2+1,\ldots,3N\Big/4-1 \\ \dfrac{C}{\varepsilon+\alpha H}Q_i & \text{if } i=3N\Big/4,\ldots,N-1. \end{cases}$$

*Proof.* For $i = 1,\ldots,N/4$

$$-L_\tau^{N,M}S_i = -\frac{2\varepsilon}{\widehat{h}_i}\left[\left(\frac{S_{i+1}-S_i}{h_{i+1}}\right)-\left(\frac{S_i-S_{i-1}}{h_i}\right)\right] - a_{i-\frac{1}{2}}\left(\frac{S_i-S_{i-1}}{h_i}\right)+b_{i-\frac{1}{2}}S_{i-\frac{1}{2}}.$$

Since $S_i = \left(1+\frac{\alpha h_i}{\varepsilon}\right)S_{i-1}$, $S_i - S_{i-1} = \frac{\alpha h_i}{\varepsilon}S_{i-1}$ and $a_i \leq -\gamma_1 \leq -2\alpha$. Thus

$$-L_\tau^{N,M}S_i = -\frac{2\alpha}{\widehat{h}_i}(S_i-S_{i-1}) - a_{i-\frac{1}{2}}S_{i-1}+b_{i-\frac{1}{2}}S_{i-\frac{1}{2}} \geq \frac{C}{\varepsilon+\alpha h_i}S_i.$$

Next, for $i = N/4+1,\ldots,N/2-1$

$$-L_\tau^{N,M}S_i = -\varepsilon\delta_x^2 S_i - a_i D_x^0 S_i + b_i S_i + D_t^- S_i \geq \frac{C}{\varepsilon+\alpha h}S_i.$$

Now, for $i = N/2+1,\ldots,3N/4-1$

$$-L_\tau^{N,M}Q_i = -\varepsilon\delta_x^2 Q_i - a_i D_x^0 Q_i + b_i Q_i + D_t^- Q_i$$

$$= -\frac{\varepsilon}{h}\left[\left(\frac{Q_{i+1}-Q_i}{h_{i+1}}\right)-\left(\frac{Q_i-Q_{i-1}}{h_i}\right)\right] - a_i\left(\frac{Q_{i+1}-Q_{i-1}}{\widehat{h}_i}\right) + b_i Q_i.$$

Since $Q_{i+1}-Q_i = \frac{-\alpha h_{N-i}}{\varepsilon}Q_{i+1}$ and $a_i \geq \gamma_2 \geq 2\alpha$. Thus

$$-L_\tau^{N,M}Q_i \geq \frac{\alpha}{h}(Q_{i+1}-Q_i)+\frac{a_i}{2}\frac{\alpha}{\varepsilon}(Q_{i+1}-Q_{i-1}) \geq \frac{C}{(\varepsilon+\alpha h)}Q_i.$$

For $i = 3N/4,\ldots,N-1$

$$-L_\tau^{N,M}Q_i \geq -\frac{2\varepsilon}{\widehat{h}_i}\left[\left(\frac{Q_{i+1}-Q_i}{h_{i+1}}\right)-\left(\frac{Q_i-Q_{i-1}}{h_i}\right)\right] - a_{i+\frac{1}{2}}\left(\frac{Q_{i+1}-Q_i}{h_{i+1}}\right)$$

and $Q_{i+1}-Q_i = \frac{-\alpha h_{N-i}}{\varepsilon}Q_{i+1}$, $a_i \geq \gamma_2 \geq 2\alpha$. Therefore

$$-L_\tau^{N,M}Q_i \geq \frac{2\alpha}{\widehat{h}_i}(Q_{i+1}-Q_i)+a_{i+\frac{1}{2}}\frac{\alpha}{\varepsilon}Q_{i+1} \geq \frac{C}{(\varepsilon+\alpha h_{N-i})}Q_i.$$

**Lemma 4.5** Let $\sigma_0 \geq 2/\alpha$. Then for $\{x_i\}_0^N$, the following inequalities are satisfied:

$$\left(\frac{S_i}{S_{N/2}}\right) \leq CN^{\frac{-4(1-2i)}{N}}, \quad i = N/4,\ldots,N/2-1 \tag{4.14}$$

and

$$\left(\frac{Q_i}{Q_{N/2}}\right) \leq CN^{-4\left(\frac{2i}{N-1}\right)}, \quad i = N/2+1,\ldots,3N/4. \tag{4.15}$$

*Proof.* Since $h_i = h$ for $i = N/4,\ldots,N/2-1$, thus

$$\left(\frac{S_i}{S_{N/2}}\right) = \left(1-\frac{\alpha h}{\varepsilon+\alpha h}\right)^{(N/2-i)}.$$

Taking log on both sides to obtain

$$\left(\frac{S_i}{S_{N/2}}\right) \leq \exp\left((N/2-i)\left(\frac{-\alpha h}{\varepsilon+\alpha h}\right)\right) \leq CN^{-4(1-2i/N)}.$$

Since sequence $N^{\frac{8(\alpha\sigma_0)^2(1-2i/N)(N^{-1}\ln N)}{(1+4\alpha\sigma_0 N^{-1}\ln N)}}$ is bounded. Further, for $i = N/2+1,\ldots,3N/4$

$$\left(\frac{Q_i}{Q_{N/2}}\right) = \frac{\prod_{j=1}^{N-i}\left(1+\frac{\alpha h}{\varepsilon}\right)}{\prod_{j=1}^{N/2}\left(1+\frac{\alpha h}{\varepsilon}\right)} = \left(1+\frac{\alpha h}{\varepsilon}\right)^{-(i-N/2)}.$$

Similarly, it is easy to follow that $\left(\frac{Q_i}{Q_{N/2}}\right) \leq CN^{-4(2i/N-1)}$.

Now, we will calculate the errors for the layers components $W_L$ and $W_R$ in $((0,1-\sigma]\cup[1+\sigma,2))\times(0,T]$.

**Lemma 4.6** Let $\alpha \leq \gamma/2$ and $\sigma_0 \geq 2/\alpha$. Then under the assumptions (4.7) and (4.8), the errors associated to the layer components satisfy

$$\begin{cases} \left|W_{L,i,k+1}-w(x_i,t_{k+1})\right| \leq CN^{-2} & \text{for } i=1,\ldots,N/4 \\ \left|W_{R,i,k+1}-w(x_i,t_{k+1})\right| \leq CN^{-2} & \text{for } i=3N/4,\ldots,N-1. \end{cases}$$

*Proof.* From (4.13), we have

$$U_{N/2,k+1} = V_{R,N/2,k+1}+W_{R,N/2,k+1} = V_{L,N/2,k+1}+W_{L,N/2,k+1}$$

$$\left|U_{N/2,k+1}\right| \leq \left|V_{L,N/2,k+1}\right|+\left|W_{L,N/2,k+1}\right|.$$

Using Theorem 2.4 and Lemma 4.2 to obtain $\left|W_{N/2,k+1}\right| \leq C$. Consider

$$\phi_{L,i}^{k+1} = -C\left(\frac{S_i}{S_{N/2}}\right) \quad \text{for} \quad i=0,\ldots,N/2.$$

For $i = 0,\ldots,N/2$ and using Lemma 4.4, we have

$$L_\tau^{N,M}\left(\phi_{L,i}^{k+1} \pm W_{L,i,k+1}\right) = L_\tau^{N,M}\phi_{L,i}^{k+1} \pm L_\tau^{N,M}W_{L,i,k+1} \geq \frac{C}{\varepsilon+\alpha H}\frac{S_i}{S_{N/2}} \geq 0$$

and $W_{L,0,k+1} = 0 = W_{L,i,0}$ for $i \leq N/2$. Thus

$$\phi_{L,0}^{k+1} \pm W_{L,0,k+1} = \phi_{L,0}^{k+1} = \frac{-CS_0}{S_{N/2}} \leq 0.$$

Similarly, we can show $\phi_{L,i}^0 \pm W_{L,i,0} \leq 0$. Therefore using Lemma 4.1, we get

$$|W_{L,i,k+1}| \leq C\left(\frac{S_i}{S_{N/2}}\right) \quad \text{for} \quad i = 1,\ldots,N/2-1. \tag{4.16}$$

Further

$$|W_{L,i,k+1}| \leq C\left(\frac{S_i}{S_{N/2}}\right) \leq C\left(\frac{S_{N/4}}{S_{N/2}}\right)$$

and from Lemma 4.5, we have

$$|W_{L,i,k+1}| \leq CN^{-2}. \tag{4.17}$$

Since $\alpha \leq \gamma/2 \leq \gamma$, $\sigma = \sigma_0 \varepsilon \ln N$, $\sigma_0 \geq 2/\alpha$ and using Theorem 2.4, we have

$$|w(x_i,t_{k+1})| \leq C\exp\left(\frac{-(1-x_i)\gamma_1}{\varepsilon}\right) \leq CN^{-2} \text{ for } i=1,\ldots,N/4. \tag{4.18}$$

Combine (4.17) and (4.18) to obtain

$$|W_{L,i,k+1} - w(x_i,t_{k+1})| \leq CN^{-2} \quad \text{for} \quad i=1,\ldots,N/4.$$

Now, consider $\phi_{R,i}^{k+1} = -C\left(\frac{Q_i}{Q_{N/2}}\right)$ for $i = N/2,\ldots,N$. Then using Lemma 4.4, for $i = 3N/4,\ldots,N-1$ to obtain

$$L_\tau^{N,M}\left(\phi_{R,i}^{k+1} \pm W_{R,i,k+1}\right) = L_\tau^{N,M}\phi_{R,i}^{k+1} \pm L_\tau^{N,M}W_{R,i,k+1} \geq \frac{C}{\varepsilon+\alpha H}\frac{Q_i}{Q_{N/2}} \geq 0$$

and $W_{R,N,k+1} = 0 = W_{R,i,0}$. Thus $\phi_{R,N}^{k+1} \pm W_{R,N,k+1} = \phi_{R,N}^{k+1} \leq 0$.

Similarly, we can show $\phi_{R,i}^0 \pm W_{R,i,0} \leq 0$. Thus using Lemma 4.1, we have $\phi_{R,i}^{k+1} \pm W_{R,i,k+1} \leq 0$ for $i = N/2,\ldots,N$. Thus

$$|W_{R,i,k+1}| \leq \left|\phi_{R,i}^{k+1}\right| \leq C\left(\frac{Q_i}{Q_{N/2}}\right) \text{ for } i=N/2+1,\ldots,N-1 \leq C\left(\frac{Q_{N/4}}{Q_{N/2}}\right) \text{ for } i=3N/4,\ldots,N-1$$

since $Q$ is decreasing and from Lemma 4.5, we have $\left(\frac{Q_{N/4}}{Q_{N/2}}\right) \leq CN^{-2}$. Thus

$$|W_{R,i,k+1}| \leq CN^{-2}. \tag{4.19}$$

Now, using Theorem 2.4 and doing the same calculation as we do for $W_L$, we have

$$|w(x_i,t_{k+1})| \leq C\exp\left(\frac{-(1-x_i)\gamma_2}{\varepsilon}\right) \leq CN^{-2}. \tag{4.20}$$

Using (4.19) and (4.20), we obtain $\left|W_{R,i,k+1} - w(x_i, t_{k+1})\right| \leq CN^{-2}$.

Next, we will state and prove some lemmas required to obtain $\varepsilon$ uniform error bounds.

**Lemma 4.7** The following inequalities hold true:

$$\exp(-\alpha(1-x_i)/\varepsilon) \leq \left(\frac{S_i}{S_{N/2}}\right), \quad i=1,\ldots,N/2-1 \tag{4.21}$$

and

$$\exp(-\alpha(x_i-1)/\varepsilon) \leq \left(\frac{Q_i}{Q_{N/2}}\right), \quad i=N/2+1,\ldots,N-1. \tag{4.22}$$

*Proof*. Follows from Lemma 2.5 in [40], for each $j$, we obtain

$$\exp\left(\frac{-\alpha h_j}{\varepsilon}\right) = \left(\exp\left(\frac{\alpha h_j}{\varepsilon}\right)\right)^{-1} \leq \left(1+\frac{\alpha h_j}{\varepsilon}\right)^{-1}. \tag{4.23}$$

Multiplying (4.23), for $j = i+1,\ldots,N/2$ to obtain

$$\prod_{i+1}^{N/2}\exp\left(\frac{-\alpha h_j}{\varepsilon}\right) \leq \prod_{i+1}^{N/2}\left(1+\frac{\alpha h_j}{\varepsilon}\right)^{-1} \leq \frac{S_i}{S_{N/2}}.$$

Further, multiplying (4.23), for $j = N-(i+1),\ldots, N/2$, we have

$$\prod_{N-(i+1)}^{N/2}\exp\left(\frac{-\alpha h_j}{\varepsilon}\right) \leq \prod_{N-(i+1)}^{N/2}\left(1+\frac{\alpha h_j}{\varepsilon}\right)^{-1} \leq \frac{Q_i}{Q_{N/2}}.$$

**Lemma 4.8** The difference operators $D_x^B$ and $D_x^F$ satisfy the following inequalities:

$$D_x^B S_{N/2} \geq \frac{C}{\varepsilon+\alpha h} S_{N/2}, \quad -D_x^F Q_{N/2} \geq \frac{C}{\varepsilon+\alpha h} Q_{N/2}.$$

*Proof.* Since $(S_i - S_{i-1}) = \frac{\alpha h_i}{\varepsilon} S_{i-1}$, we have

$$D_x^B S_{N/2} = \frac{1}{2h}\left(S_{N/2-2} - 4S_{N/2-1} + 3S_{N/2}\right)$$
$$\geq \frac{\alpha}{\varepsilon}\frac{S_{N/2}}{\left(1+\frac{\alpha h}{\varepsilon}\right)^2}\left(1+\frac{\alpha h}{\varepsilon}\right) \geq \frac{C}{\varepsilon+\alpha h} S_{N/2}.$$

Since $Q_{i+1} - Q_i = \frac{-\alpha h_{N-i}}{\varepsilon} Q_{i+1}$, we have

$$-D_x^F Q_{N/2} = \frac{1}{2h}\left(Q_{N/2+2} - 4Q_{N/2+1} + 3Q_{N/2}\right)$$
$$\geq \frac{\alpha}{\varepsilon}\frac{\left(1+\frac{\alpha h}{\varepsilon}\right)}{\left(1+\frac{\alpha h}{\varepsilon}\right)^2} Q_{N/2} \geq \frac{C}{\varepsilon+\alpha h} Q_{N/2}.$$

**Theorem 4.9** Let $u$ and $U_{i,k+1}$ be the solution of (2.1) and (4.5), respectively. Then under the assumptions (4.7), (4.8) and $\alpha \leq \gamma/2$ and $\sigma_0 \geq 2/\alpha$, we get

$$|U_{i,k+1} - u(x_i, t_{k+1})| \leq \begin{cases} C\left(N^{-2} + \Delta t\right) & \text{for } i = 1, \ldots, N/4, 3N/4, \ldots, N-1 \\ C\left(N^{-2}\ln^2 N + \Delta t\right) & \text{for } i = N/4+1, \ldots, 3N/4-1. \end{cases} \tag{4.24}$$

*Proof.* We compute the error separately in the layer region and outside the layer region.

**Case 1:** For $i = 1, \ldots, N/4, 3N/4, \ldots, N-1$.Follows from the triangle inequality to $U - u$ and using Lemma 4.3 and Lemma 4.6 to obtain

$$|U_{i,k+1} - u(x_i, t_{k+1})| \leq |V_{L,i,k+1} - v(x_i, t_{k+1})|$$
$$+ |W_{L,i,k+1} - w(x_i, t_{k+1})|$$
$$\leq C\left(N^{-2} + \Delta t\right).$$

**Case 2:** Here, we need to find out the error estimate $\left|U_{i,k+1} - u(x_i, t_{k+1})\right|$ for $i = N/4+1, \ldots, 3N/4-1$. For $i = N/4, 3N/4$, From Case 1

$$|U_{i,k+1} - u(x_i, t_{k+1})| \leq C\left(N^{-2} + \Delta t\right). \tag{4.25}$$

Consider

$$L_\tau^{N,M}\left(U_{i,k+1} - u(x_i, t_{k+1})\right) = L_\tau^{N,M} U_{i,k+1} - L_\tau^{N,M} u(x_i, t_{k+1})$$
$$= \left(\varepsilon\left(\frac{\partial^2}{\partial x^2} - \delta x^2\right) + a\left(\frac{\partial}{\partial x} - D_x^0\right) - \left(\frac{\partial}{\partial t} - D_t^-\right)\right) u(x_i, t_{k+1}).$$

Using Theorem 2.4, we have

$$\left|L_\tau^{N,M}(U_{i,k+1} - u(x_i, t_{k+1}))\right| \leq h \int_{x_{i-1}}^{x_{i+1}} \left(\varepsilon\left|\frac{\partial^4 u}{\partial x^4}\right| + \left|\frac{\partial^3 u}{\partial x^3}\right|\right) dx$$
$$+ C\Delta t \left\|\frac{\partial^2 u}{\partial t^2}\right\|_\infty$$
$$\leq C\left[h^2 + \frac{h}{\varepsilon^2}\exp\left(\frac{-(1-x_i)\gamma_1}{\varepsilon}\right)\sinh\left(\frac{h\gamma_1}{\varepsilon}\right)\right] + \Delta t. \tag{4.26}$$

From (3.5), we have

$$2\sigma_0\gamma^* \leq \frac{N}{\ln N} 4\sigma_0\gamma_1 \leq \frac{2N}{\ln N}, \quad \frac{4\sigma\gamma_1}{\varepsilon \ln N} \leq \frac{2N}{\ln N} \text{ and } \frac{h\gamma_1}{\varepsilon} \leq 2.$$

Since $\sinh x \leq cx$, $x \in [0, 2]$. This implies $\sinh\left(\frac{h\gamma_1}{\varepsilon}\right) \leq C\left(\frac{\gamma_1 h}{\varepsilon}\right)$. Therefore (4.26) becomes

$$\left|L_\tau^{N,M}\left(U_{i,k+1} - u(x_i, t_{k+1})\right)\right|$$
$$\leq C\left[\left(h^2 + \frac{h^2}{\varepsilon^3}\exp\left(\frac{-(1-x_i)\gamma_1}{\varepsilon}\right)\right) + \Delta t\right]. \tag{4.27}$$

Similarly, for $i = N/2+1, \ldots, 3N/4-1$

$$\left|L_\tau^{N,M}\left(U_{i,k+1} - u(x_i, t_{k+1})\right)\right|$$
$$\leq C\left[\left(h^2 + \frac{h^2}{\varepsilon^3}\exp\left(\frac{-(x_i-1)\gamma_2}{\varepsilon}\right)\right) + \Delta t\right]. \tag{4.28}$$

Further at $x_{N/2} = 1$

$$\left|L_\tau^{N,M}\left(U_{N/2,k+1} - u\left(x_{N/2}, t_{k+1}\right)\right)\right|$$
$$= \left|\tilde{f}_{\tau,N/2,k+1} - L_\tau^{N,M} u_{N/2,k+1}\right| \leq C\left(\frac{h^2}{\varepsilon^3} + \Delta t\right).$$

Define the following discrete function

$$\Theta_i^k = \begin{cases} -C\left(N^{-2}+\Delta t\right)(1+(x_i-(1-\sigma)))-C\dfrac{h^2}{\varepsilon^2}\left(\dfrac{S_i}{S_{N/2}}\right) & \text{for} \quad i=N/4,\ldots,N/2 \\ -C\left(N^{-2}+\Delta t\right)(1+((1+\sigma)-x_i))-C\dfrac{h^2}{\varepsilon^2}\left(\dfrac{Q_i}{Q_{N/2}}\right) & \text{for} \quad i=N/2+1,\ldots,3N/4. \end{cases}$$

Then

$$L_\tau^{N,M}\Theta_i^{k+1} = \begin{cases} -C\left(N^{-2}+\Delta t\right)L_\tau^{N,M}(x_i)-C\dfrac{h^2}{\varepsilon^2}\left(\dfrac{L_\tau^{N,M}S_i}{S_{N/2}}\right) & \text{for} \quad i=N/4,\ldots,N/2 \\ C\left(N^{-2}+\Delta t\right)L_\tau^{N,M}(x_i)-C\dfrac{h^2}{\varepsilon^2}\left(\dfrac{L_\tau^{N,M}Q_i}{Q_{N/2}}\right) & \text{for} \quad i=N/2+1,\ldots,3N/4 \end{cases}$$

$$= \begin{cases} -C\left(N^{-2}+\Delta t\right)(a_i-b_ix_i)-C\dfrac{h^2}{\varepsilon^2}\left(\dfrac{L_\tau^{N,M}S_i}{S_{N/2}}\right) & \text{for} \quad i=N/4,\ldots,N/2 \\ C\left(N^{-2}+\Delta t\right)(a_i-b_ix_i)-C\dfrac{h^2}{\varepsilon^2}\left(\dfrac{L_\tau^{N,M}Q_i}{Q_{N/2}}\right) & \text{for} \quad i=N/2+1,\ldots,3N/4. \end{cases}$$

Using assumption $\alpha \le \gamma/2$, Lemma 4.6 and Lemma 4.7 to obtain

$$L_\tau^{N,M}\Theta_i^{k+1} \ge \begin{cases} C\gamma_1\left(N^{-2}+\Delta t\right)+C\dfrac{h^2}{\varepsilon^2}\exp(-(1-x_i)\gamma_1/\varepsilon) & \text{for} \quad i=N/4,\ldots,N/2 \\ C\gamma_2\left(N^{-2}+\Delta t\right)+C\dfrac{h^2}{\varepsilon^2}\exp(-(x_i-1)\gamma_2/\varepsilon) & \text{for} \quad i=N/2+1,\ldots,3N/4. \end{cases}$$

Assumption (4.7) implies $\frac{h}{\varepsilon} \le \frac{2}{\gamma^*}$ and using Lemma 4.8, we get

$$L_\tau^{N,M}\Theta_i^{k+1} \ge \left(D_x^F - D_x^B\right)\Theta_i^{k+1} \ge 2C\left(N^{-2}+\Delta t\right)+2C\frac{h^2}{\varepsilon^3}. \tag{4.29}$$

Therefore, it follows from (4.25)−(4.29) that

$$\begin{cases} L_\tau^{N,M}\Theta_i^{k+1} \geq \left|L_\tau^{N,M}(U_{i,k+1}-u(x_i,t_{k+1}))\right| & \text{for } i=N/4+1,\ldots,3N/4-1 \\ -\Theta_i^{k+1} \geq |U_{i,k+1}-u(x_i,t_{k+1})| & \text{for } i=N/4,3N/4 \text{ and} \\ -\Theta_i^{0} \geq |U_{i,0}-u(x_i,t_0)|. \end{cases}$$

Then applying discrete maximum principle to $\Theta_i^{k+1} \pm (U_{i,k+1}-u(x_i,t_{k+1}))$ over the domain $\overline{D}^{N,M} \cap ([1-\sigma,1+\sigma]\times[0,T])$ we obtain for $i = N/4+1, \ldots, 3N/4-1$

$$|U_{i,k+1}-u(x_i,t_{k+1})| \leq C\left(\frac{h^2}{\varepsilon^2}+\Delta t\right) \leq C\left(N^{-2}\ln^2 N+\Delta t\right).$$

## 5. Numerical illustration

The performance of the proposed method is examined in this section and the theoretical estimates are numerically verified. We consider two test problems for numerical computations.

**Example 5.1** Consider the following singularly perturbed problem of class $(2.1)-(2.3)$

$$\begin{cases} \varepsilon u_{xx}+a(x)u_x-x(2-x)u(x,t)-u_t=f(x,t)+u(x-1,t), & (x,t)\in(0,2)\times(0,2], \\ u(x,0)=0, \quad x\in[0,2], \\ u(x,t)=t^2, \quad (x,t)\in[-1,0]\times[0,2], \\ u(2,t)=0, \quad t\in(0,2], \end{cases}$$

where

$$a(x)=\begin{cases} -(2+x(2-x)), & x\in[0,1] \\ (2+x(2-x)), & x\in(1,2] \end{cases}$$

and

$$f(x,t)=\begin{cases} 2(1+x^2)t^2, & (x,t)\in[0,1]\times[0,2] \\ 3(1+x^2)t^2, & (x,t)\in(1,2]\times[0,2] \end{cases}$$

**Example 5.2** Consider the following singularly perturbed problem of class $(2.1)-(2.3)$

$$\begin{cases} \varepsilon u_{xx}+a(x)u_x-5u(x,t)-u_t=f(x,t)+2u(x-1,t), & (x,t)\in(0,2)\times(0,2], \\ u(x,0)=0, \quad x\in[0,2], \\ u(x,t)=0, \quad (x,t)\in[-1,0]\times[0,2], \\ u(2,t)=0, \quad t\in(0,2], \end{cases}$$

where

$$a(x)=\begin{cases} -(4+x^2), & x\in[0,1] \\ (6-x^2), & x\in(1,2] \end{cases}$$

and

$$f(x,t)=\begin{cases} 4xt^2\exp(-t), & (x,t)\in[0,1]\times[0,2] \\ 4(2-x)t^2\exp(-t), & (x,t)\in(1,2]\times[0,2] \end{cases}$$

The exact solutions for the problems are unknown for comparison. Therefore, we use the double mesh principle to estimate the error. The maximum point−wise error ($E_\varepsilon^{N,\triangle t}$) and order of convergence ($R_\varepsilon^{N,\triangle t}$) are calculated using:

$$E_\varepsilon^{N,\triangle t} := \max\left|U^{N,\triangle t}(x_i,t_{k+1})-\tilde{U}^{2N,\triangle t/2}(x_i,t_{k+1})\right|,$$
$$R_\varepsilon^{N,\triangle t} := \log_2\left(\frac{E_\varepsilon^{N,\triangle t}}{E_\varepsilon^{2N,\triangle t/2}}\right),$$

where $U^{N,\triangle t}(x_i,t_{k+1})$ and $\tilde{U}^{2N,\triangle t/2}(x_i,t_k)$ are the approximate solutions obtained on the mesh $D^{N,M}$ and $D^{2N,2M}$, respectively. When, the perturbation parameter approaches zero, the problem's solution exhibits turning point behaviour (Figs. 1−4).

Maximum absolute error and order of convergence for problems 5.1 and 5.2 are tabulated in Tables 1 and 2. Moreover, the maximum absolute errors for problems 5.1 and 5.2 are plotted in Figs. 5 and 6, respectively. The surface plot of the numerical solution for problems 5.1 and 5.2 are plotted in Figs. 1 and 3, respectively. Also, the numerical solutions at final time step ($t = 2$) for different values of $\varepsilon$ are displayed in Figs. 2 and 4.

The numerical results tabulated in Tables 1 and 2 do not clearly depict the theoretical order of convergence for spatial discretization. It is to be noted that the error in numerical solution is due to spatial and temporal discretization. As a consequence, the errors given in Tables 1 and 2 are a combination of temporal and spatial errors, with the layer regions playing a significant role.

The hybrid difference scheme improves accuracy in space only. To verify this, we performed numerical experiments for $M = N^2$ and numerical results are tabulated in Table 3. In Table 4, we have fixed $\varepsilon = 2^{-6}$ and $N = 512$ and reduce $\Delta t$ by half and the associated errors are presented at different values of $x$. It can be observed that the errors reduce by almost half which confirms the first-order convergence in time. Figures 7 and 8 have also been illustrated to show the errors in layer region and outside layer region. This demonstrates that the numerical method is second-order spatially accurate outside of the interior layer and the errors are reduced in the layer region as claimed in Theorem 4.9. The implementation of the proposed method is done in MATLAB R2015b (The Mathworks, Inc.). The program code is uploaded in GitHub and the URL of the source code is http://shorturl.at/gjlCS.

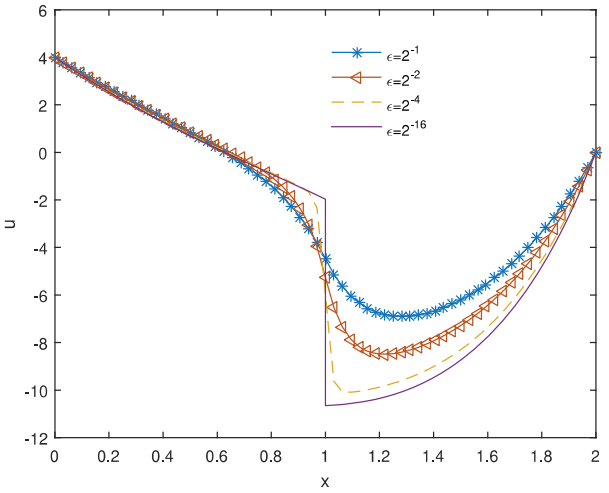



*Fig. 2. Numerical solution of Example 5.1 at $t = 2$ for different values of $\varepsilon$ when $N = 128$.*

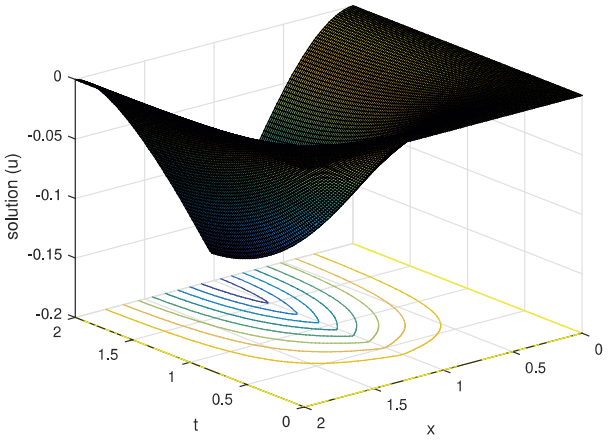



*Fig. 3. Numerical solution of Example 5.2 for $\varepsilon = 2^{-4}$ when $M = N = 128$.*

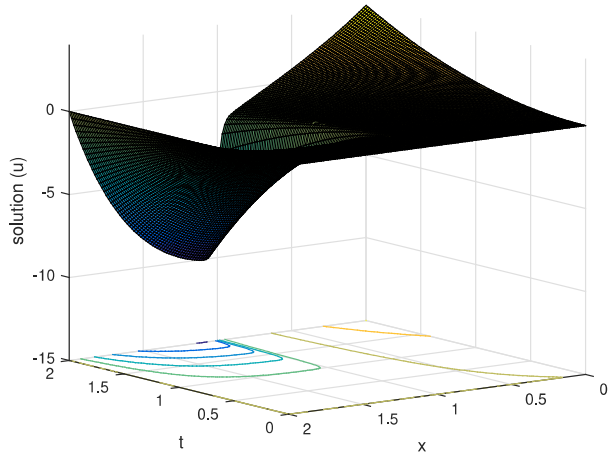



*Fig. 1. Numerical solution of Example 5.1 for $\varepsilon = 2^{-4}$ when $M = N = 128$.*

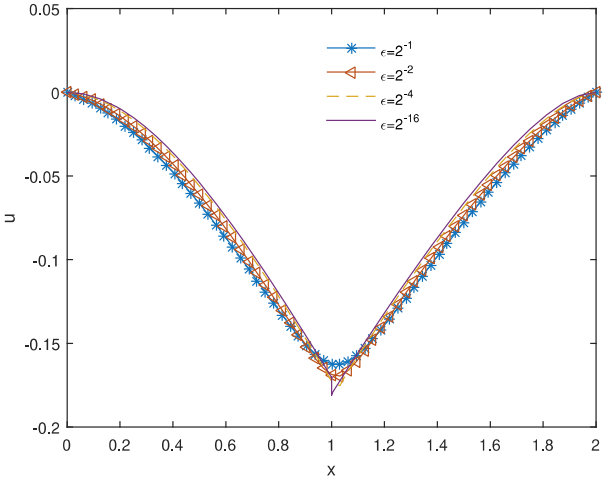



*Fig. 4. Numerical solution of Example 5.2 at $t = 2$ for different values of $\varepsilon$ when $N = 128$.*

*Table 1. Maximum absolute error and order of convergence for Example 5.1 for different values of ε, M and N when M = N.*

| $N$ | $\varepsilon = 2^{-2}$ | $2^{-4}$ | $2^{-6}$ | $2^{-8}$ | $2^{-10}$ | $2^{-12}$ |
|---|---|---|---|---|---|---|
| 32 | 1.099e-01 | 8.697e-01 | 8.723e-01 | 8.712e-01 | 8.709e-01 | 8.708e-01 |
| | 1.6274 | 1.5675 | 1.2466 | 1.2572 | 1.2601 | 1.2609 |
| 64 | 3.558e-02 | 2.934e-01 | 3.676e-01 | 3.644e-01 | 3.635e-01 | 3.633e-01 |
| | 1.3320 | 2.0478 | 1.5866 | 1.6018 | 1.6060 | 1.6071 |
| 128 | 1.413e-02 | 7.096e-02 | 1.223e-01 | 1.200e-01 | 1.194e-01 | 1.192e-01 |
| | 1.1428 | 1.9332 | 1.7309 | 1.7302 | 1.7308 | 1.7311 |
| 256 | 6.400e-03 | 1.858e-02 | 3.687e-02 | 3.619e-02 | 3.598e-02 | 3.592e-02 |
| | 1.0336 | 1.7107 | 1.6093 | 1.6294 | 1.6359 | 1.6376 |
| 512 | 3.126e-03 | 5.676e-03 | 1.208e-02 | 1.169e-02 | 1.157e-02 | 1.154e-02 |
| | 1.0066 | 1.5189 | 1.5926 | 1.6132 | 1.6207 | 1.6241 |

*Table 2. Maximum absolute error and order of convergence for Example 5.2 for different values of ε, M and N when M = N.*

| $N$ | $\varepsilon = 2^{-2}$ | $2^{-4}$ | $2^{-6}$ | $2^{-8}$ | $2^{-10}$ | $2^{-12}$ |
|---|---|---|---|---|---|---|
| 32 | 1.886e-03 | 4.150e-03 | 2.485e-03 | 2.146e-03 | 2.288e-03 | 2.323e-03 |
| | 1.4071 | 1.6239 | 1.1244 | 1.0741 | 1.0986 | 1.1044 |
| 64 | 7.113e-04 | 1.346e-03 | 1.139e-03 | 1.019e-03 | 1.068e-03 | 1.080e-03 |
| | 1.5199 | 1.6981 | 1.5245 | 1.3442 | 1.3740 | 1.3811 |
| 128 | 2.480e-04 | 4.149e-04 | 3.961e-04 | 4.015e-04 | 4.122e-04 | 4.149e-04 |
| | 1.5620 | 1.4068 | 1.5615 | 1.6111 | 1.6576 | 1.6714 |
| 256 | 8.400e-05 | 1.565e-04 | 1.342e-04 | 1.313e-04 | 1.304e-04 | 1.302e-04 |
| | 1.5102 | 1.5403 | 1.5137 | 1.6794 | 1.7007 | 1.7062 |
| 512 | 2.949e-05 | 5.380e-05 | 4.700e-05 | 4.101e-05 | 4.014e-05 | 3.991e-05 |
| | 1.4117 | 1.5972 | 1.4798 | 1.4726 | 1.4368 | 1.4272 |

## 6. Concluding remark

Singularly perturbed parabolic functional differential equation with discontinuous coefficient and source term is numerically solved. The problem's solution takes into account not just the present state of the physical system but also its history. The simultaneous presence of discontinuous data and delay makes the problem stiff. In the limiting case, the solution of the problem exhibits multi-scale character. There are narrow regions where solution derivatives grow exponentially and exhibit turning point behaviour, leading to sharp interior layers across discontinuities.

A hybrid numerical scheme composed of a central difference scheme in the layer region and a midpoint upwind scheme outside the layer region is used to discretize space variable over a specially generated mesh. Whereas an implicit finite difference scheme is used to discretize the time variable. The mesh has been chosen so that most of the mesh

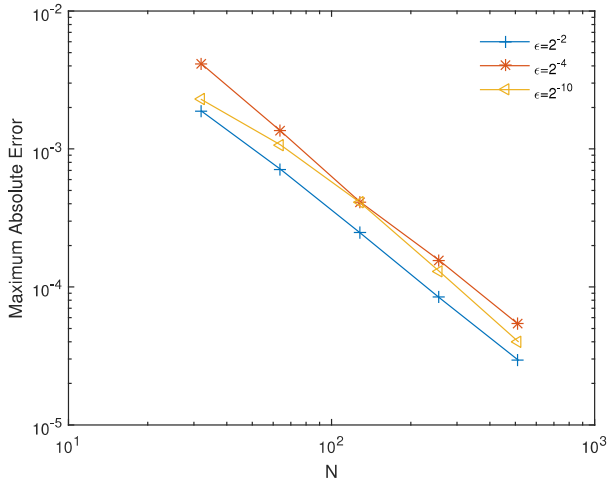



*Fig. 5. Error plot for Example 5.1.*

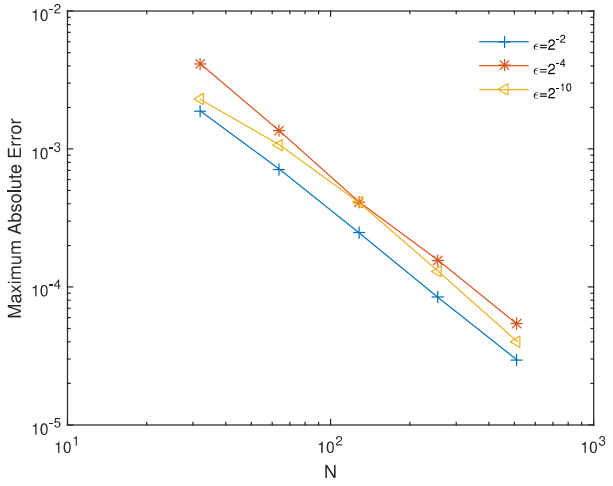



*Fig. 6. Error plot for Example 5.2.*

*Table 3. Maximum absolute error and order of convergence for Example 5.1 and 5.2 for different values of M and N when $M = N^2$ and $\varepsilon = 2^{-10}$.*

| For Example 5.1 | | | | For Example 5.2 | | |
|---|---|---|---|---|---|---|
| $N$ | *left region* $[0, 1-\sigma]$ | *interior layer region* $(1-\sigma, 1+\sigma)$ | *righ tregion* $[1+\sigma, 2]$ | *left region* $[0, 1-\sigma]$ | *interior layer region* $(1-\sigma, 1+\sigma)$ | *right region* $[1+\sigma, 2]$ |
| 32 | 7.602e-03 | 9.073e-01 | 2.237e-02 | 1.863e-04 | 2.215e-03 | 6.044e-05 |
| | 1.9665 | 1.3111 | 1.9344 | 1.9878 | 1.1088 | 1.8829 |
| 64 | 1.973e-03 | 3.656e-01 | 5.687e-03 | 4.652e-05 | 1.026e-03 | 1.632e-05 |
| | 1.9760 | 1.6430 | 1.9033 | 1.9624 | 1.3972 | 1.7645 |
| 128 | 5.083e-04 | 1.170e-01 | 1.478e-03 | 1.182e-04 | 3.895e-04 | 4.790e-06 |
| | 1.9592 | 1.7816 | 1.8513 | 1.9592 | 1.7332 | 1.5841 |
| 256 | 1.323e-04 | 3.404e-02 | 3.992e-04 | 3.064e-06 | 1.186e-04 | 1.594e-06 |
| | 1.9438 | 1.9236 | 1.8205 | 1.9382 | 1.8560 | 1.5632 |

*Table 4. Maximum absolute error and order of convergence for Example 5.2 for different values of M and x when $N = 512$ and $\varepsilon = 2^{-6}$.*

| $x$ | $M = 32$ | 64 | 128 | 256 | 512 |
|---|---|---|---|---|---|
| $x_{N/2+1}$ | 3.837e-04 | 2.058e-04 | 1.071e-04 | 5.486e-05 | 2.773e-05 |
| | 0.8988 | 0.9423 | 0.9651 | 0.9843 | 0.9981 |
| $x_{N/2+4}$ | 3.999e-04 | 2.122e-4 | 1.093e-04 | 6.530e-05 | 3.225e-05 |
| | 0.9142 | 0.9571 | 0.7432 | 1.0178 | 1.0184 |

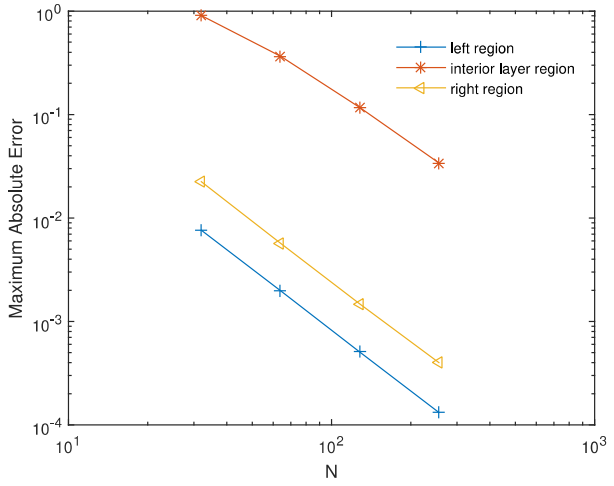



*Fig. 7. Error plot of the spatial order of convergence for example 5.1.*

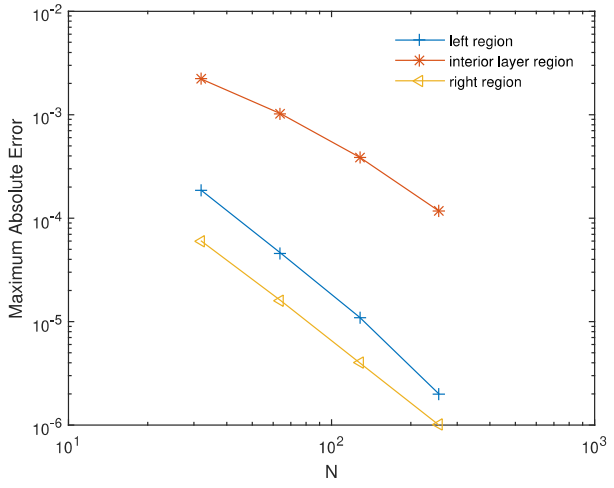



*Fig. 8. Error plot of the spatial order of convergence for example 5.2.*

points remain in the regions with rapid transitions. The proposed numerical method has been analyzed for consistency, stability and convergence. Extensive theoretical analysis is performed to obtain consistency and error estimates. It is found that the method proposed is unconditionally stable, and the convergence obtained is parameter uniform. Numerical illustrations are presented for two test examples that demonstrate the effectiveness of the technique. Convergence obtained in practical satisfies theoretical predictions.

## References

[1] der Heiden U, Walther H-O. Existence of chaos in control systems with delayed feedback. J Differ Equ 1983;47:273−95.
[2] Ansari AR, Bakr SA, Shishkin GI. A parameter robust finite difference method for singularly perturbed delay parabolic partial differential equations. J Comput Appl Math 2007;205: 552−66.
[3] Avudai-Selvi P, Ramanujam N. A parameter uniform difference scheme for singularly perturbed parabolic delay differential equation with Robin type boundary condition. Appl Math Comput 2017;296:101−15.
[4] Barbu L, Morosanu G. Singularly perturbed boundary value problems. Basel: Birkh a usar; 2007.
[5] Bashier EBM, Patidar KC. A fitted numerical method for a system of partial delay differential equations. Comput Math Appl 2011;61:1475−92.
[6] Bashier EBM, Patidar KC. A novel fitter operator finite difference method for a singularly perturbed delay parabolic partial differential equation. Appl Math Comput 2011;217: 4728−39.
[7] Bobisud L. Second-order linear parabolic equations with a small parameter. Arch Ration Mech Anal 1967;27:385−97.
[8] Clavero C, Gracia JL. A high order HODIE finite difference scheme for 1D parabolic singularly perturbed reaction diffusion problems. Appl Math Comput 2012;218:5067−80.
[9] Das A, Natesan S. Uniformly convergent hybrid numerical scheme for singularly perturbed delay parabolic convection

diffusion problems on Shishkin mesh. Appl Math Comput 2015;271:168−86.
[10] Erneux T. Applied delay differential equations. Surv. Tutor. Appl. Math. Sci. 2009;3. Springer, New York.
[11] Farrell PA, Hegarty AF, Miller JJH, O'Riordan E, Shishkin GI. Robust computational Techniques for boundary layers. Boca Raton: Chapman and Hall/CRC; 2000.
[12] Franz AL, Roy R, Shaw LB, Schwartz IB. Effect of multiple time delays on intensity fluctuation dynamics in fiber ring lasers. Phys Rev E 2008;78:16208.
[13] Gowrisankar S, Natesan S. A robust numerical scheme for singularly per-turbed delay parabolic initial boundary value problems on equidistributed grids. Electron Trans Numer Anal 2014;41:376−95.
[14] Gowrisankar S, Natesan S. $\varepsilon$-uniformly convergent numerical scheme for singularly perturbed delay parabolic partial differential equations. Int J Comput Math 2017;94:902−21.
[15] Gupta V, Kumar M, Kumar S. Higher order numerical approximation for time dependent singularly perturbed differential-difference convection-diffusion equations. Numer Methods Part Differ Equ 2018;34:357−80.
[16] Gurney WSC, Blythe SP, Nisbet RM. Nicholson's blowies revisited. Nature 1980;287:17−21.
[17] Hale JK, Lunel SMV. Introduction to functional differential equations. Berlin: Springer; 1993.
[18] Hemker PW, Shishkin GI, Shishkina LP. High-order time-accurate schemes for singularly perturbed parabolic convection-diffusion problems with robin boundary conditions. Comput Methods Appl Math 2002;2:3−25.
[19] Ikeda K. Multiple-valued stationary state and its stability of the transmitted light by a ring cavity system. Opt Commun 1979;30:257−61.
[20] Kaushik A, Sharma KK, Sharma M. A parameter uniform difference scheme for parabolic partial differential equation with a retarded argument. Appl Math Model 2010;34: 4232−42.
[21] Kaushik A and Sharma N. An adaptive difference scheme for parabolic delay differential equation with discontinuous coefficients and interior layers, J Differ Equ Appl 26 (2020), 11-12.
[22] Kumar K, Chakravarthy PP, Ramos H, Vigo-Aguiar J. A stable finite difference scheme and error estimates for parabolic singularly perturbed PDEs with shift parameters. J Comput Appl Math 2022;405:15. Paper No. 113050.
[23] Kumar M, Kumar S. High order robust approximations for singularly perturbed semilinear systems. Appl Math Model 2012;36:3570−9.
[24] Kumar M, Rao SCS. High order parameter robust numerical method for singularly perturbed reaction diffusion problems. Appl Math Comput 2010;216:1036−46.
[25] Kumar S, Kumar M. High order parameter uniform discretization for singularly perturbed parabolic partial differential equations with time delay. Comput Math Appl 2014;68:1355−67.
[26] Ladyzhenskaya OA, Solonnikov VA, Ural'ceva NN. Linear and quasilinear equations of parabolic type. Translations of mathematical monographs, 23. Providence, R.I.: American Mathematical Society; 1968. Translated from the Russian by S. Smith.
[27] Larger L, Penkovsky B, Maistrenko Y. Virtual chimera states for delayed- feedback systems. Phys Rev Lett 2013;111:54103.
[28] Linß T. Layer adapted meshes for reaction convection diffusion problems. Berlin: Springer; 2010.
[29] Mackey MC, Glass L. Oscillation and chaos in physiological control systems. Sci New Ser 1977;197:287−9.
[30] Marconi M, Javaloyes J, Barland S, Balle S, Giudici M. Vectorial dissipative solitons in vertical-cavity surface-emitting lasers with delays. Nat Photonics 2015;9:450−5.
[31] Marcus CM, Westervelt M. Stability of analog neural networks with delay. Phys Rev 1989;39:347−59.
[32] Miller JJH, O'Riordan E, Shishkin GI. Fitted numerical Methods for singular perturbation problems. Singapore: World Scientific; 2012.
[33] Miller JJH, O 'Riordan E, Shishkin GI, Shishkina LP. Fitted mesh methods for problems with parabolic boundary layers. Math Proc Roy Ir Acad 1998;98:173−90.
[34] Ramesh VP and Priyanga B. Higher order uniformly convergent numerical algorithm for time-dependent singularly perturbed differential-difference equations, Differ Equ Dyn Syst 29 (2021), 239-263.
[35] Roos HG, Stynes M, Tobiska L. Numerical methods for singularly perturbed differential equations: convection-diffusion and flow problems. New york: Springer- Verlag; 1996.
[36] Roos HG, Stynes M, Tobiska L. Robust numerical methods for singularly perturbed differential equations. Berlin: Springer; 2008.
[37] Scholl E, Schuster H. Handbook of chaos control. Weinheim: Wiley-VCH; 2008.
[38] Shivhare M, Chakravarthy PP, Ramos H, Vigo-Aguiar J. Quadratic b- spline collocation method for time dependent singularly perturbed differential- difference equation arising in the modeling of neuronal activity. Numer Methods Part Differ Equ 2021:1−22.
[39] Soriano MC, Garcia-Ojalvo J, Mirasso C, Fischer I. Complex photonics: Dynamics and applications of delay-coupled semiconductors lasers. Rev Mod Phys 2013;85:421−70.
[40] Stynes M, Roos HG. The midpoint upwind scheme. Appl Numer Math 1997;23:361−74.
[41] Stynes M, Tobiska L. The SDFEM for a convection diffusion problem with a boundary layer: Optimal error analysis and enhancement of accuracy. SIAM J Numer Anal 2003;41: 1620−42.
[42] Wu J. Theory and applications of partial functional differential equations. Berlin: Springer; 1996.

Journal of Scientific & Industrial Research
Vol. 81, September 2022, pp. 932-939
DOI: 10.56042/jsir.v81i09.59538

# A Multivariate Analysis of Agricultural Electronic Trading Adoption

Sanjay Chaudhary[1]* & P K Suri[2]

[1]Delhi Technological University, and Vivekananda Institute of Professional Studies, Pitampura, Delhi 110 034, India

[2]Delhi School of Management, Delhi Technological University, Delhi 110 032, India

*Received 21 January 2022; revised 16 August 2022; accepted 16 August 2022*

The study addresses the knowledge gap related to the scarce literature on digitalization in India's agricultural marketing. A field survey of five hundred National Agriculture Market users is undertaken to understand the theoretical constructs of wholesale electronic trading adoption in a realistic backdrop of a large digital project. The Partial Least Squares-Structural Equation Modelling (PLS-SEM) methodology is used for the statistical analysis. It demonstrates the positive effect of variables: 'Trust', 'Cost', 'Social Influence', 'Perception-Ease of Use', 'Perception-Usefulness', and 'Facilitating Conditions' on the adoption. The study brings out a simple agricultural wholesale e-trading adoption framework. It extends the existing theoretical knowledge base concerning technology adoption in new contexts (wholesale electronic trading, agriculture, India). It expands the scope of the theory by adding new constructs, 'Trust' and 'Cost'. The study's recommendations are expected to help practitioners in decision-making. It shall help practitioners of developing countries prioritize using scarce resources to deliver the intended benefits to the farming community in terms of administrative ease, user convenience, expanded market reach, faster cycle time,

**Keywords:** Digitalization, ICT, National agriculture market, PLS-SEM

## Introduction

The digitalization trend is evident in the agriculture sector. The interaction of fast-changing latest Information and Communication Technology (ICT) and the oldest profession (agriculture, traditional ecosystem) opens multiple issues. Digitalizing the business-to-business (B2B) procurement stage is essential to improving the agricultural supply chain.

### Digitalization in the Procurement

In the procurement stage, agricultural commodity sales are progressively changing from the verbal public sale (auction) mode to the digital e-trading mode in India. Wholesale e-trading platforms are more open and transparent than physical markets. E-trading has accelerated and expanded the procurement process. It is also expected to strengthen the seller's bargaining power by giving them online reach to intra-market, inter-market, and inter-state buyers. As a result, farmers are set to receive a higher price when selling produce to the markets with shortage/high demand through the online e-trading platform rather than through a limited number of regional agents at the farm gate.[1,2]

*Author for Correspondence
E-mail: one.sanjay@gmail.com

The research focus is derived from the empirical finding that the B2B e-commerce platforms such as electronic trading (e-trading) shall focus first on building a critical number of users (farmers, traders) and quality. As the number of users/transactions on the e-commerce platform grows to the critical number, the platform becomes viable and more valuable to other stakeholders. The other stakeholders include application developers, exporters, transporters, logistics and value-add service providers.[3]

Post literature review, the research gaps are identified as follows: Research on the agricultural B2B e-commerce (e-trading) adoption is scarce, the case studies on agricultural e-trading are few, and there is little research in India on agricultural e-trading benefits for farmers. By responding to the following queries, the research gaps can be filled: What are the driving factors behind adopting e-trading in the Indian agricultural sector? Using the learning from the case of the National Agriculture Market (eNAM)[4], how can the adoption of e-trading be promoted? These research questions led to the research objective of this study.

### Research Objectives

Consequently, the study has the following objectives:

- To validate the agricultural e-trading adoption framework in the National Agriculture Market context.
- Make suggestions for ways to increase wholesale agricultural e-trading adoption.

**Proposed Framework**

For validation, the proposed adoption framework for agricultural e-trading in India (Fig. 1) is generated using the Total Interpretive Structural Modelling (TISM) methodology.[5] It is the extended form of the Unified Theory of Acceptance and Use of Technology (UTAUT).[6,7]

The seven latent variables (Constructs) are construed as significant direct or indirect determinants of adoption in agricultural wholesale trading (Table 1). The constructs are considered, analyzed, and eventually managed in the agricultural sector adoption context.

The null and alternative hypotheses are conceptualized as follows:

*$H_{01}$: 'Perception -Usefulness' does not have an effect on the 'Behavioural Intention' to adopt the e-trading platform.*

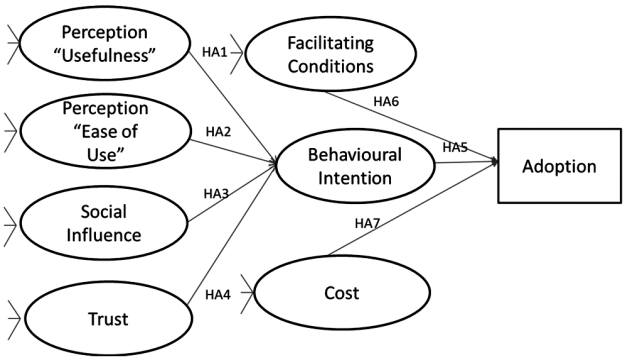



Fig. 1 — Proposed adoption framework

*$H_{A1}$: 'Perception - Usefulness' have an effect on the 'Behavioural Intention' to adopt the e-trading platform.*

Similarly, the null hypothesis $H_{0I}$ will follow the same form, where I = {2, 3, 4, 5, 6, and 7} for the null hypotheses against which the alternative hypotheses are:

*$H_{A2}$: 'Perception - Ease of Use' have an effect on the 'Behavioural Intention' to adopt the e-trading platform.*

*$H_{A3}$: 'Social Influence' have an effect on the 'Behavioural Intention' to adopt the e-trading platform.*

*$H_{A4}$: 'Trust' have an effect on the 'Behavioural Intention' to adopt the e-trading platform.*

*$H_{A5}$: 'Behavioural Intention' have an effect on the 'Adoption' of the e-trading platform.*

*$H_{A6}$: The 'Facilitating Conditions' have an effect on the 'Adoption' of the e-trading platform.*

*$H_{A7}$: 'Cost' has an effect on the 'Adoption' of the e-trading platform.*

## Materials and Methods

The Partial Least Squares-Structural Equation Modelling (PLS-SEM) method, a statistical analysis technique of the primary survey data, is used to numerically validate the proposed framework and confirm the key driver constructs for the dependent variable 'Adoption'.

PLS-SEM is preferred over Covariance-Based Structural Equation Modelling (CB-SEM) for data analysis for two reasons. First, the exploratory study aims to identify key driver constructs. Second, the data for the nominal dependent variable ('Adoption') are not normally distributed.[25,26]

Table 1 — Constructs in the research framework

| Code | Construct | In-Study Explanation | References |
|---|---|---|---|
| PU | Perception - Usefulness | eNAM provides users with benefits such as quality transparency, better pricing, and faster market transactions. | 6, 8, 9, 10, 11 |
| PEU | Perception - Ease of Use | eNAM is easy to learn and simple to use | 9, 10, 11, 12, 13 |
| SI | Social Influence | eNAM is endorsed and promoted by a community of acquaintances, pioneering farmers, affiliated traders, and community leaders. | 6, 11, 12, 14, 15 |
| T | Trust | User confidence in eNAM portal content, trade, and information. It is also a belief that the eNAM is credible. Farmers and other stakeholders stand to gain from the management's activities. | 12, 13, 16, 17, 18, 19 |
| C | Cost | Transaction Costs on the eNAM platform | 12, 19, 20, 21 |
| FC | Facilitating Conditions | The infrastructure, which includes electronic weighing, quality test laboratories, auction halls, logistics facilitation, training, and customer service, supports the use of eNAM. | 6, 11, 22, 23, 24 |
| BI | Behavioural Intention | The extent to which a farmer/trader has made a deliberate plan to e-trade. | 6, 7, 9, 22 |
| U | Adoption | Actual usage of eNAM | 6, 7, 9, 22 |

A case study of the eNAM wholesale e-trading platform provides practical context to validate the proposed framework's theoretical basis.[27] The eNAM is the government-supported de-facto pan-country platform that is much larger in scope and size than the platform launched by the corporates. It is expected to reach all 7,320 Agricultural Produce & Livestock Market Committee (APMC) market and sub-market yards over the next five years. It has registered 17.3 million farmers, 0.22 million traders, and two thousand one hundred forty Farmer Producer Organizations (FPOs), with cumulative trade reaching INR 1220 billion by February 2021.(4) The inter-market and inter-state trade of 175 commodities are possible on this platform. The recommendations to improve eNAM adoption are based on an understanding developed through the study and the interviews with experts.

Five APMC markets in two states, viz. Uttar Pradesh (Aligarh, Pilibhit, and Meerut) and Rajasthan (Nagar, Nadbai) were selected for primary data collection. These APMCs: experience many transactions, a considerable size, a large user base, and trade in similar commodities.

A structured questionnaire was developed in the local language, i.e., Hindi, with the help of a person with good knowledge of the language and agriculture. Two experts reviewed the draft questionnaire for language and content. The specific questions were modified as per the context of the study, keeping the previous research in view.[6,7] The final questionnaire measured 22 items on a 7-point Likert scale, where one implied 'No Influence' and seven meant 'Extreme Influence'. The questionnaire was subjected to internal consistency improvement and context-specific adjustments based on a pilot study involving 70 respondents.

The primary data was collected from the field survey with a sample size of 500 respondents. The sampling unit is the farmer registered on the eNAM portal at the APMC. The respondents are randomly selected via a random selection from the sampling frame (database in excel format) of registered farmers maintained by the five APMCs. It is expected that as a registered user of eNAM-enabled APMC, the respondent has an opinion about its functioning. The bias was eliminated, as all farmers had an equal chance of being chosen. The survey was conducted offline, either at the farmer's home or at the APMC location and over the course of one year starting in 2019. Although the process was time-consuming, it enjoyed a higher response rate from the rural population than surveys based on postal mail, telephone, or online mode.[28] The questionnaire responses were coded into Microsoft Excel and imported into SmartPLS 3.0 software for PLS-SEM analysis.

The sample size of 500 is adequate since SEM studies suggest a sample size of 200 as fair and 300 as acceptable. The G-Power software analysis recommended an appropriate sample size of 226.[25] Another suggestion is that the sample in multidimensional research is at least decuple the number of observed variables (survey items)[29,30], resulting in a reasonable sample size of 220.

## Results and Discussion

The PLS-SEM based statistical analysis of the proposed framework is conducted in two steps: measurement model validation and structural model validation.

### Demography and Data Characteristics

The 500 questionnaires were complete and considered for further analysis. The respondents were farmers registered at the following Agricultural Produce Market Committees (APMCs): Meerut, Aligarh, Nagar, Nadbai, and Pilibhit. Of 500 respondents, 42.6% dealt with vegetables, 29.4% dealt with wheat, 18% dealt with paddies, 6 dealt with mustard and 4% dealt with maize. The case study of the eNAM provides the survey data used to test the theory.

The model is recursive with no circular feedback loops. The observed independent variables are normally distributed. The model measures are reflective, where each of the six determinant latent constructs and one mediating determinant latent construct is made up of three observed independent variables. To ensure accuracy, the data have no missing values or outliers. The sample size of 500 is sufficient to give reliable PLS estimates.

### Measurement Model Validation

The measurement model is assessed using PLS-SEM best practices. All the observed variables' item loadings are greater than 0.71, showing a good contribution to the assigned construct. It provides acceptable construct reliability. Because all the Composite Reliability (CR) values are between 0.80 and 0.90, the internal consistency reliability is determined to be good. Cronbach's alpha (α) values

are greater than 0.70, thus re-determining internal consistency reliability to be good. The Average Variance Extracted (AVE) checks for convergent validity. All the AVE values are greater than 0.50; thus, convergent validity is established.[31,32] Table 2 displays the reliability and validity assessment values.

The Fornell and Larcker (FL) criterion and the Heterotrait-Monotrait (HTMT) criterion are used to ascertain discriminant validity.[33,34] As each construct's AVE is greater than the squared inter-construct correlation of that same construct and all other reflectively measured constructs, the FL criterion is satisfied. The HTMT criteria are also met, as most of the calculated values in Table 3 are less than the recommended value of 0.9, indicating sufficient discriminant validity. A few exceptions are noted, as the respondents are farmers (at varying literacy levels) and, as such, might have faced problems in differentiating between the constructs.

**Structural Model Validation**

The structural model (Fig. 2) emphasizes the interrelationship of the research framework's constructs. The association is tested using path coefficients (β) and t-statistics.

Most of the Variance Inflation Factor (VIF) values are less than three, and almost all are less than 5. Thus, the collinearity between the predictor constructs does not bias the regression results. The dependent constructs' 'Behavioural Intention' (BI) lowest coefficient of determination ($R^2$) value is 0.67, a near substantial value in behavioural sciences research. The value of $R^2$ for the primary dependent variable

Table 2 — Reliability and validity assessment

| Construct | Observed Variable (Item) | Item Loading | Composite Reliability (CR) | Cronbach's Alpha (α) | Average Variance Extracted (AVE) |
|---|---|---|---|---|---|
| Perception – Usefulness (PU) | Useful in Trade (PU1) | 0.94 | 0.9 | 0.9 | 0.85 |
| | Accomplish Task Quickly (PU2) | 0.93 | | | |
| | Price Increase (PU3) | 0.89 | | | |
| Perception – Ease of use (PEU) | Good User Interface (PEU1) | 0.89 | 0.9 | 0.85 | 0.78 |
| | Easy to use (PEU2) | 0.88 | | | |
| | Easy to Learn (PEU3) | 0.89 | | | |
| Social Influence (SI) | Influencers (SI1) | 0.88 | 0.87 | 0.78 | 0.69 |
| | Helpful Management (SI2) | 0.80 | | | |
| | Organization support (SI3) | 0.82 | | | |
| Facilitating Condition (FC) | Infrastructure (FC1) | 0.91 | 0.9 | 0.86 | 0.78 |
| | Training (FC2) | 0.89 | | | |
| | Support (FC3) | 0.87 | | | |
| Trust (T) | Accurate Information (T1) | 0.93 | 0.9 | 0.9 | 0.84 |
| | Trust in Seller (T2) | 0.92 | | | |
| | Trust in Buyer (T3) | 0.91 | | | |
| Cost ( C) | Transaction cost (C1) | 0.87 | 0.86 | 0.76 | 0.67 |
| | Value for Money (C2) | 0.83 | | | |
| | Overall Cost (C3) | 0.76 | | | |
| Behavioural Intention (BI) | Intend - with 3 months (BI1) | 0.82 | 0.88 | 0.79 | 0.7 |
| | Intend -within 1 year (BI2) | 0.90 | | | |
| | Intend – no time-period specified (BI3) | 0.81 | | | |

Table 3 — Discriminant validity

| | Fornell and Larcker Criteria | | | | | | | HTMT Criteria | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | BI | C | FC | PEU | PU | SI | T | BI | C | FC | PEU | PU | SI | T |
| BI | 0.84 | | | | | | | | | | | | | |
| C | 0.86 | 0.82 | | | | | | 1 | | | | | | |
| FC | 0.76 | 0.78 | 0.89 | | | | | 0.9 | 0.9 | | | | | |
| PEU | 0.71 | 0.76 | 0.80 | 0.88 | | | | 0.8 | 0.9 | 0.9 | | | | |
| PU | 0.73 | 0.77 | 0.79 | 0.85 | 0.92 | | | 0.8 | 0.9 | 0.8 | 0.9 | | | |
| SI | 0.70 | 0.75 | 0.77 | 0.80 | 0.78 | 0.83 | | 0.8 | 0.9 | 0.9 | 0.9 | 0.9 | | |
| T | 0.79 | 0.86 | 0.79 | 0.76 | 0.80 | 0.74 | 0.92 | 0.9 | 1 | 0.8 | 0.8 | 0.8 | 0.8 | |
| U | 0.79 | 0.81 | 0.70 | 0.73 | 0.72 | 0.68 | 0.71 | 0.8 | 0.9 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 |

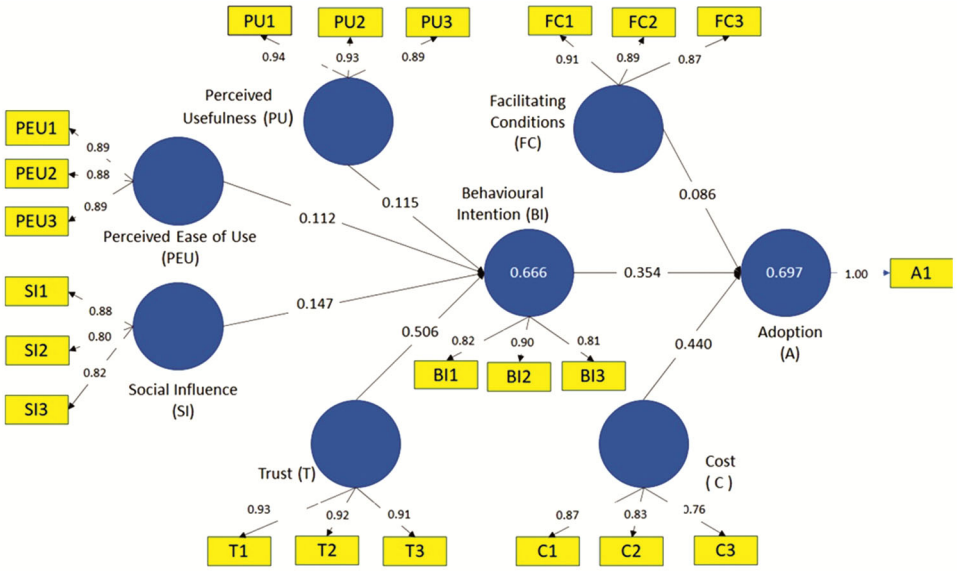



Fig. 2 — Structural model of adoption framework

***Abbreviations***: *Perceived Usefulness (PU), Perceived Ease of Use (PEU), Social Influence (SI), Trust (T), Cost (C), Facilitating Conditions (FC), Behavioural Intention (BI), Adoption (U).*

'Adoption' (U) is higher at 0.70, indicating that the model accounts for a large proportion of the dependent variable variance. The structural model has a near substantial in-sample explanatory power.[35] The structural equation's $R^2$ values and PLS-SEM fit indices are presented in Table 4.

The Non-Fuzzy Index (NFI) value is 0.81, and the Standardized Root-Mean-Square residual (SRMR) value was 0.064. The model for the adoption framework is a good fit since the SRMR value is less than 0.08 and NFI is 0.8.[36]

The PLS-SEM results for the structural model of the adoption framework (Fig. 2) are presented in Table 5. For significant P values < 0.05, all the hypotheses (Table 5) are supported. The path coefficients are also given in Table 5.

The relationship between the constructs is significant, as all the t values are more than 1.96.

The Stone-Geisser criterion ($Q^2$) is used to evaluate the model's predictive relevance. It assesses the model's reconstruction of observed values and parameter estimates. The model must have a $Q^2$ greater than zero to be of predictive relevance. $Q^2$ values ('Adoption' at 0.66, 'Behavioural Intention' at 0.66) are above the zero-threshold value, indicating that the PLS path model for the adoption framework has high predictive accuracy.[18]

As for the $f^2$ effect sizes, the construct 'Cost' has a moderate effect on the 'Adoption', whereas the 'Trust' has a moderate effect on the 'Behavioral Intention' since the $f^2$ size effect is more than 0.15 but less than 0.35. Rest other constructs have a small effect since the $F^2$ effect size is less than 0.15. The 'Behavioural Intention' partially mediates the relationship between the constructs 'Social Influence' - 'Adoption' and 'Trust' - 'Adoption' with a p-value less than 0.05 and the indirect effects 95% boot confidence interval bias-corrected does not straddle a zero in between lower limit and upper limit[37], for other constructs 'Perception – Ease of use', and 'Perception – Usefulness', the effect pass through 'behavioural intention' and there is full mediation.

Table 4 — PLS-SEM fit indices

| Model Elements | Values |
|---|---|
| R-square for Behavioural Intention BI) | 0.67 |
| R-square for Adoption (U) | 0.7 |
| Chi-square | 1991 |
| SRMR | 0.064 |
| NFI | 0.8 |

**Discussion**

Concerning the first research objective, the adoption framework is successfully tested using a multistage analysis and found suitable.

***Observations on the framework***

The construct's manifestation in the respondent sample group of farmers differs from other Industry

Table 5 — Structural results

| Hypotheses | Notation | Path Coefficients | t-Values | P-Values | Remarks |
|---|---|---|---|---|---|
| $H_{A1}$: 'Perception - Usefulness' have an effect on the 'Behavioural Intention' to adopt an e-trading platform. | PU→BI | 0.119 | 2.016 | 0.044 | Supported* |
| $H_{A2}$: 'Perception - Ease of Use' have an effect on the 'Behavioural Intention' to adopt the e-trading platform | PEU→BI | 0.111 | 2.168 | 0.031 | Supported* |
| $H_{A3}$: 'Social Influence' have an effect on the 'Behavioural Intention' to adopt the e-trading platform. | SI→BI | 0.145 | 3.134 | 0.002 | Supported* |
| $H_{A4}$: 'Trust' has an effect on the 'Behavioural Intention' to adopt the e-trading platform. | T→BI | 0.504 | 9.610 | 0.000 | Supported* |
| $H_{A5}$: 'Behavioural Intention' has an effect on the 'Adoption' of the e-trading platform. | BI→U | 0.348 | 6.301 | 0.000 | Supported* |
| $H_{A6}$: The 'Facilitating Conditions' have an effect on the 'Adoption' of the e-trading platform. | FC→U | 0.088 | 2.110 | 0.035 | Supported* |
| $H_{A7}$: 'Cost' has an effect on the 'Adoption' of the e-trading platform. | C→U | 0.443 | 9.271 | 0.000 | Supported* |

*Significant at P-values < 0.05 and t-values > 1.96 (at the 5% significance level).

groups. The adoption framework is distinct as, unlike other sectors, the agriculture sector is characterized by a lack of ecosystem enabling infrastructure, low education level of farmers, and low digital/ information technology awareness among the rural population. The validated adoption framework adds to the existing knowledge base by redefining constructs, adding two new constructs ('Trust' and 'Cost') to the UTAUT, and altering the strength of relationships between constructs.

The model accounts for 70% of the variance for the dependent variable 'Adoption'. The variance explained is better than most other competing models. The variance explained for the dependent variable in the competing models are Innovation Diffusion Theory (IDT) - 40%, Technology Acceptance Model (TAM) - 53%, Theory of Planned Behaviour (TPB) - 36%, and Theory of Reasoned Action (TRA) - 36%. Comparatively, the Unified Theory of Acceptance and Use of Technology (UTAUT) explains about 70% of the variance (adjusted R2) in usage intention[6] with four direct determinants latent construct, that is, two less determinant latent constructs used in this study. However, more studies based on the validated framework of this study may be conducted.

The collation of discussions with experts and the study analysis resulted in the following strategic and tactical recommendations in connection with the second research objective.

'*Perception – Usefulness'* may be improved by unifying the state and national markets in a single legal framework. The full benefits of the digital platform may be realized in inter-market and interstate market transactions. Such Intermarket trade is to be streamlined with the disagreement resolution at the trade-originating APMC. Farmers are looking for a quick transaction cycle and a favourable price impact. The spread of mobile applications and prompt digital payment is expected to be a positive step. *'Perception - Ease of Use'* may be enhanced by making the e-trading portal/website user-friendly and providing multilingual content. Users like the mobile app, but it has to improve its interaction with other government and partner applications.

Local agri-community active participation may facilitate the improvement of '*Trust*.' Regular community awareness camps and FPOs involvement may help in this regard. The '*Facilitating Conditions*' too need improvement. Quality test labs, commodity parameters, robust mechanisms, public computer terminals, electronic weighing machines, and trade hall with Wi-Fi Internet connectivity are available in APMCs, but under suboptimal or bare essential conditions. Storage and logistics facilities may be enhanced by the involvement of Indian Railway freight discounting. There is, as of now, no periodic appraisal of skills and competency. The implementing agency may choose to build user capacity through regular online/offline training. According to experts, rather than relying on the existing APMC management and staff, a dedicated contact centre and on-the-ground eNAM staff may improve customer care, problem resolution, and conflict management.

Increasing influencer farmer and lead trader participation in eNAM may improve *'Social Influence*'. The scope of offerings and range of services on the eNAM platform may be widened, e.g.,

farm equipment trade/rent. The eNAM platform is based on an open technical standard. The scope of offerings and range of services may be widened, identical to the Aadhar/UPI scheme. It may release its open Application Programmable Interface (API) to the public. The open API will make it easily discoverable and interoperable with numerous related applications. From a social welfare perspective, it is wise to open a more significant portion of the eNAM platform and open intellectual property rights to the public.[38]

The lower the transaction '*Cost*', the higher the '*Adoption*' of the eNAM wholesale e-trading platform. The inequity between taxes and market fees across states impedes inter-state agricultural produce transactions. Some uniformity across APMCs in different forms may result in greater user adoption of interstate trade, which has been launched on the e-trading platform. Furthermore, to compete with commission agents' informal credit system and cartels, registered farmers may be given favourable bank credit terms, access to innovations like warehouse receipt, and lifecycle-linked inter-agency credit recovery. The traders may be given flexible credit terms for buying and selling on eNAM.

## Conclusions

In the Indian agriculture sector, eNAM is a high-priority strategic intervention. The adoption framework proposed and validated in this article may aid in adopting wholesale e-trading through improved organizational readiness and recommended steps to strengthen adoption enablers among farmers. The 'Trust', 'Cost', and 'Social Influence' are identified as the significant constructs and enablers of adopting the Wholesale e-trading initiative (eNAM). Other enablers have been identified as 'Perception-Ease of Use', 'Perception-Usefulness', and 'Facilitating Conditions'. The study's recommendations are expected to help practitioners effectively deliver intended benefits to the farming community in terms of administrative ease, reduction in transaction cost, quick trade cycle, and better price realization. The theoretical details and the knowledge base generated in this study are expected to be beneficial to the researchers. The results may be refined and enriched based on learning from further studies in this direction.

Digitalization via e-trading will transform the existing agricultural marketing and supply chain activities approaches. e.g., collection, grading, trading, storage, packaging, and transport. It is expected to aid in resolving the issue of fragmented and inefficient agrarian supply chains and facilitate the ease of business and socio-economic growth of the farming community.

## References


1 Banker R, Mitra S & Sambamurthy V, The effects of digital trading platforms on commodity prices in agricultural Supply chains, *MIS Q*, **35(3)** (2011) 6–34.

2 Srinivasan G, *Quantitative models in operations and supply chain management* (PHI Learning, Delhi) 2018.

3 Alstyne M V & Parker G, Platform Business: From Resources to Relationships, *NIM Marketing Intelligence Review*, **9(1)** (2017) 24–29.

4 Department of Agriculture Cooperation and Farmers Welfare, NAM Guidelines, 2021. www.enam.gov.in/NAM/home/namguidelines (4 January 2022)

5 Chaudhary S & Suri P K, Modelling the Enablers of e-Trading Adoption in Agricultural Marketing: A TISM-Based Analysis of eNAM, *Vision,* **26(1)** (2022) 65–79.

6 Venkatesh V, Morris M G, Davis G B & Davis F D, User acceptance of information technology: Toward a unified view, *MIS Q*, **27(3)** (2003) 425–478.

7 Venkatesh V, Thong J & Xu X, Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology, *MIS Q*, **36(1)** (2012) 157–178.

8 Compeau D R, Higgins C A & Huff S, Social Cognitive Theory, and Individual Reactions to Computing Technology: A Longitudinal Study, *MIS Q*, **23(2)** (1999) 145–158.

9 Davis F D, Bagozzi R P & Warshaw P R, User Acceptance of Computer Technology: A Comparison of Two Theoretical Models, *Manag Sci*, **35(8)** (1989) 982–1002.

10 Molina-Maturano J, Verhulst N, Tur-Cardona J, Güereña D T, Gardeazábal-Monsalve A, Govaerts B & Speelman S, Understanding Smallholder Farmers' Intention to Adopt Agricultural Apps: The Role of Mastery Approach and Innovation Hubs in Mexico, *Agronomy*, **11(2)** (2021) 194.

11 Moore G C & Benbasat I, Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation, *Inform Syst Res*, **2(3)** (1991) 192–222.

12 Chaudhary S & Suri P K, Framework for agricultural e-trading platform adoption using neural networks, *Int j inf tecnol*, **13** (2021) 501–510.

13 Sui X & Geng X, Continuous usage intention to e-transaction cards in wholesale markets of agriproducts: empirical evidence from China, *Future Business Journal*, **7** (2021) 224–237.

14 Mathieson K, Predicting user intentions: Comparing the technology of planned behaviour, *Inform Syst Res*, **2(3)** (1991) 173–191.

15 Mzomwe Y M, Mariam A T, Margareth A M & Mubarack H K, ICT, and marketing for agricultural products: determinants of mobile phone usage to small scale orange farmers in Tanzania, *Business Education Journal*, **10(1)** (2021) 5–18.

16 Casalo L V, Flavian C & Guinaliu, The generation of trust in online services and product distribution: the case of Spanish electronic commerce, *J Electron Commerce Res*, **12(3)** (2011) 199–215.

17 Kuttainen C, *the Role of Trust in B2B Electronic Commerce - Evidence from Two e-Marketplaces*, Ph D Thesis, The Luleå University of Technology, Sweden, 2005.

18 Ramesh A, Banwet D K & Shankar R, Modelling the barriers of supply chain collaboration, *J Model Manag*, **5(2)** (2010) 176–193

19 Samant S & Dey K, Blockchain technology adoption, architecture, and sustainable agri-food supply chains, *J Clean Prod*, **284** (2021) 124731.

20 Clasen M & Mueller R A E, Success Factors of Agribusiness Digital Marketplaces, *Electron Market*, **16(4)** (2006) 349-360.

21 Garicano Luis & Kaplan S N, The Effects of Business-to-Business E-Commerce on Transaction Costs, *J Ind Econ*, **49(4)** (2001) 463–485.

22 Ajzen I, The Theory of Planned Behaviour, *Organ Behav Hum Decis Process*, **50(2)** (1991) 179–211.

23 Kilpatrick J & Factor R, Logistics in Canada survey: tracking the year 2000 supply chain issues and trends, *Materials Management and Distribution*, **45(1)** (2000) 16–20.

24 Tomar G, Chauhan G S & Panigrahi P K, Feasibility of m-governance in agriculture: insights from a multimodal study in rural India, *Transform Gov: People Process Policy*, **10(3)** (2016) 434–456.

25 Hair J F, Hult G T M, Ringle C M & Sarstedt M A, *Primer on Partial Least Squares Structural Equation Modelling* (Sage, California) 2016.

26 Rožman M, Tominc P & Milfelner B, A Comparative Study Using Two SEM Techniques on Different Samples Sizes for Determining Factors of Older Employees Motivation and Satisfaction, *Sustainability*, **12** (2020) 2189.

27 De Reuver M, Sørensen C & Basole R C, The digital platform: A research agenda*, J Inform Tech*, **33(2)** (2017) 124–135.

28 Malhotra N K, *Marketing Research: An Applied Orientation* (Pearson Education, Noida) 2021.

29 Hair J F, Anderson R E, Tatham R L & William C B, *Multivariate Data Analysis* (Prentice-Hall, New Jersey) 1998.

30 Kline R B, *Principles and Practice of Structural Equation Modelling* (Guilford Publications, New York) 2015.

31 Diamantopoulos A, Sarstedt M, Fuchs C, Wilczynski P & Kaiser S, Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective, *J Acad Market Sci*, **40(3)** (2012) 434–449.

32 Hair J F, Risher J J, Sarstedt M & Ringle C M, When to use and how to report the results of PLS-SEM, *Eur Bus Rev*, **31(1)** (2019) 2–24.

33 Fornell C G & Larcker D F, Evaluating structural equation models with unobservable variables and measurement error, *J Market Res*, **18(1)** (1981) 39–50.

34 Henseler J, Ringle C M & Sarstedt M, A new criterion for assessing discriminant validity in variance-based structural equation modelling, *J Acad Market Sci*, **43(1)** (2015) 115–135.

35 Hair J F, Ringle C M & Sarstedt M, PLS-SEM: Indeed, a silver bullet, *J Mark Theory Pract*, **19(2)** (2011) 139–151.

36 Ringle CM, *Advanced PLS-SEM Topics: PLS Multigroup Analysis*, Working paper, The University of Seville, Spain, 2016.

37 Preacher K J, & Hayes A F, Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models, *Behav Res Meth*, **40** (2008) 879–891.

38 Parker G & Alstyne M V, Innovation, Openness, and Platform Control, *Manag Sci*, **64(7)** (2018) 3015–3032.

*2022 2nd International Conference on Intelligent Technologies (CONIT)*
*Karnataka, India. June 24-26, 2022*

# A Novel Graphical Password Scheme to Avoid Shoulder-Surfing Attacks in Android Devices

Mayank Kaushik
*Department of Information Technology*
*Delhi Technological University*
New Delhi, India
mayankkaushik_2k18it070@dtu.ac.in

Ashish Rawat
*Department of Information Technology*
*Delhi Technological University*
New Delhi, India
ashishrawat_2k18it034@dtu.ac.in

Varsha Sisaudia
*Department of Information Technology*
*Delhi Technological University*
New Delhi, India
varsha@dtu.ac.in

Love Parashar
*Department of Information Technology*
*Delhi Technological University*
New Delhi, India
loveparashar_2k18it068@dtu.ac.in

***Abstract*—Shoulder surfing is an easy attack, where an unauthorized user tries to gain access to sensitive information. Unfortunately, none of the currently available content-based shoulder-surfing approaches for protected graphical concealed keywords are both trustworthy and professional. This paper proposes an authentication technique that operates on touchscreen devices with an aim to solve the problem of shoulder-surfing assaults over validation. Graphical passwords are required to get access to the system and to assist with its security. A graphical password is a type of password that can be used instead of a textual password to gain access to a system. It is safer than a written password, which is vulnerable to a shoulder-surfing assault. In this study, we used a falsification technique to create a system that is immune to shoulder-surfing assaults. As a result, the hacker will be perplexed when attempting to obtain and misuse the password. Using Android Studio, the proposed scheme was developed into an android application composed of shoulder-surfing resistant user registration and login process. When compared with other graphical password techniques based on images, the proposed the technique of graphical authentication mechanism with numbers/digits is more easy to memorize and has less cognitive strain on the user. Our graphical authentication technique has more complexity and larger password space than the conventional numeric password schemes, making it a superior method for safeguarding against shoulder surfing attacks. To meet customer expectations, this project seeks to design a graphical password by combining processes such as long key press, and acceleration detection mechanism of Android devices.**

***Keywords— Graphical password, shoulder surfing, authentication, login phase, registration phase***

## I. INTRODUCTION

Information technology has progressed to the point that it now makes information available to anybody who has access to the internet infrastructure. With such widespread access to information that may be sensitive or pecuniary in nature, it is preferable that access be limited to only authorised individuals. Such information should not be accessible to any unauthorized parties or malevolent users. Although not all information is confined to a certain set of users and can be made public for everybody's use, there exists some sensitive data that should be known to only intended users. Authentication is one of the most important safeguards for information confidentiality, since it limits access to only authorised individuals [1]. Textual passwords are the most basic type of authentication.

Passwords are used by us daily to protect our personal and private information from attackers or unwanted users [2]. Alphanumeric passwords are the most well-known password scheme currently used by the majority of people. As technology advances, new methods of breaking in or leaking out outdated alphanumeric passwords are being developed, and that too quite frequently. As a result, a more distinctive password type is introduced in our approach to increase the efficiency of the password in safeguarding a user's personal or private information but also to modify how an individual understands the term password, which is called as the graphical password strategy.

Text-based password schemes are vulnerable to a variety of assaults, including dictionary and shoulder surfing attacks. The need to address the integrity of such text-based systems has expanded fast as more users and organizations have become aware of possible assaults. While retaining memorability and security, password - based authentication schemes should encourage users to use passwords that are less predicted and much more secure.

User authentication [3] is most typically done with alphanumeric passwords and usernames. One disadvantage of using such a design is that passwords can be easily learned. It's fairly unusual for people to overlook the fact that if a password is difficult to guess, it's also difficult to remember. Dictionary attacks may also be used on alphanumeric passwords. Because it's difficult to recall random sequences of letters, most people choose for a common phrase or a username, not realising that their credentials may be brute-forced in a short period of time.

Humans are good at remembering pictures/visuals than remembering words or passwords. Because of this, a feasible alternative to text-based techniques has been proposed: image based password systems. Images are used regardless of alpha-numeric characters/symbols in graphical password schemes. To login successfully, the user needs to remember a collection of photos. The apparent disadvantage of such a technique is that it requires a large collection of such authentic images to be maintained in memory, although if the quantity of assets photos are high enough, the probable password void of a graphical based password should be resistant to dictionary attacks [4]. Another issue is that shoulder surfing assaults are common with these graphical

2022 International Conference on Intelligent Technologies (CONIT) | 978-1-6654-8407-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CONIT55038.2022.9848122



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:59:03 UTC from IEEE Xplore. Restrictions apply.

schemes. An adversary who is shoulder surfing attempts to predict the password by closely inspecting the customer's login window. An adversary will remember a password that is very easy to recall for an authentic user.

Shoulder surfing [2] is a commonly used attack to breach computer security by merely observing another person's inputs to an authentication mechanism such as a personal laptop, ATM, or other similar device. Thus the shoulder surfer can now access personal information of the attacked user and can pose a serious threat by misusing the gained information. For example, with popularity of digital payments and availability of unified payments interface (UPI), people have started using mobile applications for making payments in crowded market places as well. While using a phone to make UPI payment, the users need to enter UPI pin to complete the payment. Anyone with a bad intention, may observe and track your hand movements while you enter a pin for such interfaces to complete the transaction. Thus, people nearby can intentionally or unintentionally know your sensitive information. Today mobile phone users are rapidly increasing and so are the UPI payments, thus it is very easy to spy on a mobile phone and with the increase in the number of users using UPI payment, some sensitive information leaks like UPI pin can be very risky.

Nowadays, hackers are also using a binocular or some vision-enhancing devices to spy even from a long distance. Researchers have been working to make an anti-shoulder surfing authentication system. As a result, alphanumeric passwords have been phased out in favour of graphical passwords, as the human brain is better at remembering visuals rather than long texts. Another benefit of graphical passwords is that they prohibit passwords from being stolen if a keyboard logger, such as malicious programs (Rootkit), is loaded by an attacker to record text-based credentials.

As a result, it's reasonable to argue that shoulder surfing assaults constitute a significant threat to our fast growing world of electronic gadgets and we are definitely in need of a robust mechanism to counter these attacks. Any new technique can be called a good one if it is more secure and understandable to the users and have passwords that are easy to remember. Hence we propose a novel technique that can minimize these attacks on android devices.

## II. Related Work

Today, the most common method of authenticating a user to grant access to computer systems is through the use of a password. Traditional text-based password schemes, on one hand are easy to remember, but are well-known to be vulnerable to shoulder-surfing and key logger attacks. Various solutions have been proposed in order to solve this problem. Many graphical password methods have also been developed with the goal of saving memory, working quickly, and reduce vulnerability towards shoulder surfing assaults.

Dhamija and Perrig [5] presented one of the first graphical authentication systems. Their method is relied on hash ecosystem, which requires the host to maintain the rank of all users' photos in plaintext. Furthermore, selecting random photos from a large database is a time-consuming procedure, and strategies for effectively storing and selecting images. According to Sobrado and Briget [6], 4 pass-objects are situated near the exhibited items and get employed to build a crossing point that is not visible. The user was told to choose from four images of human faces via a technique dubbed "Passface" [7]. The user knows the face and clicks anywhere on it. This strategy is easily remembered for lengthy periods of time. However, an authentic predisposition for pick images based on ethnicity, gender or traditional tendencies is a key downside of such a strategy. As a result, the password becomes relatively predictable.

Brostoff and Sasse [7] conducted a research that found that consumers were less likely to use the Passface algorithm since it required lengthier time for verification as compared to text based procedures. Another drawback of the method is that it cannot be utilised by those who are visually impaired. The Jansen et al. [8-9] method was created with mobile devices in mind. During registration, a person chooses a topic (for example, sea, city, etc.) from which thumbnail photographs are displayed. At the time of authentication, the user should input the prefaced sequence of pictures beneath the registered theme with the use of a stylus, which will create a password to be authorised.

An individual picks certain photographs that are marked as "pass images" with the model using a method developed by Koike and Takada [10]. If the user-selected photos are not displayed in the grid, the user must choose "no pass-image," that may be used to create fake pictures This level of authentication is repeated numerous times., as determined by the Awase-E system at random [11].

According to Davis's research [12], users' preferred photos are fairly predictable. Davis et al. suggested a narrative system that allows a user to create a tale by selecting a password from a combination of nine categories. The categories should be different, and they should be taken from collection which portray a daily life, like eateries, automobiles, animals, and so on. The user must choose from a set of photographs from which he or she can choose throughout the registration time.

Another approach was presented by Man et al., [13] to handle the shoulder surfing problem as well. Each pass-object in this algorithm has multiple variations, each with its own different code. At the time of authentication, the end-user would submit his password as a string of code, as well as a signal indicating the pass-objects' relative placement. The technique, on the other hand, needs the end-user to memorise the numeric code at each object variant. The technique got improved by allowing the end-user to select an objects variant pass code. Hong et al. employed a grid-based picture arrangement to build up the photographs, and the programme displays the randomly produced images on the screen each time during the authentication procedure [11]. Main disadvantage of Hong et al. technique .'s was: it takes lot of time, making it unsuitable for many users.

N. Wakabayashi et al. [14] conducted some recent study in which the user utilises a secret position in conjunction with hidden key for images to match the colour. [15] proposes other way for advanced graphical passwords. The approach converts passwords chosen by the user into drawings, making them resistant to shoulder surfing assaults. For the usage of a PDA, [16] suggests a mix of recognition and recollection based approaches. 25 photos are used in the registration phase, and 30 images are used in the recognition phase. Such solutions have the disadvantage of having big databases and a crowded grid, a legitimate user's ability to manage such a big visual area is hampered.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:59:03 UTC from IEEE Xplore. Restrictions apply.

Higashikawa et al. [17] suggested a shoulder-surfing-resistant authentication mechanism based on the Pattern Lock pass pattern, which employs a problem and response authentication method as well as the short-term memory of the users. Mohammedi et al. [18] also suggested a safe and lightweight remote health care authentication technique for use in portable healthcare systems.

Other recent developments in the field of graphical user authentication have also been completed. GamePass, a gamified variant of the graphical password proposed by the authors in [19], has been tested to be less predictable and effective for encouraging users. They've left open the possibility of comparing the GamePass approaches to other graphical user authentication systems, as well as incorporating features like memorability and security, to further validate their findings.

Another recent paper [20] suggests a recall-based graphical password that forms a graphical triangle with the use of permutation, a mathematical approach for determining the number of alternative configurations. The user is prompted to enter a one-time password supplied to them in the second step of the authentication procedure.

The most of the schemes covered in the area of graphical passwords, alphanumeric, or a combination of both. The major challenges in developing a graphical password based system are: making it efficient by not foiling the memory of a media device with picture datasets; making it easy to remember for a frequent user; and preventing shoulder surfing attacks because the lot of consumers use cellphones. Shoulder surfing assaults are the second most common after dictionary attacks, and they have the potential to make a highly efficient method unattractive if they are simple to break.

## III. Methodology

This section provides the description of a simple and efficient shoulder-surfing resistant and key logger resistant graphical password scheme. This scheme is particularly for android devices that have passwords containing only digits like a unlock pin [22] or UPI password. The proposed system has two phases -:

- Registration Phase
- Login Phase

### A. Registration Phase

The registration process is the initial stage, and the user has two or three options for establishing a graphical password depicted in Fig. 1, with the help of Registration flowchart.

- The user will select a color that he must remember, as well as three other options: the length of the password, which must be a multiple of three, the location where he must long-press a button rather than merely tapping it, and lastly, the position in which he will shake his phone.
- He can create a password of any length he wishes with these four variables; memorizing the password is also easier than previously proposed schemes because he only has to memorize four values, and he receives greater security than with traditional authentication procedures.

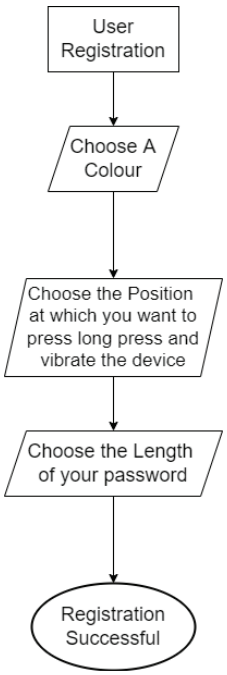



Fig. 1. Sequence Diagram for Registration Phase

- The colour can be chosen only from the given 10 colors. The password length should not be taken more than 30 digits and the long press and shaking position can only be less than the password length and greater than one.

Following these steps will set your shoulder-surfing resistant password, now the user just has to remember these 4 values to login into the system.

### B. Login Phase

The next phase is the login phase where the user will input the password to login into the system. As the user lands on the login page he will have to follow the certain steps to login as shown in Fig. 2.

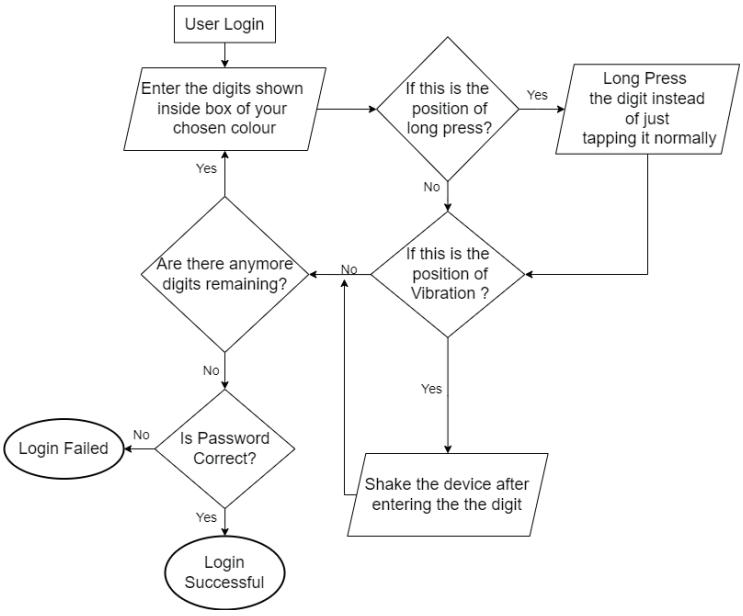
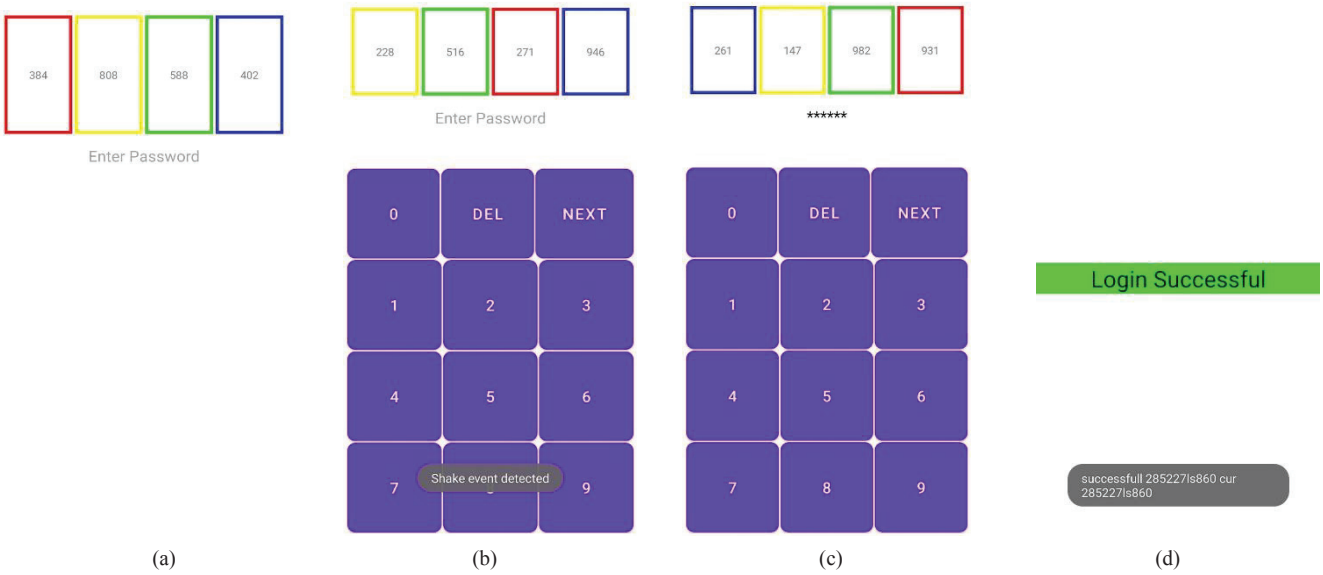



Fig. 2. Sequence Diagram for Login Phase



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:59:03 UTC from IEEE Xplore. Restrictions apply.




(a) (b) (c) (d)

Fig. 3 Different steps in Login Phase

- Fig. 3(a) shows the initial login screen where the user needs to shake the device to make the keyboard visible. The user will not see any keyboard to input the password, so he will have to shake the mobile phone to make the keyboard visible. This step will prevent hackers who don't know this graphical password technique and it is a good way to counter them.
- Now the user will have to use those values he used while registering, first one was the colour name, he observes his registered colour from the multiple boxes present on the screen and type the numbers present inside the box using our custom keyboard depicted in Fig. 3(b).
- After writing the digits then the user will have to press the next button present on the keyboard. This way we are entering the password but even if the attacker notes down our inputs and try to log in next time, he will not be able to do so as numbers inside each coloured box are random [23].
- Fig. 3(c) shows the process of entering the password where after entering some digits user has to make sure he vibrates and long press at the correct position, also the password entered is shown as a string of "*", so as to ensure another layer of security.
- As you can see that the difference between just pressing a button and long-pressing is negligible, and will be quite difficult to notice by the shoulder surfer. Also shaking of a device has not been considered in any way as user input, so the attacker cannot even imagine that shaking the device is also acting as an actual input.
- Following the same process of inputting the password until the length you initially fixed, keep in mind to press the next button after every three digits.
- If the login attempt fails consecutively three times, then the account will not be accessed until an email and OTP verification of the user is done.
- Finally, if everything goes correct the user will be able to login into the system successfully shown by Fig.3(d), without exposing its password to the shoulder surfer.

## IV. Empirical Analysis

We tried to assess the approach in the real world after designing the smartphone application, therefore we had 40 participants ranging in age from 18 to 55. We offered our participants 10 minutes and no more than two tries to guess the right password, and we provided them a recorded video of the password being typed. We split the participants into two groups: the first group was told to guess the password being a shoulder surfer and the second group was told to guess the password by hit and trial. Here we compared our technique with the Conventional numeric password scheme, which is widely used scheme for UPI pins. Fig. 4 shows the results of the analysis, in terms of percentage of people who were able to guess the password from both the techniques.

Our proposed technique outperforms traditional password/pin authentication systems in terms of security and simplicity. The initial empty layout, which only has coloured boxes with random numbers on it but no keyboard to type the password [21], can be very useful in fooling the attacker into thinking there's a problem with the device or the app, but that's not the case with our user because he knows the keyboard is only a shake away.

We compared our method to the most common alphanumeric, passface, and numeric password schemes, and found that our technique is more secure as compared to the listed one. Table 1 shows the results of this investigation.

The foundation of our approach is choosing a colour from an array of 10 options and then entering the numbers found inside the box of your selected colour on the login screen. Even if the attacker is aware of the method, the



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:59:03 UTC from IEEE Xplore. Restrictions apply.

attacker will be perplexed by the colour the user has selected and the length of the password, which may change for various users.

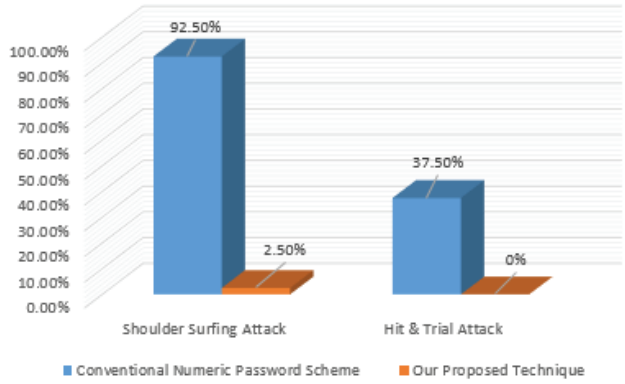



Fig. 4. Performance Analysis

TABLE I. DIFFERENT TECHNIQUES COMPARISON

| Name of Attacks | Alphanumeric-al Passcode | Pass-face [7] | Numeric-al Passcode | Proposed Technique |
|---|---|---|---|---|
| **Brute-force Attack** | Yes | Yes | Yes | No |
| **Dictionary Attack** | No | No | No | No |
| **Spyware Attack** | Yes | Yes | Yes | No |
| **Password Guessing** | Yes | Yes | Yes | Yes |
| **Shoulder-Surfing Attack** | Yes | Yes | Yes | No |

The attacker cannot brute force by selecting each color once and typing corresponding digits as the system allows only two consecutive false attempts and then blocks the account which can only be unblocked with email and OTP [24]. Shaking the smartphone to enter the password is a unique approach, and it is very hard for an attacker to detect that the user is shaking the device to write the password. To the attacker, the device's shaking may appear to be an accident, and he may not even notice it. Even if the attacker is aware that shaking is required, he must determine at when point during the password entry process he must shake the device. Most smartphones have the ability to detect long-press on a button. This simple yet amazing feature to add a variable at the time of registration that will determine the position where the user will long-press the corresponding digit instead of just pressing it. This way, a shoulder surfer will only see the digit being pressed, so even if he copies it, his password will be different from the original password.

The graphical analysis of each of the four techniques is shown in Fig. 5, with password length on the x-axis and approximate password complexity represented on the y-axis. Table II explains the techniques and their respective complexity according to the password length. For example, numeric passwords of length 1 can have up to 10 combinations i.e. 0,1,2,3,4,5,6,7,8,9 and similarly Alphanumeric passwords of length 1 will have 62 combinations as the password can be any of the 62 alphanumeric characters available. In our technique we have used only digits but there are two more factors: long press and vibration, and on top of that we have another factor of choosing a colour, making the password space of 1-digit password of our technique equals to 48. In the same way we can calculate the complexity of each technique upto 6 digits and the data is shown in Fig.5 with the help of the graph.

TABLE II. COMPLEXITY ANALYSIS COMPARISON

| Technique | Password Length in Digits | | | | | |
|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** |
| **Proposed Technique** | 48 | 576 | 6912 | 82944 | 995328 | 11943936 |
| **Numerical Passcode** | 10 | 100 | 1000 | 10000 | 100000 | 1000000 |
| **Alphanumeri-cal Passcode** | 62 | 3844 | 238328 | 14776336 | 916132832 | 56800235584 |
| **Passface [7]** | 92 | 81 | 729 | 6561 | 59049 | 531441 |

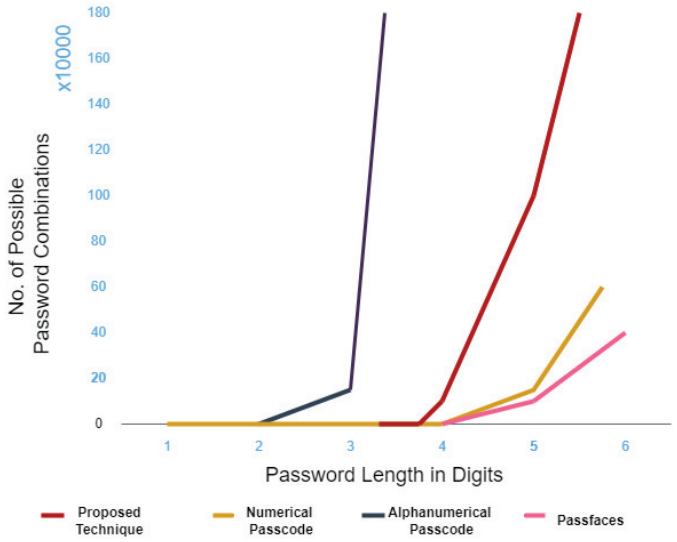



Fig. 5. Graphical Analysis of Different Techniques

Overall the complexity of the proposed graphical password based system is better than the conventional password which was prone to shoulder surfing attacks. For a password to be considered good, it should also be easier to remember which is the case with our password which only has 4 variables i.e. one colour, and three digits which are similar to conventional passwords [25] in terms of simplicity. As we can see, while the participant is unaware of the strategy, it is difficult to log into the system; but, when the participant is aware of the technique, it gives satisfactory results. It is also fair to assume that as the number of participants grows, the proportion of valid password guesses decreases.

## V. CONCLUSION AND FUTURE SCOPE

This graphical password setup introduces a new method for creating strong passwords that will withstand any password cracking technology. It is a novel method for creating a random password every time by allowing the user to select from a variety of colours and enabling the long-press and vibration features during registration. In terms of security and completeness, this new graphical method of generating the password gives more dependability and scope.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:59:03 UTC from IEEE Xplore. Restrictions apply.

Shoulder surfing was the new threat to users at first, and this threat instilled long-term anxieties in many users who frequently use their credentials in public locations and allow others to get access to their systems in order to steal their data and harm their privacy.

This strategy may be used with additional authentication mechanisms such as keyboard shuffling or utilising alphanumeric characters to establish the password, increasing the password space and making it more difficult for the attacker to guess the password. Multiple long-pressing features in three or four intervals might increase it even more. It may also be used as a smartphone app lock to evaluate its real-world usability and performance, and if it works well, an upgraded version of this technique can be used in current authentication systems.

However, with the introduction of our outlined approach, which has been demonstrated here for Android devices, we can confidently predict that it will be extremely useful for active smartphone users to make UPI payments in public locations without worry of security.

## REFERENCES


[1] Zakaria, Omar, Toomaj Zangooei, and Mohd Afizi Mohd Shukran. "Enhancing Mixing Recognition-Based and Recall-Based Approach in Graphical Password Scheme." International Journal of Advancements in Computing Technology 4, no. 15 (2012).

[2] Schaub, Florian, Ruben Deyhle, and Michael Weber. "Password entry usability and shoulder surfing susceptibility on different smartphone platforms." In Proceedings of the 11th international conference on mobile and ubiquitous multimedia, pp. 1-10. 2012.

[3] Cain, Ashley A., Liya Chiu, Felicia Santiago, and Jeremiah D. Still. "Swipe authentication: Exploring over-the-shoulder attack performance." In Advances in Human Factors in Cybersecurity, pp. 327-336. Springer, Cham, 2016.

[4] Suo, Xiaoyuan, Ying Zhu, and G. Scott Owen. "Graphical passwords: A survey." In 21st Annual Computer Security Applications Conference (ACSAC'05), pp. 10-pp. IEEE, 2005.

[5] Dhamija, Rachna, and Adrian Perrig. "Deja {Vu--A} User Study: Using Images for Authentication." In 9th USENIX Security Symposium (USENIX Security 00). 2000.

[6] Sobrado, Leonardo, and Jean-Camille Birget. "Graphical passwords. "The Rutgers Scholar, an electronic Bulletin for undergraduate research 2002.

[7] Brostoff, Sacha, and M. Angela Sasse. "Are Passfaces more usable than passwords? A field trial investigation." In People and computers XIV—usability or else!, pp. 405-424. Springer, London, 2000.

[8] Jansen, Wayne. "Authenticating mobile device users through image selection." WIT transactions on information and communication technologies 30 (2004).

[9] Jansen, Wayne, et al. "Picture password: a visual login technique for mobil devices." NIST Interagency/Internal Report (NISTIR)-7030 2003.

[10] Takada, Tetsuji, Takehito Onuki, and Hideki Koike. "Awase-e: Recognition-based image authentication scheme using users' personal photographs." In 2006 Innovations in Information Technology, pp. 1-5. IEEE, 2006.

[11] Ugochukwu, Ejike Ekeke Kingsley, and Yusmadi Yah Jusoh. "A review on the graphical user authentication algorithm: recognition-based and recall-based." Journal of Information Processing and Management 4, no. 3 (2013): 238-252.

[12] Davis, Darren, Fabian Monrose, and Michael K. Reiter. "On user choice in graphical password schemes." In USENIX security symposium, vol. 13, no. 2004, pp. 11-11. 2004.

[13] Man, Shushuang, Dawei Hong, and Manton M. Matthews. "A Shoulder-Surfing Resistant Graphical Password Scheme-Wiw." In Security and management, pp. 105-111. 2003.

[14] Wakabayashi, Naoki, Masayuki Kuriyama, and Atsushi Kanai. "Personal authentication method against shoulder-surfing attacks for smartphone." In 2017 IEEE International Conference on Consumer Electronics (ICCE), pp. 153-155. IEEE, 2017.

[15] Yu, Xingjie, Zhan Wang, Yingjiu Li, Liang Li, Wen Tao Zhu, and Li Song. "EvoPass: Evolvable graphical password against shoulder-surfing attacks." Computers & Security 70 (2017): 179-198.

[16] Mrs. Aakansha S. Gokhale, Vijaya S. Waghmare, The Shoulder Surfing Resistant Graphical Password Authentication Technique, Procedia Computer Science, Volume 79, Pages 490-498, 2016.

[17] Higashikawa, So, Tomoaki Kosugi, Shogo Kitajima, and Masahiro Mambo. "Shoulder-surfing resistant authentication using pass pattern of pattern lock." IEICE TRANSACTIONS on Information and Systems 101, no. 1 (2018): 45-52.

[18] Mohammedi, Mohamed, Mawloud Omar, and Abdelmadjid Bouabdallah. "Secure and lightweight remote patient authentication scheme with biometric inputs for mobile healthcare environments." Journal of Ambient Intelligence and Humanized Computing 9, no. 5 (2018): 1527-1539.

[19] Raptis, George E., Christina Katsini, Andrew Jian-Lan Cen, Nalin Asanka Gamagedara Arachchilage, and Lennart E. Nacke. "Better, Funner, Stronger: A Gameful Approach to Nudge People into Making Less Predictable Graphical Password Choices." In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1-17. 2021.

[20] Kaur, Amanpreet, and Khurram Mustafa. "Preference-Oriented Password-Based Authentication." In Information and Communication Technology for Competitive Strategies (ICTCS 2020), pp. 953-965. Springer, Singapore, 2022.

[21] Khedr, Walid I. "Improved keylogging and shoulder-surfing resistant visual two-factor authentication protocol." Journal of Information Security and Applications 39 (2018): 41-57.

[22] De Luca, Alexander, Bernhard Frauendienst, Sebastian Boring, and Heinrich Hussmann. "My phone is my keypad: privacy-enhanced pin-entry on public terminals." In Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7, pp. 401-404. 2009.

[23] HIMAJA, BHARATHEPUDI SRI, POTLURI UDAY TEJA, MADDIRALA VISWANADHA KASHYAP, and SAJJA POORNA CHAND. "Graphical Password to avoid Shoulder-Surfing."

[24] Papadopoulos, Athanasios, Toan Nguyen, Emre Durmus, and Nasir Memon. "Illusionpin: Shoulder-surfing resistant authentication using hybrid images." IEEE Transactions on Information Forensics and Security 12, no. 12 (2017): 2875-2889.

[25] Wiedenbeck, Susan, Jim Waters, Leonardo Sobrado, and Jean-Camille Birget. "Design and evaluation of a shoulder-surfing resistant graphical password scheme." In Proceedings of the working conference on Advanced visual interfaces, pp. 177-184. 2006.




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:59:03 UTC from IEEE Xplore. Restrictions apply.

Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-022-04022-5

ORIGINAL RESEARCH

# A novel type-II intuitionistic fuzzy clustering algorithm for mammograms segmentation

**Sonika Dahiya**[1] · **Anjana Gosain**[2]

Received: 21 January 2021 / Accepted: 30 May 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

**Abstract**
Fuzzy clustering has been gaining prominence in medical image segmentation but challenges still exist. This paper proposes a novel Type-II Intuitionistic Fuzzy C Means clustering algorithm by introducing a new membership degree called Intuitionistic Type-II membership. Intuitionistic Type-II membership combines Type-II membership with hesitation degree. Using Intuitionistic Type-II membership, the proposed algorithm shows following advantages: (1) defining clusters clearly, (2) robustness to noise and outliers, and (3) improving desired position of centroids. These advantages make Type-II Intuitionistic Fuzzy C Means clustering algorithm a preferred choice for mammogram segmentation. On some mammograms from Mammographic Image Analysis Society database, performance of Type-II Intuitionistic Fuzzy C Means clustering algorithm is compared with the performance of other fuzzy clustering algorithms such as Fuzzy C-Means, Intuitionistic Fuzzy C-Means, Type-II Fuzzy C-Means, Interval Type-II Fuzzy C-Means, and Particle Swarm Optimization Based Interval Type-II Fuzzy C-Means using Intuitionistic Fuzzy Sets. By qualitative analysis, results of Type-II Intuitionistic Fuzzy C Means are found to be better than the results of discussed algorithms as the proposed algorithm identifies shape and size of lumps in mammograms more accurately. On experimenting with synthetic data sets, it is observed that Type-II Intuitionistic Fuzzy C Means produces robust and stable results as outliers increase and average error is reduced by 84% on D15 dataset.

**Keywords** Fuzzy clustering · FCM · IFCM · T2FCM · Interval type II fuzzy clustering · Mammogram

## 1 Introduction

Fuzzy clustering is a technique of dividing data objects in fuzzy groups such that each data object may fractionally belong to a number of clusters (Gosain and Dahiya 2016). Cluster membership can be a fraction ranging from 0 to 1 and is subjected to the condition that sum of all the memberships of a data object to each of the cluster is always 1. Fuzzy clustering has wide applications such as customer segmentation (Bodyanskiy et al. 2019; Moodley et al. 2020; Munusamy and Murugesan 2020), recommender systems (Oner and Oztaysi 2018), anomaly detection and fault diagnosis (Dahiya et al. 2020; Nancy et al. 2020; Rodríguez-Ramos et al. 2018; Rodríguez Ramos et al. 2019), resource optimization (Gupta et al. 2017), sentiment classification (Phu et al. 2017), network optimization (Patil and Ragha 2020; Rajput and Kumaravelu 2019; Rajput et al. 2020), and image segmentation (Hu et al. 2020; Tongbram et al. 2021; Wen et al. 2020) etc. Images, especially medical images, are deemed fuzzy because of the ambiguity in boundaries, haziness in class definition, and imprecise grey levels. So, fuzzy clustering is very useful for segmentation of medical images.

Fuzzy-C-Means (FCM), proposed by Bezdek et al. (1984), is the earliest fuzzy clustering approach with wide applications. However, FCM is very sensitive to presence of noise and outliers. To overcome the weaknesses of FCM, a number of proposals such as Possibilistic Fuzzy C-Means(PFCM) by Pal et al. (2005), Credibilistic Fuzzy C-Means (CFCM) by Chintalapudi and Kam (1998) and Noise Clustering (NC) by Dave (1991) were made by the researchers. PFCM works on memberships and possibilities and is less sensitive to outliers. But it fails to perform well on data that has clusters of high size variability. CFCM performs better than FCM and PFCM on data contaminated with outlier by reducing the effect of outliers on resultant

✉ Sonika Dahiya
sonika.dahiya11@gmail.com

[1] Delhi Technological University, Delhi 110042, India

[2] USICT, GGS Indraprastha University, Sector 16C, Dwarka, Delhi 110075, India

Published online: 30 August 2022

clusters. However, CFCM is sensitive to initial parameters. NC performs better in comparison to FCM, PFCM, and CFCM as it creates a separate cluster for noise and outliers (Ferraro and Giordani 2019). But NC fails to identify all outliers and noise. All these approaches basically work on suppressing the effect of noise and outliers in computing clusters as discussed in research work by Gosain and Dahiya (2020) and Dahiya et al. (2021). Rhee and Hwang (2001) proposed T2FCM by introducing Type-II membership (T2m). Its basis is that the data objects closer to centroid contribute more in centroid computation, improving the resultant clusters. However, Type 2 Fuzzy C-Means (T2FCM) shows poor performance with noise and outlier contaminated data but it still has wide applications such as image classification (Xu et al. 2019), ranking job applicants (Doctor et al. 2008), developing an online decision support system (Balaji et al. 2008; Gu and Zhang 2007) etc.

In 2011, Intuitionistic Fuzzy C-Means (IFCM) was proposed by Chaira (2011) based on the concept of Intuitionistic Fuzzy Entropy (IFE) and Intuitionistic Fuzzy membership (IFm). IFCM shows better robustness in clustering accuracy even when data has high noise. Intuitionistic Fuzzy membership (IFm) is the backbone of IFCM and is the major reason of success of IFCM as it deals with the uncertainty more effectively and precisely. This makes IFCM quite effective in clustering and handling of noise and outliers. IFCM has wide applications such as denoising color archival document images (Elhedda et al. 2020), image segmentation (Chaira 2011; Verma et al. 2019; Xiangxiao et al. 2019) including medical image segmentation (Ouchicha et al. 2019) etc. IFCM is most effective in defining clusters and handling of noise and outliers. Also, IFCM showed very good results in segmentation of MRI human brain image. But IFCM performance for mammograms segmentation is not found to be effective as mammograms have limited spatial resolution which causes higher uncertainty and vagueness in boundaries.

In this research paper, we have worked on mammogram segmentation. Mammogram segmentation is an interesting area as mammography is the first step in the diagnosis of breast cancer. Mammography is very effective and affordable way of providing anatomical information such as high density, cell mass, or calcification. Presence of any abnormality may be further diagnosed for presence of cancer. According to radiologist, usually round or oval shaped masses are benign tumor that are non-cancerous, whereas partially round shaped with spikes or any irregular outlines masses are malignant tumor which are cancerous.

Segmentation of mammogram with all lumps and ducts in their accurate shape and size is an important problem. Most of the work done on breast cancer diagnosis using mammograms has addressed the issue as a classification problem using deep learning (Kavitha et al. 2022; Li et al. 2019; Parvathavarthini and Deepa 2021), artificial intelligent techniques (Zheng and Chan 2001), and various classifiers (Al-Najdawi et al. 2015; Eltoukhy et al. 2010). However, very limited work has been done on mammogram segmentation. Most of this segmentation work is done using image processing (Maitra et al. 2012; Singh and Gupta 2015).

In this paper, we proposed a new clustering algorithm-Type-II Intuitionistic Fuzzy C-Means (T2IFCM) by introducing Intuitionistic Type-II membership (IT2m). IT2m combines type-II membership with hesitation degree in Intuitionistic fuzzy sets (IFS). Using IT2m for clustering of data that is represented using IFS has some great advantages—(1) defining clusters clearly, (2) robustness to noise and outliers, and (3) improving desired position of centroids. Performance of T2IFCM is compared with FCM, IFCM, and T2FCM on some standard synthetic data sets and on some mammogram images from Mammographic Image Analysis Society (MIAS) database (Ibrahim et al. 1997). Considering the encouraging results, T2IFCM can be very effective algorithm in designing tools for early diagnosis of breast cancer.

This paper is composed of five sections. Section 2 provides brief of fuzzy clustering approaches. Section 3 elaborates the proposed algorithm—T2IFCM and Sect. 4 elaborates the simulation process and presents results on synthetic and MIAS-mammograms. Section 5 concludes the proposal using analysis of results from Sect. 4.

## 2 Background algorithms

In this section, basic description of FCM, T2FCM, IFCM, IT2FCM, IT2IFCM, PSO-IT2FCM, and PSO-IT2IFCM is provided.

### 2.1 Fuzzy-C-Means algorithm (FCM)

FCM proposed by Bezdek et al. (1984) is the first fuzzy clustering algorithm that bridged clustering to fuzzy clustering. It works on the supposition that number of clusters in the data set are pre-known. Optimal clusters are obtained by minimizing FCM objective function ($Obj_{FCM}$), subject to the constraint that sum of all the memberships of a data object to each cluster is one and the same is formulated in Eq. (2):

$$Obj_{FCM} = \sum_{j=1}^{cluster_n} \sum_{i=1}^{z} u_{ij}^{m} dist_{ij}^{2} \tag{1}$$

where $cluster_n$ is the number of clusters; $z$ is count of data objects in given dateset; $u_{ij}$ is the fuzzy membership of $d_i$ in$(c_j)$; $d_i$ is $i$th data object in the data set$D$, where$D = \{d_1, d_2, d_3, ....., d_z\}$; $c_j$ is $j$th centroid in the set of centroid C, where $C$ is represented as$\{c_1, c_2, ....., c_{cluster_n}\}$; $m$ is the fuzziness index (or fuzzifier);

$dist_{ij} = \sqrt{\sum_{k=1}^{dim} \left(d_{ik} - c_{jk}\right)^2}$ is Euclidean distance between $d_i$ and $c_j$, dim represents dimensionality of the data set –$D$.

$$\sum_{j=1}^{cluster_n} u_{ij} = 1, \forall i = 1, 2, 3, .., z \tag{2}$$

Lagrangian multipliers is used to compute the optimal solution for $Obj_{FCM}$ and the optimal solution is given by Eqs. (3) and (4):

$$c_j = \frac{\sum_{i=1}^{z} \left(u_{ij}^m d_i\right)}{\sum_{i=1}^{z} \left(u_{ij}^m\right)} \forall j \tag{3}$$

$$u_{ij} = \left( \sum_{k=1}^{cluster_n} \left( \frac{dist_{ij}^2}{dist_{kj}^2} \right)^{\frac{1}{(m-1)}} \right)^{-1} \tag{4}$$

## 2.2 Intuitionistic fuzzy-C-means (IFCM)

IFCM introduces intuitionistic fuzzy entropy (IFE) which is based on hesitation degree ($\pi_{ij}$). IFE is formulated as:

$$IFE = \sum_{j=1}^{cluster_n} \pi_j^* e^{1-\pi_j^*} \tag{5}$$

where $cluster_n$ is the number of clusters and $\pi_j^*$ is formulated as:

$$\pi_j^* = \frac{1}{z} \sum_{i=1}^{z} \pi_{ij}, i \in [1, z] \tag{6}$$

where $z$ is the size of dataset and $\pi_{ij}$ is hesitation degree which is formulated as follows:

$$\pi_{ij} = 1 - u_{ij} - \left(1 - u_{ij}^{\alpha}\right)^{\frac{1}{\alpha}}, \alpha > 0 \tag{7}$$

where is $u_{ij}$ is fuzzy membership and $\alpha$ is a tuning parameter for algorithms using Intuitionistic fuzzy sets. $u_{ij}$ and $\pi_{ij}$ expresses fraction of belonging and not belonging of $i$th data object ($d_i$) to $j$th centroid ($c_j$) respectively (Chaira 2011). IFCM incorporates IFE and intuitionistic fuzzy membership ($u_{ij}^*$) (explained in Eq. (9)) to form the following objective function:

$$Obj_{IFCM} = \sum_{j=1}^{cluster_n} \sum_{i=1}^{z} (u_{ij}^*)^m dist_{ij}^2 + \sum_{j=1}^{cluster_n} \pi_j^* e^{1-\pi_j^*} \tag{8}$$

where $z$ is the size of the data set, $cluster_n$ is the number of clusters, $dist_{ij}$ is Euclidean distance between data object $d_i$ and centroid $c_j$ (as explained in Sect. 2.1), and intuitionistic fuzzy membership ($u_{ij}^*$) is formulated as:

$$u_{ij}^* = u_{ij} + \pi_{ij} \tag{9}$$

where $u_{ij}$ is fuzzy membership and $\pi_{ij}$ is hesitation degree explained in Eq. (7). On minimization of IFCM objective function (Eq. (8)), subject to constraint specified in Eq. (2), centroid is updated as per Eq. (10) and fuzzy membership ($u_{ij}$) is updated as per Eq. (4):

$$c_j = \frac{\sum_{i=1}^{z} \left(u_{ij}^* d_i\right)}{\sum_{i=1}^{z} \left(u_{ij}^*\right)} \tag{10}$$

Algorithm 1 explains step by step computation of clusters using IFCM algorithm.

**Algorithm 1: IFCM Algorithm**

**Input Arguments:** Data Set ($D_{[z,dim]}$), number of clusters (cluster_n), Maximum Iterations (mt), fuzziness index (m), alpha ($\alpha$)

**Output Arguments:** Clusters Centroids ($C_{[cluster_n, dim]}$), Membership Matrix ($U_{[cluster_n,z]}$)

Initialize default parameters (fuzziness index, Maximum Iterations, Minimum_Improvement) to [2; 100; 1e−5] or to user values

Randomly initialize U, subject to membership constraint (*2*)

for i ← 1 to mt

  Begin

    Compute Hesitation Degree using eq. (7)

    Compute Intuitionistic fuzzy membership degree as per Eq. (9), using $u_{ij}$ and Hesitation Degree ($\pi_{ij}$)

    Update C using Eq. (10)

    Compute $Obj_fun_i$ using Eq. (8)

    Update U using Eq. (4)

    If ($Obj_fun_i$ - $Obj_fun_{(i-1)}$) < Minimum_Improvement

      Begin

        Break

      EndIf

  Endfor

Return C and U

## 2.3 Type-II fuzzy-c-means (T2FCM)

T2FCM introduces type-II membership (T2m) based on the logic that contribution of data objects in updation of a centroid should be respective to their membership on that centroid (Rhee and Hwang 2001). T2m is mathematically formulated as:

$$a_{ij} = u_{ij} - \frac{\left(1 - u_{ij}\right)}{2} \tag{11}$$

where $a_{ij}$ is the T2m and $u_{ij}$ is the fuzzy membership. T2m is represented as $a_{ij}$. To apply T2FCM logic, T2m is used for centroid updation. Hence, centroid updation equation for T2FCM is:

$$c_j = \frac{\sum_{i=1}^{z}\left(a_{ij}^{m} d_i\right)}{\sum_{i=1}^{z}\left(a_{ij}^{m}\right)} \tag{12}$$

where $d_i$ is $i$th data object in the data set $D$, where $D = \{d_1, d_2, d_3, \ldots\ldots, d_z\}$, z is the size of $D$, and $c_j$ is $j$th centroid in the set of centroid C, where $C$ is represented as$\{c_1, c_2, \ldots\ldots, c_{cluster_n}\}$.

Algorithm 2 explains step by step computation of clusters using T2FCM algorithm.

**Algorithm 2: T2FCM Algorithm**

**Input Arguments:** Data Set ($D_{[z,dim]}$), number of clusters (cluster_n), Maximum Iterations (mt), fuzziness index (m)

**Output Arguments:** Clusters Centroids ($C_{[cluster_n,\ dim]}$), Membership Matrix ($U_{[cluster_n,z]}$)

Initialize default parameters (fuzziness index, Maximum Iterations, Minimum_Improvement) to [2; 100; 1e−5] or to user values

Assign random membership to each data object $d_j$, subject to membership constraint (*2*)

for i ← 1 to mt

  Begin

    Compute Type-II membership using Eq. (11)

    Update $c_j$ for $\forall j \in [1, 2, \ldots, cluster_n]$ using Eq. (12).

    Compute $Obj_fun_i$ using Eq. (11)

    Update U using Eq. (4)

    If (Obj_fun(i) - Obj_fun(i−1)) < Minimum_Improvement

      Begin

        Break

      EndIf

  Endfor

Return C and U

## 2.4 Interval type-2 FCM (IT2FCM)

IT2FCM, an extension of FCM, uses two different fuzzifiers : $m_1$ and $m_2$, that indicate distinct fuzzy degrees (Dang et al. 2015; Hwang and Rhee 2007). The Objective function for IT2FCM is formulated as:

$$\begin{cases} J_1 = \sum_{i=1}^{z}\sum_{j=1}^{cluster_n} u_{ij}^{m_1} dist_{ij} \\ J_2 = \sum_{i=1}^{z}\sum_{j=1}^{cluster_n} u_{ij}^{m_2} dist_{ij} \end{cases} \tag{13}$$

where *cluster_n* is the number of clusters, $z$ is size of the data set, $u_{ij}$ is fuzzy membership, and $dist_{ij}$ is Euclidean distance between $d_i$ and $c_j$.

To solve the objective function, iterations are run and the left memberships $u_{ij}^{L}$ and right memberships $u_{ij}^{R}$ are computed as follows:

$$u_{ij}^{L} = \frac{\sum_{l=1}^{M} u_{il}}{M}, \quad u_{il} = \begin{cases} \overline{u}_{ij}(d_i), & if(d_{il})\ uses\ \overline{u}_{ij}\ for\ v_j^{L} \\ \underline{u}_{ij}(d_i), & Otherwise \end{cases} \tag{14}$$

$$u_{ij}^{R} = \frac{\sum_{l=1}^{M} u_{il}}{M}, \quad u_{il} = \begin{cases} \overline{u}_{ij}(d_i), & if(d_{il})\ uses\ \overline{u}_{ij}\ for\ v_j^{R} \\ \underline{u}_{ij}(d_i), & Otherwise \end{cases} \tag{15}$$

where $M$ represent number of attributes of each pattern and $\overline{u}_{ij}(d_i)$, $\underline{u}_{ij}(d_i)$ are calculated as follows:

$$\begin{cases} \overline{u}_{ij}(d_i) = \max\left\{ \left[\sum_{k=1}^{cluster_n}\left(\frac{dist_{ij}}{dist_{ik}}\right)^{\frac{2}{m_1-1}}\right]^{-1}, \left[\sum_{k=1}^{cluster_n}\left(\frac{dist_{ij}}{dist_{ik}}\right)^{\frac{2}{m_2-1}}\right]^{-1} \right\} \\ \underline{u}_{ij}(d_i) = \min\left\{ \left[\sum_{k=1}^{cluster_n}\left(\frac{dist_{ij}}{dist_{ik}}\right)^{\frac{2}{m_1-1}}\right]^{-1}, \left[\sum_{k=1}^{cluster_n}\left(\frac{dist_{ij}}{dist_{ik}}\right)^{\frac{2}{m_2-1}}\right]^{-1} \right\} \end{cases} \tag{16}$$

Accurate centroids and memberships are computed using type-reduction and defuzzification techniques:

$$v_j = \frac{v_j^{L} + v_j^{R}}{2} \tag{17}$$

$$u_{ij}(d_i) = \frac{u_{ij}^{R}(d_i) + u_{ij}^{L}(d_i)}{2} \tag{18}$$

## 2.5 Interval type-2 FCM using intuitionistic fuzzy sets (IT2IFCM)

IT2IFCM, an extension of IFCM, uses two different fuzzifiers : $m_1$ and $m_2$, that shows distinct fuzzy degrees (Nguyen et al. 2014). The Objective function for T2IFCM is:

$$\begin{cases} J_1(u,v) = \sum_{i=1}^{z}\sum_{j=1}^{cluster_n} u_{ij}^{*m_1} dist_{ij}^2 + \sum_{j=1}^{cluster_n} \pi_j^{*} e^{1-\pi_j^{*}} \\ J_2(u,v) = \sum_{i=1}^{z}\sum_{j=1}^{cluster_n} u_{ij}^{*m_2} dist_{ij}^2 + \sum_{j=1}^{cluster_n} \pi_j^{*} e^{1-\pi_j^{*}} \end{cases} \tag{19}$$

where *cluster_n* is the number of clusters, $z$ is size of the data set, $u_{ij}$ is intuitionistic fuzzy membership, $dist_{ij}$ is Euclidean distance between $d_i$ and $c_j$, $\pi_j^{*}$ is defined in Eq. (6).

## 2.6 Particle swarm optimization based interval type-2 FCM (PSO-IT2FCM)

Eberhart and Kennedy (1995), Clerc (2010) proposed an optimization method—PSO. In this method, particles look into feasible solution area until there is an optimal solution. Here, both experimental output of particles and exchanged information from particles neighbourhood are accounted as proposed by Zhao et al (2019). PSO search is expressed as:

$$V_i(t+1) = w \times V_i(t) + c_1 rand_1 \times \left(Pbest_i(t) - p_i(t)\right) + c_2 rand_2 \times (gbest(t) - p_i(t)) \tag{20}$$

$$p_i(t+1) = p_i(t) + V_i(t+1) \tag{21}$$

where $V_i$ is velocity vector and $p_i$ is position vector. Here, we have '$w$' as weight factor, $c_1$ and $c_2$ as acceleration coefficients, and $t$ as iteration counter. $rand_1$ and $rand_2$ shows two random numbers between 0 and 1. $Pbest_i$ indicates the best personal position corresponding to particle $i$ and $gbest$ is the global optimal position in the whole population.

### 2.7 Particle swarm optimization based interval type-2 FCM using intuitionistic fuzzy sets (PSO-IT2IFCM):

Interval type-2 fuzzy clustering fitness function with intuitionistic fuzzy information is designed as:

$$f_{API} = \left(\frac{1}{cluster_n} \times \frac{E_1}{E_c} \times D_{max}\right)^2 \tag{22}$$

where $cluster_n$ is the number of clusters, $E_c$ denotes the interval type-2 fuzzy compactness of the partition, and $E_1$ is the special case of $E_c$ when c is equal to 1. $D_{max}$ measures the maximum fuzzy separation of the partition as discussed by Zhao et al (2019). Then, PSO is used to find optimal solution to Interval Type-2 FCM in IFS data sets. Fitness function used to calculate $Pbest_i$ and $gbest$ is the objective function of IT2IFCM.

## 3 Proposed Algorithm—T2IFCM

A new fuzzy clustering algorithm, T2IFCM, is proposed that defines cluster boundary effectively, shows robustness to the presence of noise and outlier, and improves cluster centroid position. So, it produces remarkable results for mammogram segmentation.

T2IFCM introduces a new membership—Intuitionistic Type-II membership (IT2m) and uses Intuitionistic Fuzzy Sets (IFS) for cluster computation. IFS—'S' was explained by Chaira (2011) and is expressed as follows:

$$S = \left\{\left(d, u_S(d), v_S(d)\right) \mid d \in D\right\} \tag{23}$$

where $d$ is a data object, $D$ is the set of data objects, $u_S(d)$ is membership degree, and $v_S(d)$ is non-membership degree of the data object –$'d'$ in the given set, and the constraint on these memberships is:

$$0 \le u_S(d) + v_S(d) \le 1 \tag{24}$$

With the above understanding of IFS, Atanassov (2020) introduced Hesitation Degree which arises due to lack of knowledge in defining the membership degree of each data object $'d'$ in set $S$ and can be formulated as:

$$\pi_{ij} = 1 - u_{ij} - \left(1 - u_{ij}^{\alpha}\right)^{1/\alpha}, \quad \alpha > 0 \tag{25}$$

where $u_{ij}$ is a membership degree, $\left(1 - u_{ij}^{\alpha}\right)^{1/\alpha}$ is non-membership degree, and $\pi_{ij}$ is hesitation degree. Hesitation degree deals with uncertainty in fuzzy environment, especially the boundary region among the clusters.

IT2m is formulated as follows:

$$b_{ij}^{*} = a_{ij} + \pi_{ij} \tag{26}$$

where $b_{ij}^{*}$ is IT2m, $\pi_{ij}$ is Hesitation Degree as explained in Eq. (25), and $a_{ij}$ is T2m as explained in Eq. (11). T2m (Type-II membership) was introduced in T2FCM and its basis is that higher membership values contribute more than smaller membership values in updating the centroid. Because of T2m, cluster centroid in T2FCM converges to a more desirable location than the centroid position obtained through fuzzy membership ($u_{ij}$). Using IT2m, we proposed T2IFCM algorithm that improves the conventional IFCM and T2FCM methods in the following three ways.

1. T2IFCM helps in better definition of cluster boundary.
2. Cluster centers that are estimated by IT2m tend to converge to a more desirable location than cluster centers obtained by FCM/IFCM/T2FCM method even in the presence of noise.
3. T2IFCM is more robust to noise and outliers than other popular fuzzy clustering algorithms FCM/IFCM/T2FCM.

### 3.1 Clustering formulation

For the proposed T2IFCM algorithm, formulation of optimal clusters is based on minimization of following Objective function ($J_{T2IFCM}$):

$$J_{T2IFCM} = \sum_{j=1}^{cluster_n} \sum_{i=1}^{z} (u_{ij}^{*})^m dist_{ij}^2 + \sum_{j=1}^{cluster_n} \pi_j^{*} e^{1-\pi_j^{*}} \tag{27}$$

where $u_{ij}^{*}$ is IFm; $dist_{ij}^2$ is square of Euclidean distance between $d_i$ and $c_j$; $\pi_j^{*}$ is defined in Eq. (6); $z$ is ||D||; $m$ is fuzziness index; $cluster_n$ is count of clusters. Minimization of $J_{T2IFCM}$ is justified by suitable selection of membership

matrix (U). For this, $u_{ij}^*$ is iteratively updated using following equation:

$$u_{ij}^* = u_{ij} + \pi_{ij} \tag{28}$$

where $u_{ij}$ is computed as per Eq. (29) and $\pi_{ij}$ is computed using Eq. (25).

$$u_{ij} = \left( \sum_{k=1}^{cluster_n} \left( \frac{dist_{ij}^2}{dist_{kj}^2} \right)^{\frac{1}{(m-1)}} \right)^{-1} \tag{29}$$

For each iteration, firstly fuzzy membership is computed using Eq. (29) and then $\pi_{ij}$ is computed using Eq. (25). Then IFm and $a_{ik}$ are computed using Eq. (28) and Eq. (11) respectively, followed by IT2m computation as per Eq. (26). Then centroid is updated using Eq. (30). Iterations stop when minimum improvement in $J_{T2IFCM}$ for consecutive iteration is less than $\varepsilon$ where $\varepsilon$ is a user-defined value and $\varepsilon$ value is taken as 0.00005 for this simulation.

$$c_j = \frac{\sum_{i=1}^{z} \left( b_{ij}^{*m} d_i \right)}{\sum_{i=1}^{z} \left( b_{ij}^{*m} \right)} \tag{30}$$

Algorithm 3 explains step by step process of T2IFCM.

**Algorithm 3: T2IFCM Algorithm**

**Input Arguments:** Data Set ($D_{[z,dim]}$), number of clusters (cluster_n), Maximum Iterations (mt), fuzziness index (m), alpha ($\alpha$)

**Output Arguments:** Clusters Centroids ($C_{[cluster_n,\ dim]}$), Membership Matrix ($U_{[cluster_n,z]}$)

Initialize default parameters (fuzziness index, Maximum Iterations(mt), Minimum_Improvement) to [2; 100; 1e−5] or to user values

Assign random membership to each data object $d_j$, subject to membership constraint (*2*)

for i ← 1 to mt

  Begin

    Compute Fuzzy membership using eq. (29)

    Compute Hesitation degree using eq. (25)

    Compute IFm using eq. (28)

    Compute Type-II fuzzy membership using eq. (11)

    Compute IT2m using eq. (26)

    Update C using eq. (30)

    Compute $\pi_j^*$ using Eq. (6).

    Compute $Obj_fun_i$ using eq. (27)

    Update U using eq. (29)

    If($Obj_fun_i$ - $Obj_fun_{(i-1)}$) < Minimum_Improvement

      Begin

        Break

      EndIf

  Endfor

Return C and U

**Theorem:** *To find optimal clusters, it is a necessary to minimize $J_{T2IFCM}$ subject to the conditions specified in Eq. (2). On minimizing Eq. (27) under constraint in Eq. (2), we get Eq. (29) for membership computation and Eq. (30) for centroid computation.*

***Proof:*** Differentiate $J_{T2IFCM}$ wrt $u_{ij}^*$ and $C_i$ and the derivatives are set to zero. Thus, we derive Eqs. (29) and (30). Step by step calculation is shown in Appendix.

## 4 Experimental analysis and results

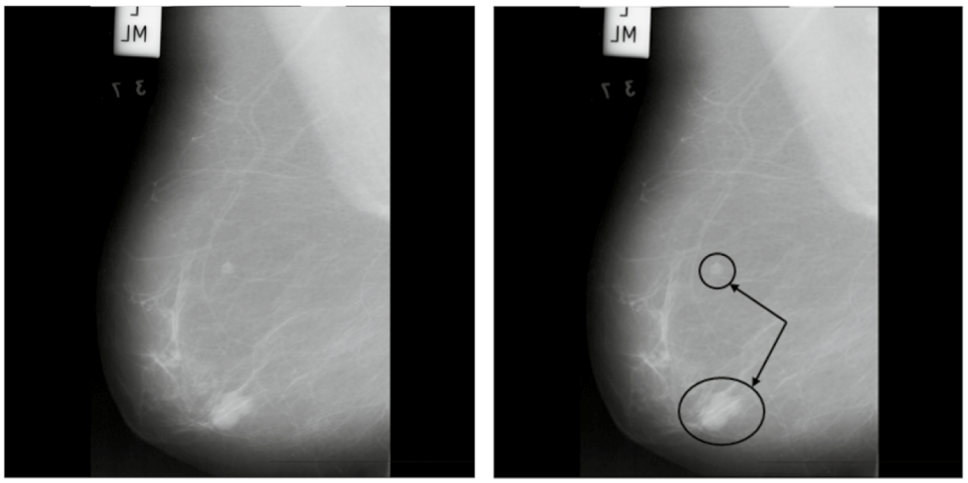
In this section, simulation results on synthetic data sets and on some mammograms from Mammographic Image Analysis Society (MIAS) database (Ibrahim et al. 1997) are presented. Proposed approach, T2IFCM, is compared with the FCM, T2FCM, IFCM, IT2FCM, IT2IFCM, PSO-IT2FCM, and PSO-IT2IFCM. All these algorithms are implemented using MATLAB-2017 on a system with intel Core i5 Processor-1600 MHz and 8GB RAM.

### 4.1 MIAS mammograph images

Experiment is done on several mammograms taken from MIAS database (Ibrahim et al. 1997) and any mammogram image can be segmented in 4 parts: background, fatty tissues, ducts, and lump (Tot et al. 2000). Figure 1a, mammogram from MIAS Database that is referred as mdb005. mdb005 has two lumps and few ducts and these lumps are highlighted in Fig. 1b. Figure 2a–d are the results from segmentation by FCM, IFCM, T2FCM, and T2IFCM respectively. Figure 3 shows segmentation results on mdb005 by IT2FCM, PSO-IT2FCM, IT2IFCM, and PSO-IT2IFCM. T2FCM detects neither the lumps nor the ducts, and it only segments the breast region and the background. Though PSO-IT2FCM performs marginally better than T2FCM, but still shows poor performance. FCM, IFCM, IT2FCM, IT2IFCM and PSO-IT2IFCM perform better than T2FCM and PSO-IT2FCM as they detect one lump but miss the other smaller lump. Also, all these algorithms almost fail to identify the ducts. However, T2IFCM detects both the lumps in very appropriate shape and size as well as shows almost all the ducts.

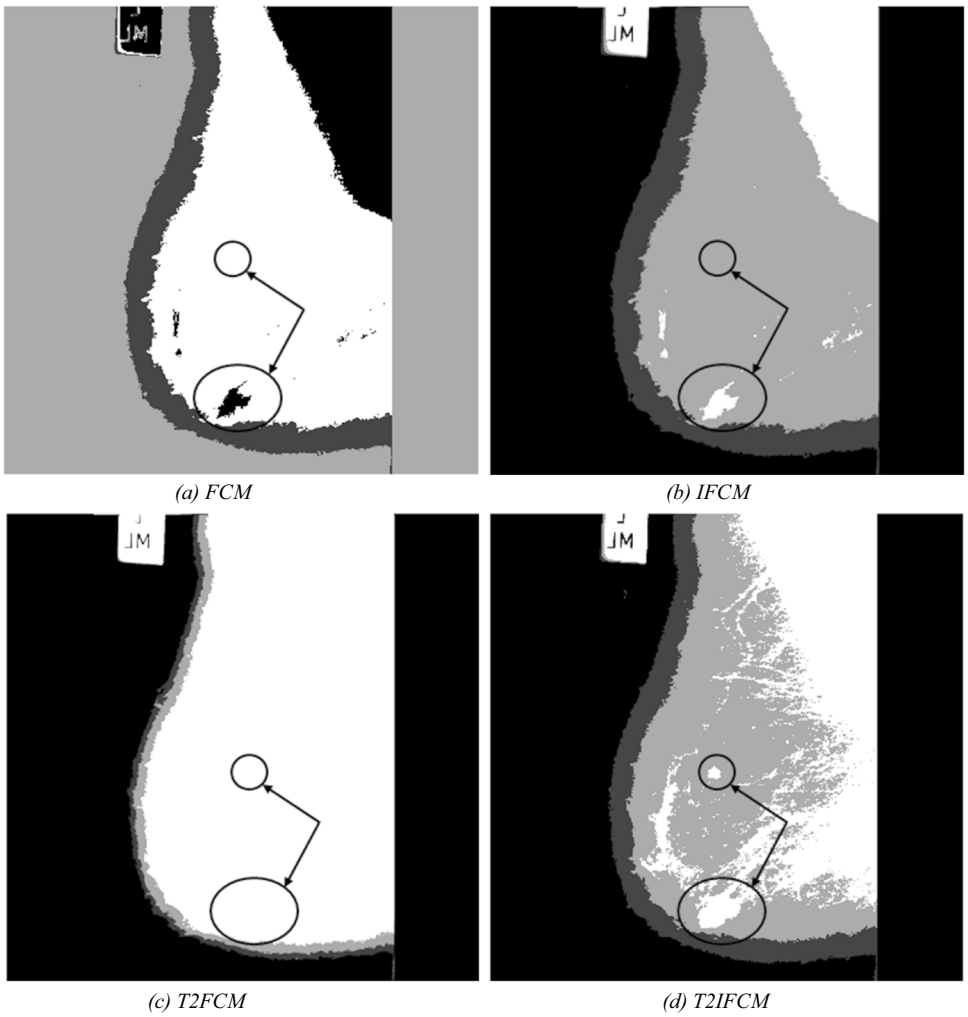
Figure 4 shows segmentation results on mdb010. mdb010 has one lump and good density of ducts. Figure 4b highlights the lump of Fig. 4a. Fig. 4c–f show segmentation results by FCM, IFCM, T2FCM, and T2IFCM respectively.

Figure 5 shows segmentation results on mdb010 by IT2FCM, PSO-IT2FCM, IT2IFCM, and PSO-IT2IFCM. T2FCM drastically fails as it only identifies the breast region in the image.



*(a) Original Image* *(b) Lumps are marked*

**Fig. 1** Mammogram image: mdb005

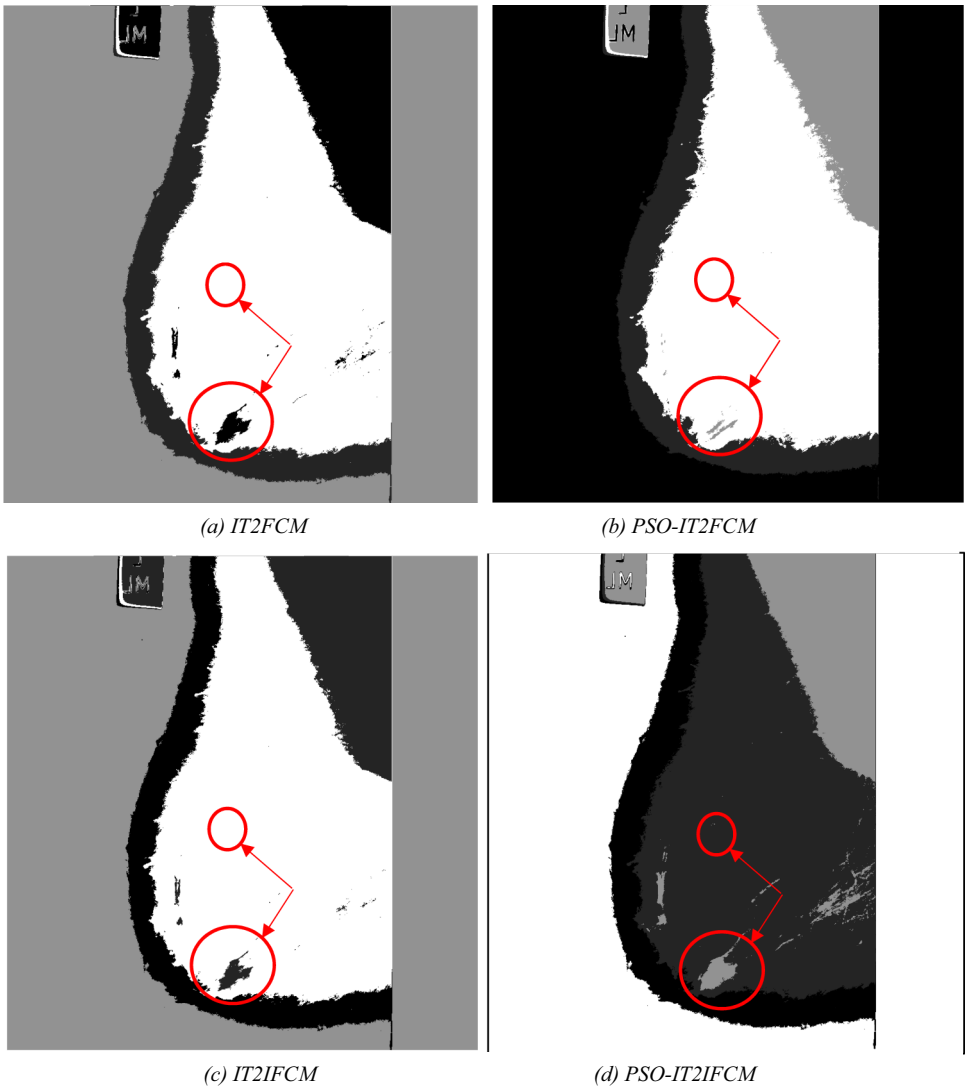


*(a) FCM* *(b) IFCM*

*(c) T2FCM* *(d) T2IFCM*

**Fig. 2** Segmentation results on mdb005

Also, IT2IFCM and PSO-IT2IFCM show similar poor performance. FCM, IFCM, IT2FCM, and PSO-IT2FCM identify the lump as well as most of the ducts, however, shape of lump by IFCM and PSO-IT2FCM is relatively better than shape and size identified by FCM and IT2FCM. T2IFCM outperforms IFCM and PSO-IT2FCM as it



*(a) IT2FCM* *(b) PSO-IT2FCM*

*(c) IT2IFCM* *(d) PSO-IT2IFCM*

**Fig. 3** Segmentation results on mdb005 by interval type-2 based algorithms

identifies lump in its most accurate shape and size and identifies all ducts.

Figure 6a shows Poisson noisy image of mdb005, and Fig. 6b highlights the lumps in Fig. 6a. Figure 6c–f show segmentation results of FCM, IFCM, T2FCM, and T2IFCM.

Figure 7 shows segmentation results on noisy mdb005 by IT2FCM, PSO-IT2FCM, IT2IFCM, and PSO-IT2IFCM. T2FCM detects neither the lumps nor the ducts, and it only segments the breast region and the background. All Interval Type-2 algorithms such as IT2FCM, PSO-IT2FCM, IT2IFCM, and PSO-IT2IFCM are drastically impacted by the presence of noise as they show just one lump in poor shape. FCM, IFCM, and T2IFCM perform better than all other algorithms but results of T2IFCM are much better than FCM and IFCM.

## 4.2 Synthetic data sets

Five synthetics data sets—D10, D12, D13, D14, and D15 are considered to analyse the performance of FCM, IFCM, T2FCM, and T2IFCM. D10 and D12, Diamond data sets with two identical clusters, are referred from a popular clustering paper by Pal et al. (2005). D13, D14, and D15 (Kaur et al. 2013) are variations of D12 with subsequent increment of outliers in D12. So, these data sets have increasing number of outliers, i.e., D10 with no outlier and D15 with highest number of outliers among the stated data sets. The ideal centroids corresponding to the two clusters in these data sets are provided by Pal et al. (2005):

$$C_{IDEAL} = \begin{bmatrix} 3.34 & 0 \\ -3.34 & 0 \end{bmatrix} \tag{31}$$

The average error is calculated as:

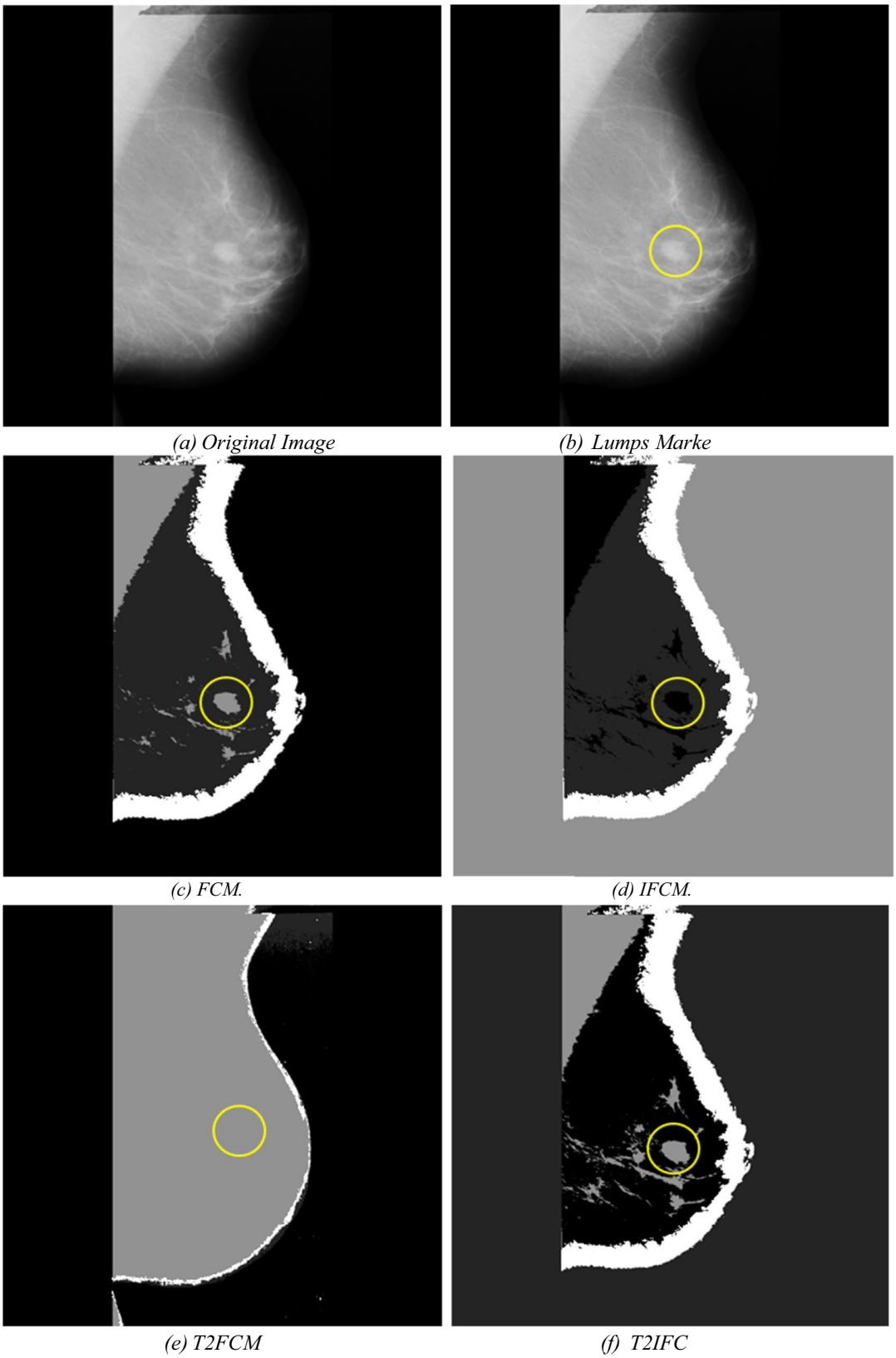
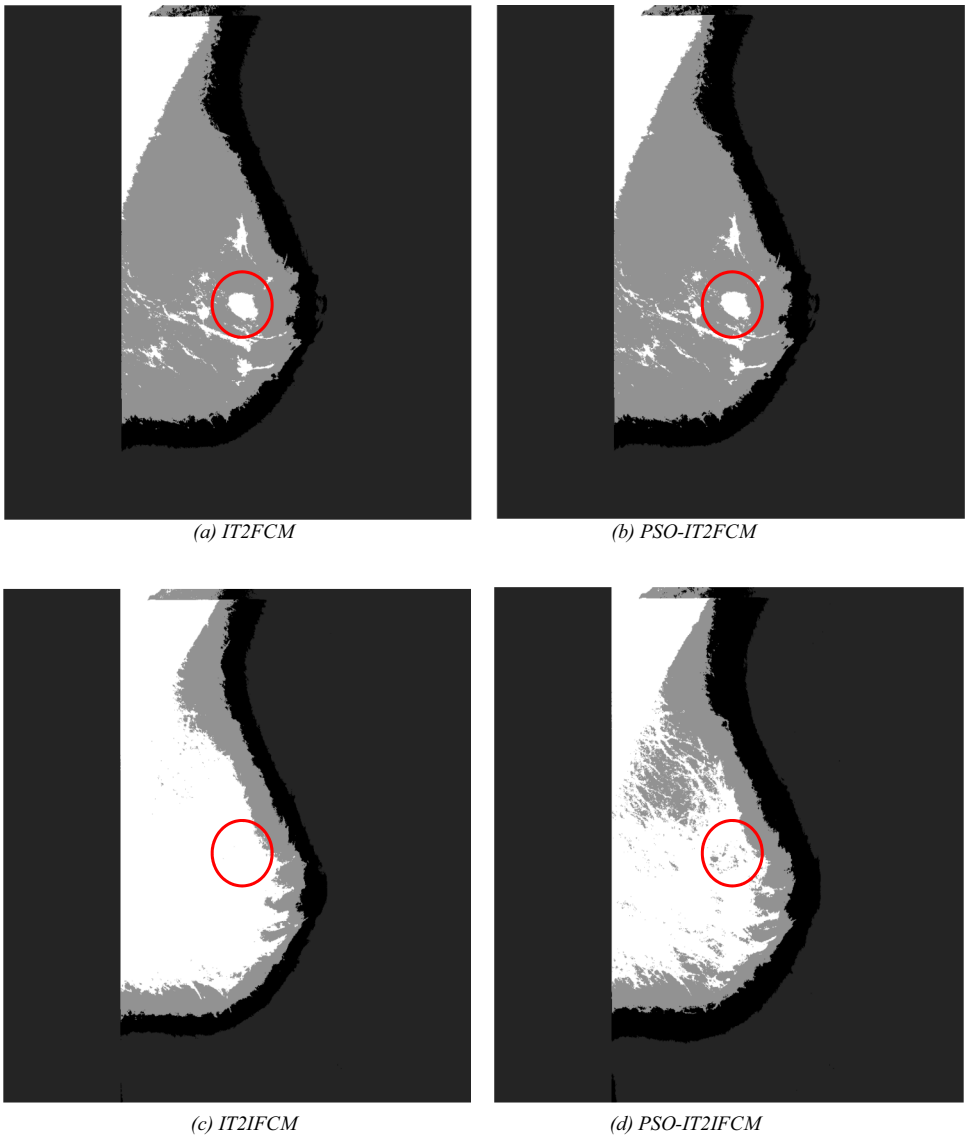



**Fig. 4** Segmentation results on mdb010



(a) IT2FCM (b) PSO-IT2FCM

(c) IT2IFCM (d) PSO-IT2IFCM

**Fig. 5** Segmentation results on mdb010 by interval type-2 Based algorithms

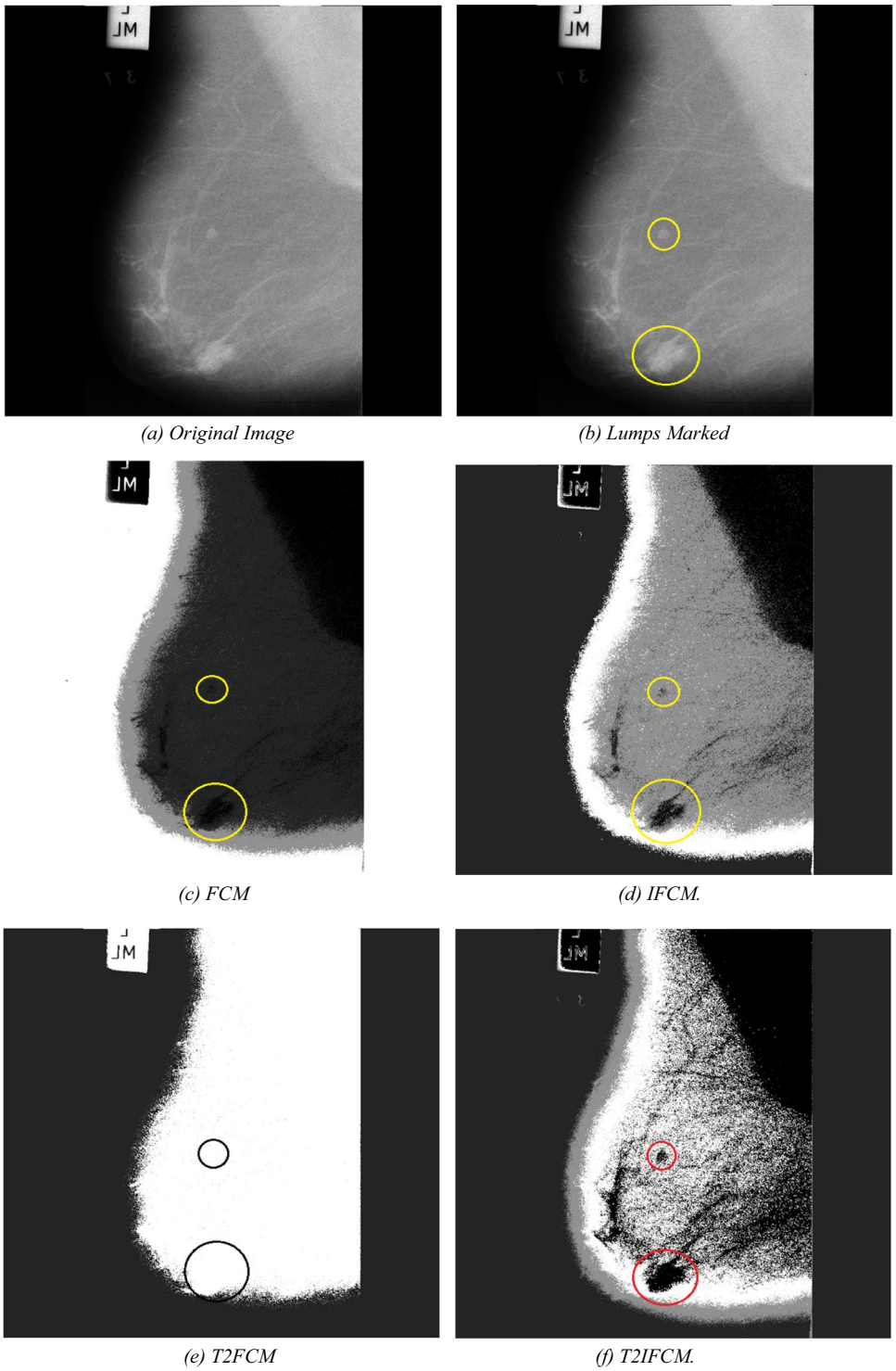
$$\text{AverageError} = \|C_{IDEAL} - C_*\|^2 \tag{32}$$

where $C_*$ is centroid computed by FCM, T2FCM, IFCM, or T2IFCM.

Figure 8 shows chart that display average error of various algorithms on different data sets such as D10, D12, D13, D14, and D15. From this chart, it can be observed that increase in average error is minimum for T2IFCM as the number of outliers increases. FCM performance gets severely affected with subsequent increase in outliers. Performance of T2FCM and IFCM lies somewhere between T2IFM and FCM. Table 1 and Fig. 9i show clustering results of FCM, IFCM, T2FCM, and T2IFCM on D14 data set. $\alpha$ is a tuning parameter for algorithms using Intuitionistic fuzzy sets to compute Hesitation Degree as per Eq. (25). Values of $\alpha$ in Tables 1 and 2 are the best values for respective algorithm. These values were arrived by hit & trial method. Similarly, Table 2 and Fig. 9ii show clustering results on D15 data set. Analyzing these tables and figures, it is observed that IFCM and T2IFCM are able to identify right clusters on D14 data set, but FCM and T2FCM fail to do so. Similarly, only T2IFCM has been able to identify right clusters on D15 data set. Thus, T2IFCM has been most robust to such densely outlier contaminated data sets.

## 5 Conclusion

This paper presents a novel fuzzy clustering algorithm-T2IFCM. T2IFCM introduces Intuitionistic Type-II membership (IT2m) which is a combination of hesitation degree and Type 2 fuzzy membership. T2IFCM, which uses IT2m



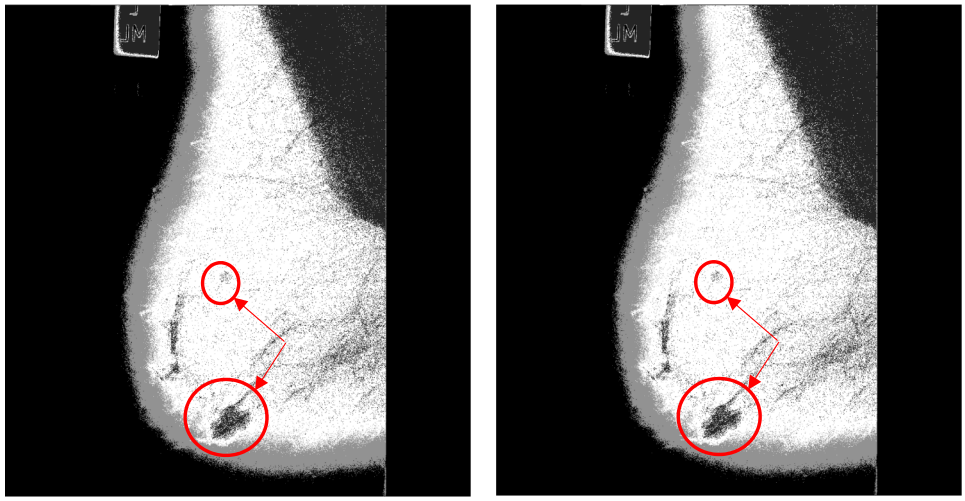
*(a) Original Image* *(b) Lumps Marked*

*(c) FCM* *(d) IFCM.*

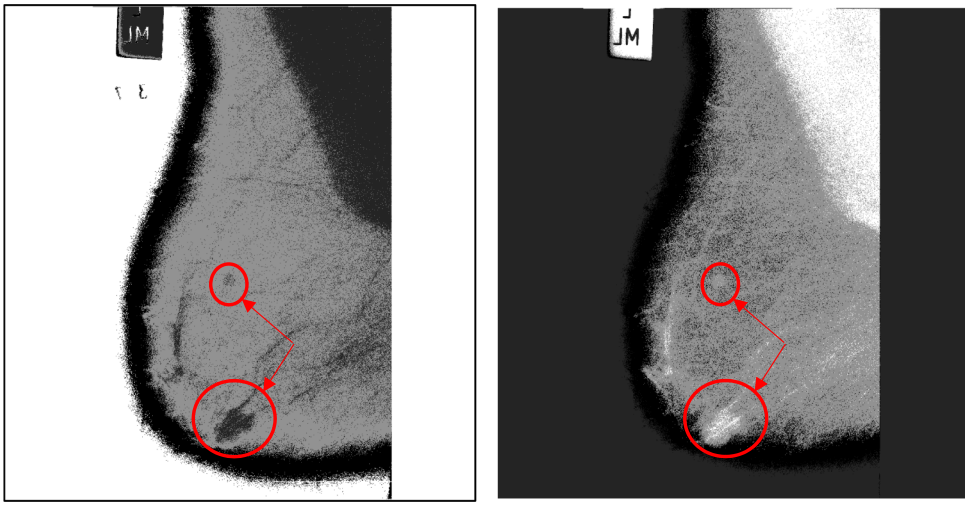
*(e) T2FCM* *(f) T2IFCM.*

**Fig. 6** Segmentation results on noisy mdb005

for clustering of data, has number of advantages over other popular algorithms—FCM/IFCM/T2FCM. Experiments are conducted on synthetic data sets and on a number of mammograms (normal as well as noisy) taken from MIAS database. Mammogram segmentation results of T2IFCM are compared with results of FCM, T2FCM, and IFCM, IT2FCM, IT2IFCM, PSO-IT2FCM, and PSO-IT2IFCM. It is observed that T2IFCM effectively segments the lumps in



*(a) IT2FCM* *(b) PSO-IT2FCM*



*(c) IT2IFCM* *(d) PSO-IT2IFCM*

**Fig. 7** Segmentation results on noisy mdb005 by interval type-2 based algorithms

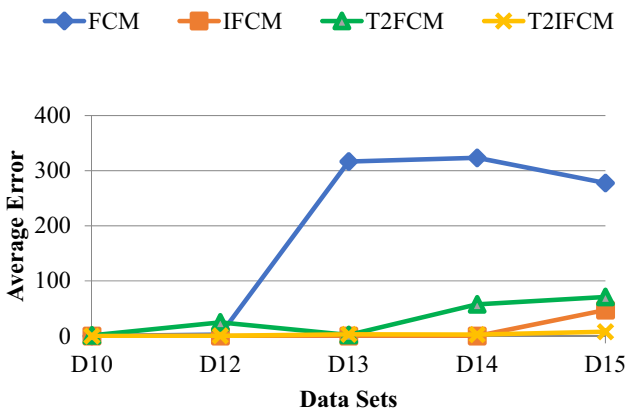
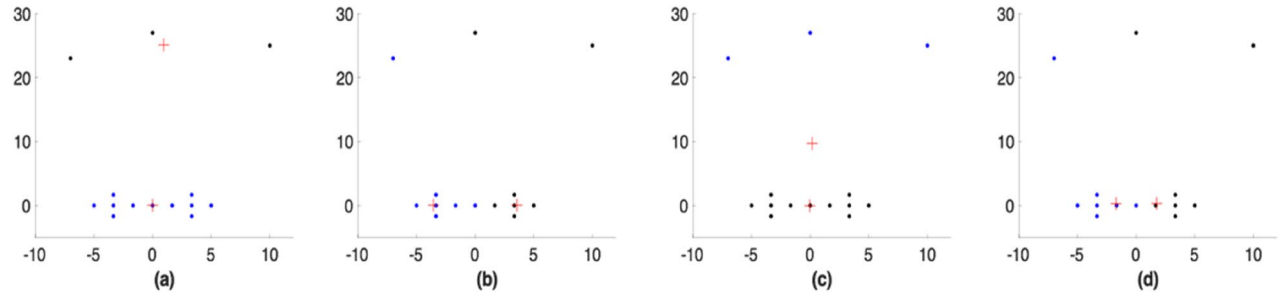



**Fig. 8** Performance comparison chart

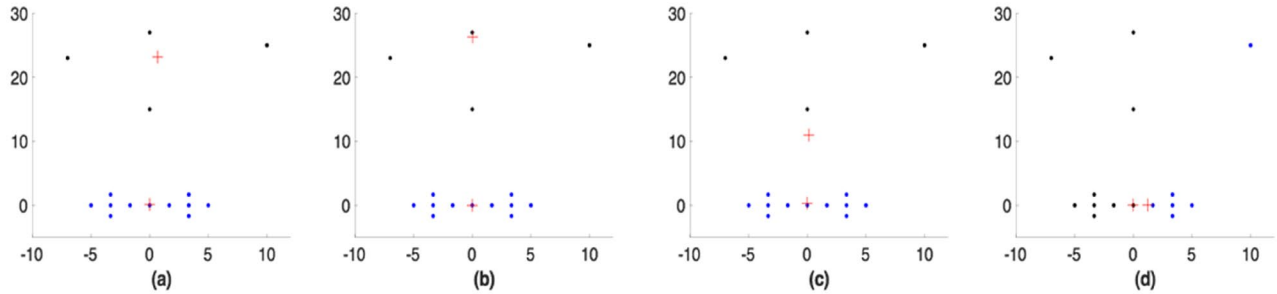
**Table 1** Resultant centroids and average error on D14

| Algorithm | Cluster1 & Cluster2 | | Average error on D14 |
|---|---|---|---|
| | cx* | cy* | |
| FCM | 0.0009220884<br>0.9450915975 | 0.0452794477<br>25.0937044825 | 323.2967007698 |
| IFCM ($\alpha = 8.55$) | −3.5657953160<br>3.5728307772 | 0.0463526489<br>0.0526587375 | 0.0550576031 |
| T2FCM | −0.0345681842<br>0.1721022034 | −0.0347763796<br>9.6986543495 | 57.5132807633 |
| T2IFCM ($\alpha = 1.25$) | −1.7090027956<br>1.7460055529 | 0.2895239093<br>0.3227248627 | 2.6944728046 |

their accurate shape and size in comparison to other algorithms. From result analysis of synthetic data sets, proposed algorithm outperforms other conventional algorithms when the data sets are highly contaminated with outliers. Thus, proposed approach is robust and effective in dealing with




*(i)D14: (a) FCM (b) IFCM (c) T2FCM (d) T2IFCM*




*(ii)D15: (a) FCM (b) IFCM (c) T2FCM (d) T2IFCM*

**Fig. 9** Results on synthetic data set: D14 and D15

**Table 2** Resultant centroids and average error on D15

| Algorithm | Cluster1 & Cluster2 | | Average error on D15 |
|---|---|---|---|
| | cx* | cy* | |
| FCM | 0.0047458445<br>0.6756914861 | 0.1227332578<br>23.1738580428 | 277.6643123309 |
| IFCM ($\alpha = 8.55$) | 0.2000000000<br>0.2000000000 | 6.0000000000<br>6.0000000000 | 47.1956000000 |
| T2FCM | −0.0218842151<br>0.1298341500 | 0.3103551466<br>10.9756094054 | 70.9376896418 |
| T2IFCM ($\alpha = 1.25$) | −0.0217766120<br>1.2226395768 | 0.0266913797<br>0.0361044729 | 7.7479187885 |

outlier contaminated data, and mammograms that has poor contrast and vague boundaries.

## Appendix: Type-II intuitionistic fuzzy c means clustering

### Proof of Type-II Intuitionistic fuzzy c means clustering

Proof: Proof of T2IFCM is given here in this Appendix. Minimization of Eq. (17), subject to constraint in Eq. (2) is done using Lagrangian method. Following is the Lagrangian equation with respect to Eq. (17):

$$J_{T2IFCM} = \sum_{j=1}^{cluster_n}\sum_{i=1}^{z}(u_{ij}^{*})^{m}dist_{ij}^{2} + \sum_{j=1}^{cluster_n}\pi_{j}^{*}e^{1-\pi_{j}^{*}} + \sum_{j=1}^{cluster_n}\lambda_{j}\left(\sum_{i=1}^{z}u_{ij}^{*}-1\right) \tag{33}$$

where $\lambda_j$ for $j = 1, 2, ..., cluster_n$ are Lagrangian multipliers.

### Partial derivation of $J_{T2IFCM}$ wrt $u^*_{ij}$

$$\frac{\partial J_{T2IFCM}}{\partial u^*_{ij}} = mu^{*(m-1)}_{ji} dist^2_{ij} + 0 + \lambda_i - 0 \tag{34}$$

Equate Eq. (34) to zero:

$$\frac{\partial J_{T2IFCM}}{\partial u^*_{ij}} = 0 \tag{35}$$

$$mu^{*(m-1)}_{ji} dist^2_{ij} + \lambda_i = 0 \tag{36}$$

$$u^*_{ij} = \left( \frac{-\lambda_i}{m.dist^2_{ij}} \right)^{\frac{1}{(m-1)}} \tag{37}$$

To fulfil the constraint in Eq. (2)

$$u^*_{ij} = \frac{u^*_{ij}}{\sum_{k=1}^{cluster_n} u^*_{ik}} \tag{38}$$

In view of Eq. (38), Eq. (37) can be written as:

$$u^*_{ij} = \frac{\left( \frac{-\lambda_i}{m.dist^2_{ij}} \right)^{\frac{1}{(m-1)}}}{\sum_{k=1}^{cluster_n} \left( \frac{-\lambda_i}{m.dist^2_{ij}} \right)^{\frac{1}{(m-1)}}} \tag{39}$$

$$u^*_{ij} = \frac{\left( \frac{1}{dist^2_{ij}} \right)^{\frac{1}{(m-1)}}}{\sum_{k=1}^{cluster_n} \left( \frac{1}{m.dist^2_{ij}} \right)^{\frac{1}{(m-1)}}} \tag{40}$$

$$u^*_{ij} = \left( \sum_{k=1}^{cluster_n} \left( \frac{dist^2_{ij}}{dist^2_{ik}} \right)^{\left(\frac{1}{m-1}\right)} \right)^{-1} \tag{41}$$

### Partial derivation of $J_{T2IFCM}$ wrt $c_j$:

Rewrite Eq. (33) as:

$$J_{T2IFCM} = \sum_{j=1}^{cluster_n} \sum_{i=1}^{z} (u^*_{ij})^m \|d_i - c_j\|^2 + \sum_{j=1}^{cluster_n} \pi^*_j e^{1-\pi^*_j} + \sum_{j=1}^{cluster_n} \lambda_j \left( \sum_{i=1}^{z} u^*_{ij} - 1 \right) \tag{42}$$

$$\frac{\partial J_{T2IFCM}}{\partial c_j} = \sum_{i=1}^{z} -2(u^*_{ij})^m \left| d_i - c_j \right|^2 + 0 \tag{43}$$

Equate Eq. (43) to zero:

$$\sum_{i=1}^{z} -2(u^*_{ij})^m \left| d_i - c_j \right| = 0 \tag{44}$$

$$\sum_{i=1}^{z} -2(u^*_{ij})^m d_i + \sum_{i=1}^{z} 2(u^*_{ij})^m c_j = 0 \tag{45}$$

$$c_j = \frac{\sum_{i=1}^{z} \left( u^{*m}_{ij} d_i \right)}{\sum_{i=1}^{z} \left( u^{*m}_{ij} \right)} \tag{46}$$

IT2m (mathematically notated as $b^*_{ij}$) is computed as per Eq. (26) and on the basis of T2IFCM concept, Eq. (46) is updated as:

$$c_j = \frac{\sum_{i=1}^{z} \left( b^{*m}_{ij} d_i \right)}{\sum_{i=1}^{z} \left( b^{*m}_{ij} \right)} \tag{47}$$

## References


Al-Najdawi N, Biltawi M, Tedmori S (2015) Mammogram image visual enhancement, mass segmentation and classification. Appl Soft Comput J 35:175–185. https://doi.org/10.1016/j.asoc.2015.06.029

Atanassov K (2020) Interval valued intuitionistic fuzzy sets past, present and future. Stud Comput Intell 835:87–110. https://doi.org/10.1007/978-3-030-31041-7_5

Balaji PG, Srinivasan D, Tham CK (2008) Coordination in distributed multi-agent system using type-2 fuzzy decision systems. IEEE Int Conf Fuzzy Syst. https://doi.org/10.1109/FUZZY.2008.4630688

Bezdek JC, Ehrlich R, Full W (1984) FCM: the fuzzy c-means clustering algorithm. Comput Geosci 10(2–3):191–203. https://doi.org/10.1016/0098-3004(84)90020-7

Bodyanskiy YV, Tyshchenko OK, Mashtalir SV (2019) Fuzzy clustering high-dimensional data using information weighting. In: Rutkowski L, Scherer R, Korytkowski M, Pedrycz W, Tadeusiewicz R, Zurada J (eds) Artificial intelligence and soft computing. Lecture notes in computer science, vol 11508. Springer, Cham

Chaira T (2011) A novel intuitionistic fuzzy c means clustering algorithm and its application to medical images. Appl Soft Comput 11(2):1711–1717. https://doi.org/10.1016/j.asoc.2010.05.005

Chintalapudi KK, Kam M (1998) Credibilistic fuzzy c means clustering algorithm. Proc IEEE Int Conf Syst Man Cybernet 2:2034–2038. https://doi.org/10.1109/icsmc.1998.728197

Clerc M (2010) Particle swarm optimization. Wiley, Chichester

Dahiya S, Nanda H, Artwani J, Varshney J (2020) Using clustering techniques and classification mechanisms for fault diagnosis. Int J Adv Trends Comput Sci Eng. https://doi.org/10.30534/ijatcse/2020/188922020

Dahiya S, Gosain A, Mann S (2021) Experimental analysis of fuzzy clustering algorithms. Adv Intell Syst Comput 1177:311–320. https://doi.org/10.1007/978-981-15-5679-1_29

Dang TH, Ngo LT, Pedrycz W (2015) Interval type-2 fuzzy c-means approach to collaborative clustering. IEEE Int Conf Fuzzy Syst. https://doi.org/10.1109/FUZZ-IEEE.2015.7337932

Dave RN (1991) Characterization and detection of noise in clustering. Pattern Recogn Lett 12(11):657–664. https://doi.org/10.1016/0167-8655(91)90002-4

Doctor F, Hagras H, Roberts D, Callaghan V (2008) A type-2 fuzzy based system for handling the uncertainties in group decisions for ranking job applicants within human resources systems. IEEE Int Conf Fuzzy Syst. https://doi.org/10.1109/FUZZY.2008.4630412

Eberhart R, Kennedy J (1995) New optimizer using particle swarm theory. In: Proceedings of the international symposium on micro machine and human science, pp 39–43. https://doi.org/10.1109/mhs.1995.494215

Elhedda W, Mehri M, Mahjoub MA (2020) Hyperkernel-based intuitionistic fuzzy c-means for denoising color archival document images. Int J Doc Anal Recogn 23(3):161–181. https://doi.org/10.1007/s10032-020-00352-2

Eltoukhy MM, Faye I, Samir BB (2010) Breast cancer diagnosis in digital mammogram using multiscale curvelet transform. Comput Med Imaging Graph 34(4):269–276. https://doi.org/10.1016/j.compmedimag.2009.11.002

Ferraro MB, Giordani P (2019) A review and proposal of (fuzzy) clustering for nonlinearly separable data. Int J Approx Reason 115:13–31. https://doi.org/10.1016/j.ijar.2019.09.004

Gosain A, Dahiya S (2016) Performance analysis of various fuzzy clustering algorithms: a review. Procedia Comput Sci 79:100–111. https://doi.org/10.1016/j.procs.2016.03.014

Gosain A, Dahiya S (2020) A new robust fuzzy clustering approach: DBKIFCM. Neural Process Lett 52(3):2189–2210. https://doi.org/10.1007/s11063-020-10345-1

Gu L, Zhang YQ (2007) Web shopping expert using new interval type-2 fuzzy reasoning. Soft Comput 11(8):741–751. https://doi.org/10.1007/s00500-006-0117-z

Gupta R, Muttoo SK, Pal SK (2017) fuzzy c-means clustering and particle swarm optimization based scheme for common service center location allocation. Appl Intell 47(3):624–643. https://doi.org/10.1007/s10489-017-0917-0

Hu F, Chen H, Wang X (2020) An intuitionistic kernel-based fuzzy c-means clustering algorithm with local information for power equipment image segmentation. IEEE Access 8:4500–4514. https://doi.org/10.1109/ACCESS.2019.2963444

Hwang C, Rhee FCH (2007) Uncertain fuzzy clustering: interval type-2 fuzzy approach to c-means. IEEE Trans Fuzzy Syst 15(1):107–120. https://doi.org/10.1109/TFUZZ.2006.889763

Ibrahim N, Fujita H, Hara T, Endo T (1997) Automated detection of clustered microcalcifications on mammograms: CAD system application to MIAS database. Phys Med Biol 42(12):2577–2589. https://doi.org/10.1088/0031-9155/42/12/021

Kaur P, Soni AK, Gosain A (2013) Robust kernelized approach to clustering by incorporating new distance measure. Eng Appl Artif Intell 26(2):833–847. https://doi.org/10.1016/j.engappai.2012.07.002

Kavitha T, Mathai PP, Karthikeyan C, Ashok M, Kohar R, Avanija J, Neelakandan S (2022) Deep learning based capsule neural network model for breast cancer diagnosis using mammogram images. Interdiscip Sci Comput Life Sci 14(1):113–29. https://doi.org/10.1007/s12539-021-00467-y

Li H, Zhuang S, Li D-A, Zhao J, Ma Y (2019) Benign and malignant classification of mammogram images based on deep learning. Biomed Signal Process Control 51:347–54. https://doi.org/10.1016/j.bspc.2019.02.017

Maitra IK, Nag S, Bandyopadhyay SK (2012) Technique for preprocessing of digital mammogram. Comput Methods Programs Biomed 107(2):175–188. https://doi.org/10.1016/j.cmpb.2011.05.007

Moodley R, Chiclana F, Caraffini F, Carter J (2020) A product-centric data mining algorithm for targeted promotions. J Retail Consum Serv. https://doi.org/10.1016/j.jretconser.2019.101940

Munusamy S, Murugesan P (2020) Modified dynamic fuzzy c-means clustering algorithm – application in dynamic customer segmentation. Appl Intell 50(6):1922–1942. https://doi.org/10.1007/s10489-019-01626-x

Nancy P, Muthurajkumar S, Ganapathy S, Santhosh Kumar SVN, Selvi M, Arputharaj K (2020) Intrusion detection using dynamic feature selection and fuzzy temporal decision tree classification for wireless sensor networks. IET Commun. https://doi.org/10.1049/iet-com.2019.0172

Nguyen DD, Ngo LT, Pham LT (2014) Interval type-2 fuzzy c-means clustering using intuitionistic fuzzy sets. In 2013 3rd World Congress on Information and Communication Technologies, WICT 2013, p. 299–304. doi: https://doi.org/10.1109/WICT.2013.7113152

Oner SC, Oztaysi B (2018) An interval type 2 hesitant fuzzy MCDM approach and a fuzzy c means clustering for retailer clustering. Soft Comput 22(15):4971–4987. https://doi.org/10.1007/s00500-018-3191-0

Ouchicha C, Ammor O, Meknassi M (2019) Unsupervised brain tumor segmentation from magnetic resonance images. In Proceedings - 2019 International Conference on Wireless Networks and Mobile Communications, WINCOM 2019. doi: https://doi.org/10.1109/WINCOM47513.2019.8942589

Pal NR, Pal K, Keller JM, Bezdek JC (2005) A possibilistic fuzzy c-means clustering algorithm. IEEE Trans Fuzzy Syst 13(4):517–530. https://doi.org/10.1109/TFUZZ.2004.840099

Parvathavarthini S, Deepa D (2021) A hybrid artificial neural network classifier based on feature selection using binary dragonfly optimization for breast cancer detection. IOP Conf Series Mater Sci Eng 1055(1):012107. https://doi.org/10.1088/1757-899x/1055/1/012107

Patil SD, Ragha L (2020) Adaptive fuzzy-based message dissemination and micro-artificial bee colony algorithm optimised routing scheme for vehicular ad hoc network. IET Commun 14(6):994–1004. https://doi.org/10.1049/iet-com.2019.0388

Phu VN, Dat ND, Tran VTN, Chau VTN, Nguyen TA (2017) Fuzzy c-means for English sentiment classification in a distributed system. Appl Intell 46(3):717–738. https://doi.org/10.1007/s10489-016-0858-z

Rajput A, Kumaravelu VB (2019) Scalable and sustainable wireless sensor networks for agricultural application of internet of things using fuzzy c-means algorithm. Sustain Comput Inf Syst 22:62–74. https://doi.org/10.1016/j.suscom.2019.02.003

Rajput A, Kumaravelu VB, Murugadass A (2020) Smart monitoring of farmland using fuzzy-based distributed wireless sensor networks. In: Subramanian B, Chen SS, Reddy K (eds) Emerging technologies for agriculture and environment. Lecture notes on multidisciplinary industrial engineering. Springer, Singapore

Rhee FCH, Hwang C (2001) A type-2 fuzzy c-means clustering algorithm. In Annual Conference of the North American Fuzzy Information Processing Society - NAFIPS 4, p. 1926–1929. doi: https://doi.org/10.1109/NAFIPS.2001.944361

Rhee FCH, Hwang C (2001) A Type-2 Fuzzy C-Means Clustering Algorithm. In Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569), vol. 4, p. 1926–29. IEEE

Rodríguez Ramos A, Bernal de Lázaro JM, Prieto-Moreno A, da José Silva Neta A, Llanes-Santiago O (2019) An approach to robust fault diagnosis in mechanical systems using computational intelligence. J Intell Manuf 30(4):1601–15. https://doi.org/10.1007/s10845-017-1343-1

Rodríguez-Ramos A, da JoséSilva Neto A, Llanes-Santiago O (2018) An approach to fault diagnosis with online detection of novel

faults using fuzzy clustering tools. Expert Syst Appl 113:200–212. https://doi.org/10.1016/j.eswa.2018.06.055

Singh AK, Gupta B (2015) A novel approach for breast cancer detection and segmentation in a mammogram. Procedia Comput Sci 54:676–682. https://doi.org/10.1016/j.procs.2015.06.079

Tongbram S, Shimray BA, Surajkumar Singh L, Dhanachandra N (2021) A novel image segmentation approach using FCM and whale optimization algorithm. J Ambient Intell Hum Comput. https://doi.org/10.1007/s12652-020-02762-w

Tot T, Tabár L, Dean PB (2000) The pressing need for better histologic-mammographic correlation of the many variations in normal breast anatomy. Virchows Arch 437(4):338–344. https://doi.org/10.1007/s004280000301

Verma H, Gupta A, Kumar D (2019) A modified intuitionistic fuzzy c-means algorithm incorporating hesitation degree. Pattern Recogn Lett 122:45–52. https://doi.org/10.1016/j.patrec.2019.02.017

Wen J, Xuan S, Li Y, Peng Q, Gao Q (2020) Image segmentation algorithm based on neutrosophic fuzzy clustering with non-local information. IET Image Proc 14(3):576–584. https://doi.org/10.1049/iet-ipr.2018.5949

Xiangxiao L, Honglin O, Lijuan Xu (2019) Kernel-distance-based intuitionistic fuzzy c-means clustering algorithm and its application. Pattern Recognit Image Anal 29(4):592–597. https://doi.org/10.1134/S1054661819040199

Xu J, Feng G, Zhao T, Sun X, Zhu M (2019) Remote sensing image classification based on semi-supervised adaptive interval type-2 fuzzy c-means algorithm. Comput Geosci 131:132–143. https://doi.org/10.1016/j.cageo.2019.06.005

Zhao F, Chen Y, Liu H, Fan J (2019) Alternate PSO-based adaptive interval type-2 intuitionistic fuzzy c-means clustering algorithm for color image segmentation. IEEE Access 7:64028–64039. https://doi.org/10.1109/ACCESS.2019.2916894

Zheng L, Chan AK (2001) An artificial intelligent algorithm for tumor detection in screening mammogram. IEEE Trans Med Imaging 20(7):559–567. https://doi.org/10.1109/42.932741

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

# A Passive Grounded MOSCAP- Memrisor Emulator

Rahul Kumar Gupta, IEEE Student Member
Research Scholar
Electronics and Communication Engineering
Delhi Technological University- India
rahulkumargupta_2k21phdec08@dtu.ac.in

Prof. M. S. Chaudhary, Professor
Electronics and Communication Engineering
Delhi Technological University
India
msc_1976@yahoo.com

Dr. Sachin Taran, Assistant Professor
Electronics and Communication Engineering
Delhi Technological University- India
sachintaran@dtu.ac.in

Dr. Varun Saxena, Assistant Professor
Electronics and Communication Engineering
JNU-Delhi, India
varunsaxena@jnu.ac.in

2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) | 978-1-6654-9781-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CONECCT55679.2022.9865831

***Abstract*— In this paper, a grounded Passive MOSCAP-Memristor Emulator is presented consisting of only three MOSFETs and one MOSCAP without biasing. Earlier the researchers have presented memristor emulators consisting of existing blocks and passive elements. The proposed circuit's structure is very simple compared to the existing one. This circuit is simulated on Cadence Virtuoso GPDK 180nm technology with sinusoidal and pulse input signals. The operating frequency of this circuit is up to 30MHz. This circuit analyzed the hysteresis loop and variation of current according to the voltage direction. The expected figure of merits of the memristor is successfully obtained and presented the memristor application in filters.**

***Keywords— MOSCAP, Memristor, Emulator, Grounded Memristor***

## I. INTRODUCTION

A memristor is defined as an electrical component that limits the flow of electrical current and retains or recollects the amount of charge that has flowed previously. It has two terminals. Memristors are a prevalent topic from a research point of view because they keep memory without power and are non-volatile.

Chua has introduced a component named memristor. It is placed as the fourth fundamental element [1,2] and left a remark for further research. Leon Chua found the first memristor in 1971 [1]. After a long gap, Hewlett Packard (HP) researchers demonstrated the first physical TiO2 memristor [3]. Memristor is used in many different applications, e.g., programmable circuits [4], neuro applications [5], chaotic circuits [6], non-volatile memories [7], etc., because of its non-volatile characteristics. Researchers are still focusing on memristor emulator circuits using MOSFET because of the fabrication complexity and high cost. Further, it is still not commercialized. Many researchers have presented a lot of emulators of memristor, which consists of active elements and existing circuits. The existing blocks used for the implementation of the memristor emulator are OPAMP (Operational Amplifier), OTA (Operational-Transconductance-Amplifier), DDCC (Differential Difference Current Conveyor), CCII (Second-generation Current Conveyor), and CFOA (Current Feedback Operational Amplifier) [8-16].

These existing emulator circuits are complex in structure and require passive components. John et al. [17] have been presented a memristor circuit consisting of active (BJT) and passive components without external biasing and working frequency up to a few kHz. Eila Gale [18] has described the idea to produce different non-idealizes of memristor, including memory conversation theory. Babacan [19] has presented a straightforward structure of memristor emulator consisting of only three N-type MOSFETs and one capacitor. It is entirely passive with a frequency range of up to 100KHz. One researcher has introduced a DTMOS based memristor emulator (floating) operating on shallow frequency [20]. The First DTMOS was highlighted by Assaderaghi et al. [21], in which the body of MOS is attached with the gate terminal to obtain DTMOS (Dynamic Threshold MOS). Vista and Ranjan [22] have also presented a simple memristor circuit consisting of three MOSFETs and biasing, an external current source, and a capacitor. Srivastav, Gupta et al. [23] have also presented a floating MOS-only memristor operating in passive mode.

The proposed paper presented a grounded memristor emulator consisting of four MOSFETs. Here, one MOSFET is used as a Capacitor (MOSCAP). It is a straightforward, simple structure, and static power consumption is zero because of no external biasing. This emulator circuit is simulated on Cadence Virtuoso GPDK 180nm technology, and the expected figure of merit has been successfully obtained and presented. The operating frequency of the proposed circuit is up to 30Mhz, and the static power consumption will be zero because of its passive nature. This paper also presents some applications of this memristor emulator circuit in analog and digital domains.

Further, section II describes the proposed emulator circuit and its working. Section III describes the simulation results such as hysteresis loop, transient response, process corner analysis, temperature analysis, and non-volatility of the proposed circuit. The last, sections IV, describes the conclusion of the presented paper.



978-1-6654-9781-7/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:56:50 UTC from IEEE Xplore. Restrictions apply.

## II. Proposed Emulator Circuit

This paper presents a circuit shown in Fig. 1 consisting of three MOS transistors, N1, N2, and N3, and one MOSCAP P1, which satisfies the characteristics of a memristor. N1, N2, N3 are NMOS transistors, and P1 is PMOS transistors. This circuit has nonlinear characteristics, and it also has a pinched hysteresis loop for the sinusoidal input signal. This emulator circuit is proposed as a passive grounded memristor, so external biasing is no use. N1, N2 are controlled by MOSCAP, and N3 has controlled the charging and discharging of MOSCAP. This MOSCAP behaves like a passive capacitor and attains a nonvolatile nature. The W/L ratio of each transistor is represented in Table 1.

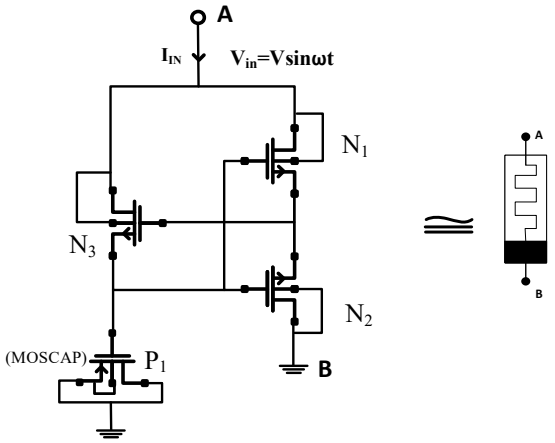



Fig. 1. Circuit Diagram of Presented Memristor Emulator

Table 1 W/L Ratios of NMOS and PMOS

| Transistor | W/L Ratio (in µm/ µm) |
|---|---|
| N1 | 4/2 |
| N2 | 10/1 |
| N3 | 1/1 |
| P1 | 10/10 |

## III. Simulation Results

Here, we have analyzed all three fingerprints of the memristor and simulated using 180nm GPDK CMOS technology (cadence spectre software). First, we have examined the relationship between memristor voltage and current. The signals such as voltage and current are measured from the input terminal because it is a grounded memristor in which the other terminal of the memristor is grounded. The V-I relationship is presented in Fig. 2, which exhibits a hysteresis loop at different frequencies 100KHz to 30MHz. The input signal (sinusoidal with 1.5Vpp) is applied to the emulator circuit. Here we have also presented the input voltage and respective current in Fig. 3 at 1MHz frequency in which phase difference is zero.

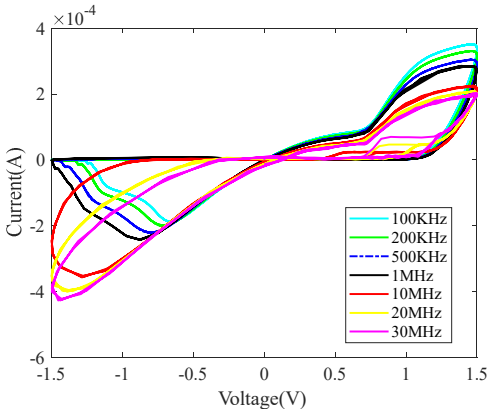



Fig. 2. Simulated Hysteresis Loop ( at different frequencies)

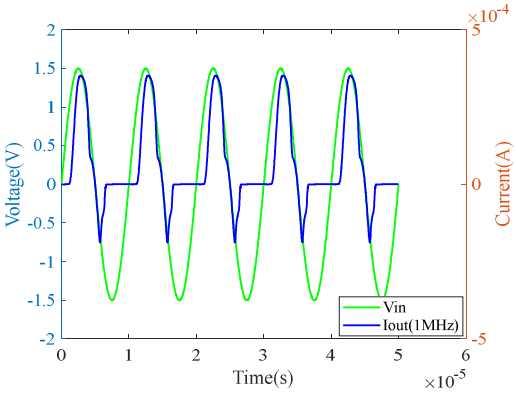



Fig. 3. Transient Response (at f = 1MHz)

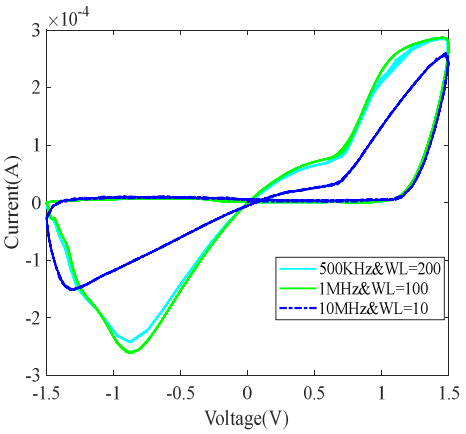



Fig. 4. Simulated Hysteresis Loop ( at different frequencies with different MOSCAP W/L Ratio)

In Fig 4, we have presented the second fingerprint at different MOSCAP W/L ratios for varying the capacitance value concerning frequency and obtained the expected results. We have given the behavior of the memristor emulator at different W/L ratios of MOSCAP for observing the variation of the MOS capacitance effect in Fig. 6. We have also presented



978-1-6654-9781-7/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:56:50 UTC from IEEE Xplore. Restrictions apply.

the third fingerprint of the memristor that is non-volatile in Fig. 5 with a pulse signal (amplitude 1V, Pulse width 500ns, period 10µs, and simulation time 200µs) obtained the rising staircase voltage across the MOSCAP. The non-volatile quality justifies the memristor memory.

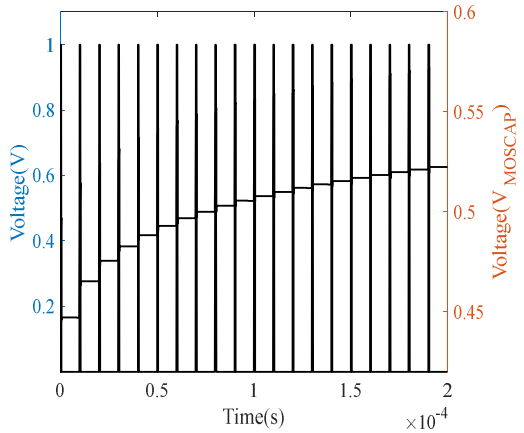



Fig. 5. Non-volatility of Memristor

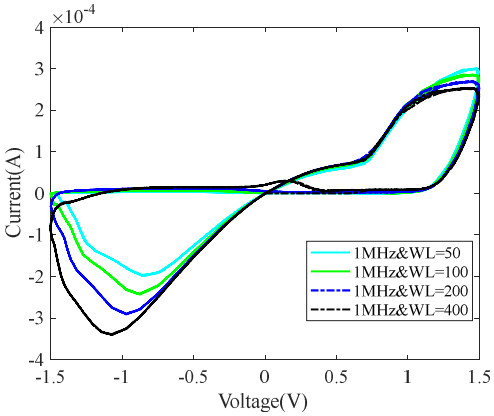



Fig. 6. Simulated Hysteresis Loop ( at different MOSCAP W/L Ratio)

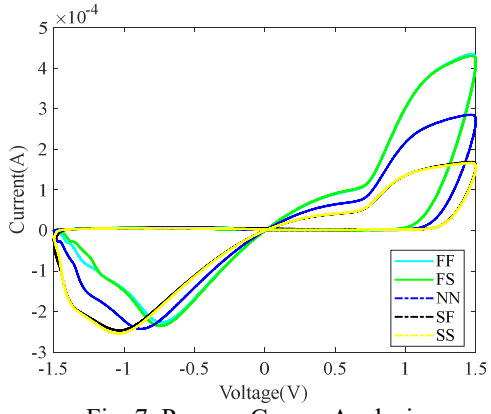



Fig. 7. Process Corner Analysis

We have presented the different corners (FF, SS, NN, SF, FS) in which F stands for Fast, S stands for Sl,,ow and N stands for Normal for process corner analysis. These corner simulations are shown in Fig. 7. The performance of memristor at different temperatures (-10℃ to 70℃) has also been demonstrated in Fig. 8.

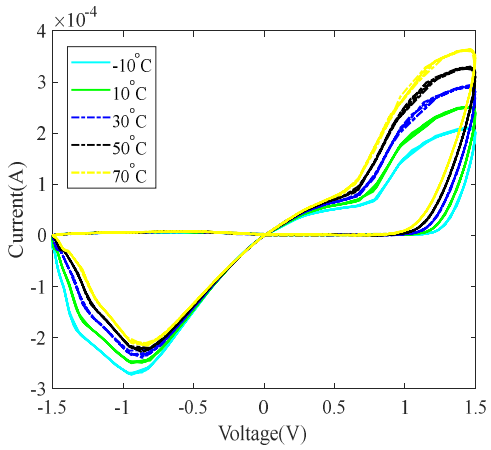



Fig. 8. Performance of Proposed Circuit at different temperature

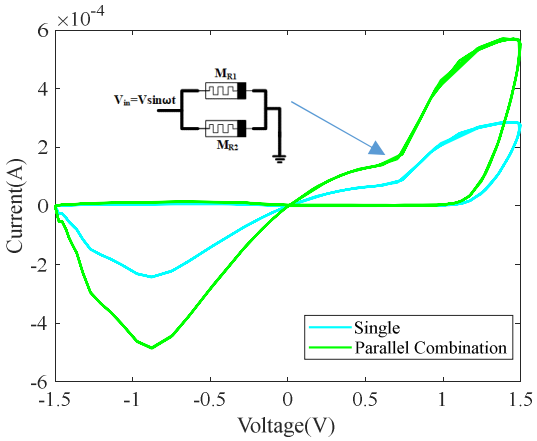



Fig. 9. Two Memristor Parallel Combination Effect on Hysteresis Loop

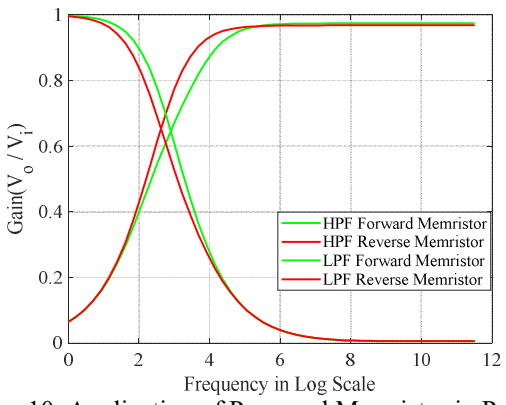


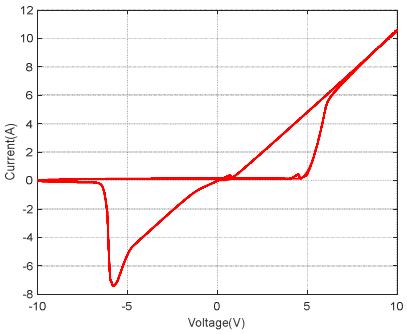

Fig. 10. Application of Proposed Memristor in Passive High Pass and Low Pass Filter



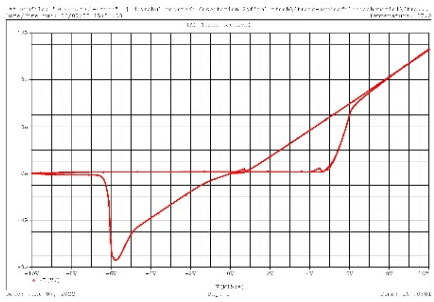
978-1-6654-9781-7/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:56:50 UTC from IEEE Xplore. Restrictions apply.

In Fig. 9, we have presented the behavior of a parallel combination of two memristors and observed the given result with a single memristor emulator. The memristance of the parallel combination is lower than the single memristor. Here, we have only presented only parallel combination because it is a grounded memristor emulator circuit. We have analyzed one application of the proposed memristor in passive filter circuits (LPF and HPF). The response of filters (Low Pass and High Pass) have been shown in Fig. 10 and obtained the expected results. Additionally, the proposed grounded memristor emulator is verified experimentally on Cadence Orcad-Capture using the existing MOSFET IRF840 at 100KHz frequency with 22nF capacitor and it is shown in Fig.11(a-b) MATLAB plot as well as orcad-capture result. The results validate the memristor model.




(a)



(b)

Fig. 11. (a) Experimental Result Plotted in MATLAB (b) Orcad-Capture Result using IRF840

Table 2. Comparison Table

Subsequently, we compared the proposed memristor emulator circuit with previous fully floating, partially floating, or grounded studies. This comparison is listed in Table 2, contains no transistors used, floating or grounded, no passive components used, simulated/experimented results, and operating frequency. Table 2 details a comparative summary of selected previous MREs [16, 19, 22, 23-27] based on active/passive elements, circuit structure, technology, power and frequency range. MOSFET counts is more in [16, 25, 26, 27]. The operating frequency is not much in [16, 19, 22, 24, 26]. This work presents 4-MOSFET MRE utilizing MOSCAP which operates on high frequency and zero static power.

## IV. Conclusion

In this work, an alternate approach memristor emulator circuit consists of MOSCAP as a driving capacitor for N1 and N2 MOSFETs. Here, only four transistors, N1, N2, N3, and P1, are proposed in the circuit with no external biasing. Since there is no external biasing, it is an entirely passive memristor. It is a very simple structure compared to others presented in the literature. Some of them are using existing blocks such as multipliers for nonlinearity behavior. The simulated favorable results of the proposed circuit have been given. Process corner and temperature variation analyses were also performed and presented in this paper. At the end of this paper, a simple application of memristor as a low pass and high filter has also been shown. In summary, all simulated results are satisfactory and compatible with earlier studies. Because of its simple structure, the proposed circuit will be very helpful in the future.

## References


[1] Chua, L. O., Memristor: the missing circuit element. IEEE Transactions on Circuit Theory, 18(5), pp. 507–512, 1971. https://doi.org/10.1109/tct.1971.1083337.

[2] L.O. Chua, S.M. Kang, “Memristive devices and systems,” In Proceedings of IEEE, vol. 64, no. 2, pp. 209-223, 1976. https://doi.org/10.1109/PROC.1976.10092

[3] D. B. Strukov, G. S. Sneider, D. R. Stewart and R. S. Williams, “The missing memristor found,” Nature, vol. 453, pp. 80–83, 2008.

[4] Y. V. Pershin and M. Di Ventra, “Practical approach to programmable analog circuits with memristors,” *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 8, pp. 1857–1864, Aug. 2010.

[5] G. C. Adam, B. D. Hoskins, M. Prezioso, F. Merrikh-Bayat, B. Chakrabarti, and D. B. Strukov, “3-D memristor crossbars for analog and neuromorphic computing applications,” *IEEE Trans. Electron Devices*, vol. 64, no. 1, pp. 312–318, Jan. 2017.

[6] B. Muthuswamy and P. P. Kokate, “Memristor-based chaotic circuits,” *IETE Tech. Rev.*, vol. 26, no. 6, pp. 417–429, 2009.

[7] H. A. F. Almurib, T. N. Kumar, and F. Lombardi, “Design and evaluation of a memristor-based look-up table for non-volatile field programmable gate arrays,” *IET Circuits, Devices Syst.*, vol. 10, no. 4, pp. 292–300, 2016.

[8] Y. Babacan, A. Yesil, and F. Kacar, “Memristor emulator with tunable characteristic and its experimental results,” AEU-Int. J. Electron. Commun., vol. 81, pp. 99–104, Nov. 2017, doi: 10.1016/j.aeue.2017.07.012.

[9] Y. Babacan, F. Kaçar, and K. Gürkan, “A spiking and bursting neuron circuit based on memristor,” Neurocomputing, vol. 203, pp. 86–91, Aug. 2016, doi: 10.1016/j.neucom.2016.03.060.

[10] D. Yu, H. H.-C. Iu, A. L. Fitch, and Y. Liang, “A floating memristor emulator based relaxation oscillator,” IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 61, no. 10, pp. 2888–2896, Oct. 2014, doi: 10.1109/TCSI.2014.2333687.

[11] H. Kim, M. P. Sah, C. Yang, S. Cho, and L. O. Chua, “Memristor emulator for memristor circuit applications,” IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 59, no. 10, pp. 2422–2431, Oct. 2012, doi: 10.1109/TCSI.2012.2188957.

[12] M. T. Abuelma’atti and Z. J. Khalifa, “A new memristor emulator and its application in digital modulation,” Analog Integr. Circuits Signal




978-1-6654-9781-7/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:56:50 UTC from IEEE Xplore. Restrictions apply.

Table 2. Comparison Table

| Reference | Existing Active Devices | Total Transistors | Passive Elements | Type of Memristor | Technology File | Power Dissipation | Working Frequency |
|---|---|---|---|---|---|---|---|
| [16] | DDCC | Forty | R→2, C→1 | Grounded | 35nm TSMC | -NA- | 1MHz |
| [19] | - | Three | C→1 | Grounded | 180nm TSMC | 0 | 100KHz |
| [22] | - | Three | C→1 | Floating | 180nm TSMC | 6.725nW | 13MHz |
| [23] | - | Three | 0 | Floating/ Grounded | 180nm GPDK | 0 | 20MHz/ 30MHz |
| [24] | OTA | Seventeen | C→1 | Floating | 180nm TSMC | -NA- | 1 MHz |
| [25] | VDTA | Sixteen | R→2, C→1 | Floating | 180nm TSMC | -NA- | 50 MHz |
| [26] | | Sixteen | C→1 | Floating | 180nm TSMC | 8.05µW | Few Hz |
| [27] | - | Seven | C→1 | Grounded | 180nm TSMC | -NA- | 50MHz |
| Presented Circuit | - | Four | 0 | Grounded | 180nm GPDK | 0 | 30MHz |


[13] Process., vol. 80, no. 3, pp. 577–584, 2014, doi: 10.1007/s10470-0150660-6.

[14] M. T. Abuelma'atti and Z. J. Khalifa, "A continuous-level memristor emulator and its application in a multivibrator circuit," AEU-Int. J. Electron. Commun., vol. 69, no. 4, pp. 771–775, Apr. 2015, doi: 10.1016/j.aeue.2014.12.011.

[15] H. Sözen and U. Çam, "Electronically tunable memristor emulator circuit," Analog Integr. Circuits Signal Process., vol. 89, no. 3, pp. 655–663, 2016, doi: 10.1007/s10470-016-0785-2.

[16] Sa´nchez-Lo´pez, C., Mendoza-Lopez, J., Carrasco-AGuilar, M. A., & Mun˜iz-Montero, C. (2014). A floating analog memristor emulator circuit. IEEE Transactions on Circuits and Systems II: Express Briefs, 61(5), 309–313. https://doi.org/10.1109/TCSII.2014.2312806

[17] A. Ye¸sil, Y. Babacan, and F. Kaçar, "A new DDCC based memristor emulator circuit and its applications," Microelectron. J., vol. 45, no. 3, pp. 282–287, 2014, doi: 10.1016/j.mejo.2014.01.011.

[18] John Kalomiros, Stavros G. Stavrinides Fernando Corinto, "A two-transistor non-ideal memristor emulator." 2016 5th International Conference on Modern Circuits and Systems Technologies (MOCAST), 2016. https://doi.org/10.1109/MOCAST.2016.7495164

[19] Ella Gale, " Non Ideal Memristor for non ideal world." Physica Status solidi(a), Volume 212, issue-2, page no. 229-238, 2015. https://doi.org/10.1002/pssa.201400169.

[20] Yunus Babacan, "Memristor: Three MOS Transistor and one Capacitor." IEEE Conference Paper, December, 2017.

[21] Yunus Babacan, Farat Kacar, "FCS Based Memristor Emulator with Associative Learning Circuit Application" IU-JEEE Journal ( Volume: 17 , 2 , April 2017 ), Page(s): 3433 – 3437.

[22] Assaderaghi, F., Sinitsky, D., Parke, S. A., Bokor, J., Ko, P. K., & Chenming Hu. (1997). Dynamic threshold-voltage MOSFET (DTMOS) for ultra-low voltage VLSI. IEEE Transactions on Electron Devices, 44(3), 414–422. doi:10.1109/16.556151

[23] John Vista, Ashish Ranjan, "A Simple Floating MOS-Memristor for High-Frequency Applications." IEEE Transactions on Very Large Scale Integration (VLSI) Systems ( Volume: 27, Issue: 5, May 2019 ) Pages1186 – 1195

[24] P. Srivastava, R. K. Gupta, R. K. Sharma and R. K. Ranjan, "MOS-Only Memristor Emulator," Circuits Systems and Signal Processing, Springer, vol. 39, pp. 5848-5861, 2020.

[25] A. Yesil, "Floating Memristor Employing Single MO-OTA with Hard-Switching Behavior," J. Circuits, Syst. Comput., vol. 28, no. 02, pp. 1950026, 2019.

[26] A. Yesil, Y. Babacan and F. Kacar, "Design and experimental evolution of memristor with only one VDTA and one capacitor," IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 38, no. 6, pp. 1123-1132, Jun 2019.

[27] Y. Babacan and F. Kaçar, "Floating memristor emulator with subthreshold region," Analog Integr. Circuits Signal Process., vol. 90, no. 2, pp. 471–475, 2017.

[28] A. Yesil, "A New Grounded Memristor Emulator Based on MOSFET-C," International Journal of Electronics and Communications, vol. 91, pp. 143-149, 2018.




978-1-6654-9781-7/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:56:50 UTC from IEEE Xplore. Restrictions apply.

Social Network Analysis and Mining (2022) 12:137
https://doi.org/10.1007/s13278-022-00961-1

ORIGINAL ARTICLE

# A self-attention hybrid emoji prediction model for code-mixed language: (Hinglish)

**Gadde Satya Sai Naga Himabindu[1] · Rajat Rao[1] · Divyashikha Sethia[1]**

Received: 19 June 2022 / Revised: 27 July 2022 / Accepted: 21 August 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2022

**Abstract**
Emojis are an essential tool for communication, and various resource-rich languages such as English use emoji prediction systems. However, there is limited research on emoji prediction for resource-poor and code-mixed languages such as Hinglish (Hindi + English), the fourth most used code-mixed language globally. This paper proposes a novel *Hinglish Emoji Prediction (HEP)* dataset created using Twitter as a corpus and a hybrid emoji prediction model *BiLSTM attention random forest (BARF)* for code-mixed Hinglish language. The proposed BARF model combines deep learning features with machine learning classification. It begins with BiLSTM to capture the context and then proceeds to self-attention to extract significant texts. Finally, it uses random forest to categorize the features to predict an emoji. The self-attention mechanism aids learning since Hinglish, a code-mixed language, lacks proper grammatical rules. The combination of deep learning and machine learning algorithms and attention is novel to emoji prediction in the code-mixed language(Hinglish). Results on the HEP dataset indicate that the BARF model outperformed previous multilingual and baseline emoji prediction models. It achieved an accuracy of 61.14%, precision of 0.66, recall of 0.59, and F1 score of 0.59.

**Keywords** Emoji prediction · Hinglish · Code mixed · Deep learning · Hybrid model

## 1 Introduction

As the Internet and social media platforms have grown in popularity, individuals have become used to expressing various emotional tendencies and feelings through social media platforms. As a new visual language, emojis are essential for conveying emotion and amplifying the visual impact of short text messages (Barbieri et al. 2016). To communicate on social media platforms, people usually use informal short text messages to give them a personal touch, which results in ambiguity. Emojis reduce this ambiguity in short text messages by conveying the intended tone. For example, *see you tomorrow* , here tone is not clear, so without context, it is on the reader how he takes it. However, the use of emojis can reduce ambiguity in it. *see you tomorrow* 😁 conveys excitement and joy while *see you tomorrow* 😡 conveys anger. This ability of emojis to convey tone and emotion in short text messages made them extremely popular.

Emojis' extensive use and popularity have created a new language of symbols that is constantly expanding, with new emojis arising with varying meanings day by day. Although hundreds of emojis are available, users cannot efficiently use them due to the time-consuming task of selecting an emoji from hundreds of options. This problem has prompted research into emojis and their relationship to text (Aoki and Uchida 2011). With the advancement in natural language processing (NLP) and its applications, emoji prediction has become one of the most exciting social media research topics. The emoji prediction task attempts to predict relevant emoji(s) that can fit the context based on the text input.

However, the increasing popularity of social media platforms in linguistically distinct demographic regions proposes a new challenge for emoji prediction due to the code-mixing of languages. Code mixing is the mixing of two or more languages while communicating. Native speakers and code-mixers tend to change the script of their native language. Users write original code-mix text in Roman script,

✉ Gadde Satya Sai Naga Himabindu
gssnhimabindu@gmail.com

Rajat Rao
rajatrao006@gmail.com

Divyashikha Sethia
divyashikha@dtu.ac.in

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

Published online: 14 September 2022

resulting in no fixed grammatical norms and phonetic differences due to regional influence. It becomes quite challenging in the Indian context as there are 22 official languages and numerous phonetic variations due to diverse cultural and regional influences. Hinglish is an important code-mix language (Parshad et al. 2016). Previous research focuses on semantic analysis, but there is a need to have an emoji prediction model for Hinglish. This paper attempts to explore emoji prediction for the Hinglish language for the first time to the best of our knowledge.

The main contributions of this paper are summarized as follows:

- *Hinglish Emoji Prediction (HEP) dataset* This work creates a new HEP dataset,[1] which is a collection of Hinglish tweets having emojis from Twitter. The annotated Hinglish (Hindi–English) code-mixed dataset for emoji prediction can enable future researchers to contribute to this domain.
- *BiLSTM attention random forest (BARF) model* Proposal of a hybrid BARF emoji prediction model for Hinglish, one of the most used code-mixed languages globally. Hinglish presents challenges with no fixed grammatical rules and spelling variations due to Roman script. "Jhooth bolna galat baat hai," for example, implies that lying is wrong. Due to many possible pronunciations, the term "jhooth" can have several phonetic variants such as "jhuth," "jhoot," or "jhut." These different terms may have distinct meanings in different contexts, resulting in ambiguity. The proposed model uses BiLSTM, self-attention, and random forest to predict our dataset's top 40 frequently used emojis. Deep learning methods, along with machine learning methods, enhance prediction accuracy. Deep learning methods are good at extracting features and learning semantic expressions, which were otherwise done manually in traditional methods. BARF uses self-attention to aid learning as it helps to overcome the problem of no fixed grammatical rules in Hinglish.
- *Comparison of BARF with other models* Compared our model with various machine learning models, including Naive Bayes and random forest, as well as deep learning models such as CNNs, LSTMs, and bidirectional LSTMs (with and without attention).

The paper is structured as follows: Section 2 contains background information and relevant work on this topic. Section 3 describes the approach used in this paper to conduct the tests, including pre-processing, embeddings, and models. Section 4 lists the experimental parameters used to reproduce the work. Section 5 details the outcomes and contains conclusions based on the data.

[1] https://github.com/Himabindugssn/HEP.

## 2 Related work

Emojis are a type of ideographic character widely used on social media platforms. They help express emotions intuitively and alter the overall semantics of short text messages. Emojis not only reduce ambiguity in short text messages but also help to convey the tone. Therefore, emojis play an important role in short text messages. However, emojis do not have grammar rules, so their usage is subjective. Emojis have been the subject of a number of academic research, as discussed in the following section. Initially, the focus was on the descriptive analysis of the usage of emojis, and they were explored as emotional annotations in plain text (Vidal et al. 2016). With the advancements in natural language processing (NLP) and further research on emojis, emoji prediction received more attention.

### 2.1 Embeddings and emoji prediction

Barbieri et al. (2016a) used distributional embeddings to study emoji semantics. Using the Skip-gram model, the authors trained emoji embeddings on 100 million English Twitter tweets. The model indicated how embeddings could improve pair similarity and clustering accuracy. The proposed methodology was further used in different languages to study emoji usage (Barbieri et al. 2016, 2016b). Pohl et al. (2017) proposed a similar embedding model for computing emoji similarity. The model allowed to place related emojis together, thus making it easy for the user to use them. The neural embedding model gave a good performance for computing similarity between emojis. Eisner et al. (2016) proposed emoji2vec, pre-trained embeddings for emojis. Embedding methods like skip-gram and word2vec are not efficient in the case of infrequent emojis. The model directly trained embeddings from the Unicode description of emojis. Emoji2vec outperforms the neural embedding model proposed by Barbieri et al. (2016a). Wijeratne et al. (2017) further improved the embedding model for emojis by incorporating different representation methods, namely emoji description, emoji definition, and emoji sense labels. The authors created the EmoSim508 dataset, which is publicly available, consisting of 508 emoji pairs for performing emoji similarity calculation tasks.

### 2.2 Deep neural network and emoji prediction

For modeling emoji semantics neural models are used because of their efficiency in learning features. The majority of the neural models use word embeddings generated from FastText, word2vec, or GloVe. Xie et al. (2016) explored emoji recommendation tasks in cross-conversational systems. They proposed a hierarchical long short-term memory

(LSTM) network to model contextual information for the emoji prediction task. Their proposed model outperformed other LSTM models, namely Single-LSTM and Flattened-LSTM. Barbieri et al. (2017) first proposed an automatic method to predict emojis for given short text input. They proposed a bidirectional long short-term memory (BiLSTM) model (Graves and Schmidhuber 2005) with word and character-based representations (Ling et al. 2015) which outperformed bag-of-words (BOW) baseline and skip-gram vector average-based baseline model.

### 2.3 Attention mechanism and emoji prediction

Felbo et al. (2017) used a variant of the LSTM neural network along with an attention mechanism (Yang et al. 2016) to detect sentiment, emotion, and sarcasm through emojis. Based on this model, Barbieri et al. (2018) proposed a label-wise attention mechanism. It improved the robustness of Recurrent Neural Network(RNN) models in datasets with unbalanced distributions, as ltekin and Rama (2018) showed that they, do not perform better than SVMs on the emoji prediction task.

Guibon et al. (2018) proposed an emoji prediction model which could be trained and tested on actual data from text messaging applications. The proposed multi-label random forest (RF) classifier model with BOW/character representation outperformed BiLSTM networks.

For the multi-label emoji prediction task, Wu et al. (2018) proposed a Hierarchical neural model with an attention mechanism. The proposed model used Convolutional Neural Networks (CNN) to learn hidden word representations, a CNN and LSTM-based word encoder to learn sentence representations, and an attention mechanism. The proposed model outperformed the Support Vector Machine (SVM), CNN, and Hierarchical LSTM model in the emoji prediction task.

### 2.4 Multilingual emoji prediction

Barbieri et al. (2018) introduced a multilingual emoji prediction task at SemEval in which there were two subtasks, one for emoji prediction in English and one for Spanish. Out of the 49 participating teams, 22 teams submitted for Spanish subtask. The majority of the top-performing teams preferred CNN or LSTM-based neural networks. Hence they are considered baseline models for testing the validity of the proposed model. However, Tubingen-Oslo ltekin and Rama (2018) performed best in both tasks, which used an SVM classifier with bag-of-n-grams features. Performance metrics of these models can be seen in Tables 1 and 2 where BiLSTM + Attention represents the system proposed by NTUA-SLP Baziotis et al. (2018) and SVM with n-gram is taken from the system proposed by ltekin and Rama (2018).

On the similar lines to SemEval, EVALITA 2018 evaluation campaign proposed an emoji prediction task for Italian Language Ronzano et al. (2018). Five different teams made 12 submissions in total for this task. Most top-performing teams employed neural architectures and achieved good accuracy, especially using the BiLSTM model.

For the first time, Tomihira et al. (2018) explored the emoji prediction task for the Japanese language. They proposed an encoder–decoder with attention that outperformed CNN and RNN-based models. Another study by Tomihira et al. (2020) explored the Bidirectional Encoder Representations from Transformers (BERT) model, which performed better than the conventional model like FastText, CNN, the Attention BiLSTM. They also compared the Japanese BERT model with the English BERT model, though it scored less than the English BERT score, which may be lower due to the Japanese dataset.

Choudhary et al. (2018) addressed the issue of primarily ignored resource-poor languages for emoji prediction and sentiment analysis by creating a corpus for Hindi, Bengali, and Telugu. Choudhary et al. (2018) introduced Classification of Emojis using Siamese Network Architecture(CESNA), a twin BiLSTM network-based emoji prediction model for resource-poor languages. They trained Hindi and Telugu (resource-poor language) and English and Spanish(resource-rich language) simultaneously in the same emoji space.

Liebeskind and Liebeskind (2019) explored the emoji prediction task for the Hebrew language, considering it a single-label classification problem. They investigated different dimension reduction methods used to associate similar words to similar vector representations against machine learning algorithms. They found that common word embedding dimension reduction methods are not optimal. They showed that n-grams and character n-grams representations significantly outperform other vector representations for emoji prediction tasks in Hebrew.

Peng and Zhao (2021) explored emoji prediction for the Chinese language. They proposed an encoder–decoder model that utilizes attention for emoji prediction. BiLSTM-CNN was used for the encoder to understand the input sentence's global and local semantic information. Simultaneously the attention mechanism increases the weight of words with a significant contribution, which helps to improve prediction accuracy. The decoder used two RNNs in different directions to decode and predict the emojis.

While there are many studies on resource-rich languages, the field of code-mixed (Hinglish) text is still relatively new and unexplored. Most past research is on monolingual datasets since a large corpus of annotated data is readily available. The fundamental challenge in dealing with code-mixed situations is a lack of well-labeled datasets, besides ambiguity in code-mixed language. With the growing number

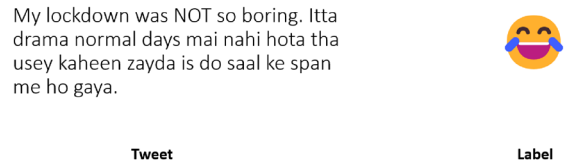



**Fig. 1** Example of text and corresponding emoji from our HEP dataset

of non-native English speakers on social media, sentiment analysis, hate speech detection on regional languages, and code-mixed data have gained traction. According to a thorough review of data from English–Hindi bilingual Facebook users (Bali et al. 2014), 17.2% of all postings, accounting for about one–fourth of the words in their dataset, exhibited some code-mixing.

While sentiment analysis (Vijay et al. 2018), Hate speech detection (Mathur et al. 2018) for Hinglish has been explored in the past, but emoji prediction has not been explored yet. Hence, this paper investigates the emoji prediction for Hinglish code-mixed language.

## 3 Proposed methodology

### 3.1 Dataset

This work proposes creating a new Hinglish Emoji Prediction (HEP) dataset for the Hinglish emoji prediction task. It comprises of retrieval of tweets containing Emojis using the Twitter API, with Twitter serving as the corpus.

The HEP dataset comprises plain text as input and emojis as labels, as shown in Fig. 1. It uses the emojis extracted from tweets as labels since manual annotation can introduce bias. The former method helps us to understand the real-world usage of emojis in a better way.

Out of 2,62,408 tweets, there are 86,072 usable tweets after filtering. The retrieved tweets are from a radius of 1000 kilometers, using Delhi(India) as the center, as the probability of obtaining Hinglish tweets is higher in this region.

The annotated tweets are divided into three categories: English, Hinglish, and Others (foreign script tweets, tweets of languages other than English and Hinglish). The following is the distribution of the three classes: Tweets in English-14,006, Hinglish-51,756, and Others-20,310.

The HEP dataset comprises tweets extracted using various hashtags such as Demonetization, NamasteTrump, Election 2019, Olympic Games, and Farmer Protest related to critical events that capture various emotions, thereby helping us maintain a balanced dataset.

Hindi keywords like aapka, accha, ajeeb, batao, chutti, dard, gussa, haal, hai, hasna, humari, koon, pareshan, pyaar, shaadi, suno, theri, wale, yaad, yaar,hai yaar, hamari, accha ,ajeeb, padhai, chup, halat, etc. are also used for extracting tweets.

### 3.2 Preprocessing

The data cleaning process removes all the unnecessary details as they are not aiding in learning, with the following steps:

- Removal of all emojis from the text.
- Removal of URLs from the text.
- Removal of all user tags, hashtags from the text.
- Removal of extra spaces from the entire text.
- Change all text characters to lower case.

It retains stop words since users switch from English to specific Hindi terms to enhance the tweet's content and personalize it. The removal of these words results in the loss of vital information. The model treats Hinglish as a stand-alone language under the proposed method without considering grammatical rules. The generated tweets are treated as individual words and phrases to generate word vector representation.

### 3.3 Proposed model

For emoji prediction, this paper proposes a BiLSTM attention random forest (BARF) model that employs BiLSTM, self-attention, and random forest, as shown in Fig. 2.

#### 3.3.1 Embedding layer

For input, the proposed model uses FastText pre-trained word embeddings. FastText is an open-source framework that enables the learning of unsupervised text representations and supervised text classification. The benefit of FastText over Word2Vec is that the former treats a single word as a character-level n-gram, whereas the latter treats a single word as a single vector in space.

The model must learn each n-gram with a unique representation so that uncommon words can share n-grams with other words. For out-of-vocabulary terms, which are present in the testing set but not in the training set, the model learns the representation of each n-gram in an earlier stage and represents a missing word as a concatenation of vectors of n-grams. Model experiments with alternative word embedding approaches such as Word2Vec, Glove, and FastText. The results indicate that FastText performs better since it can overcome vocabulary errors

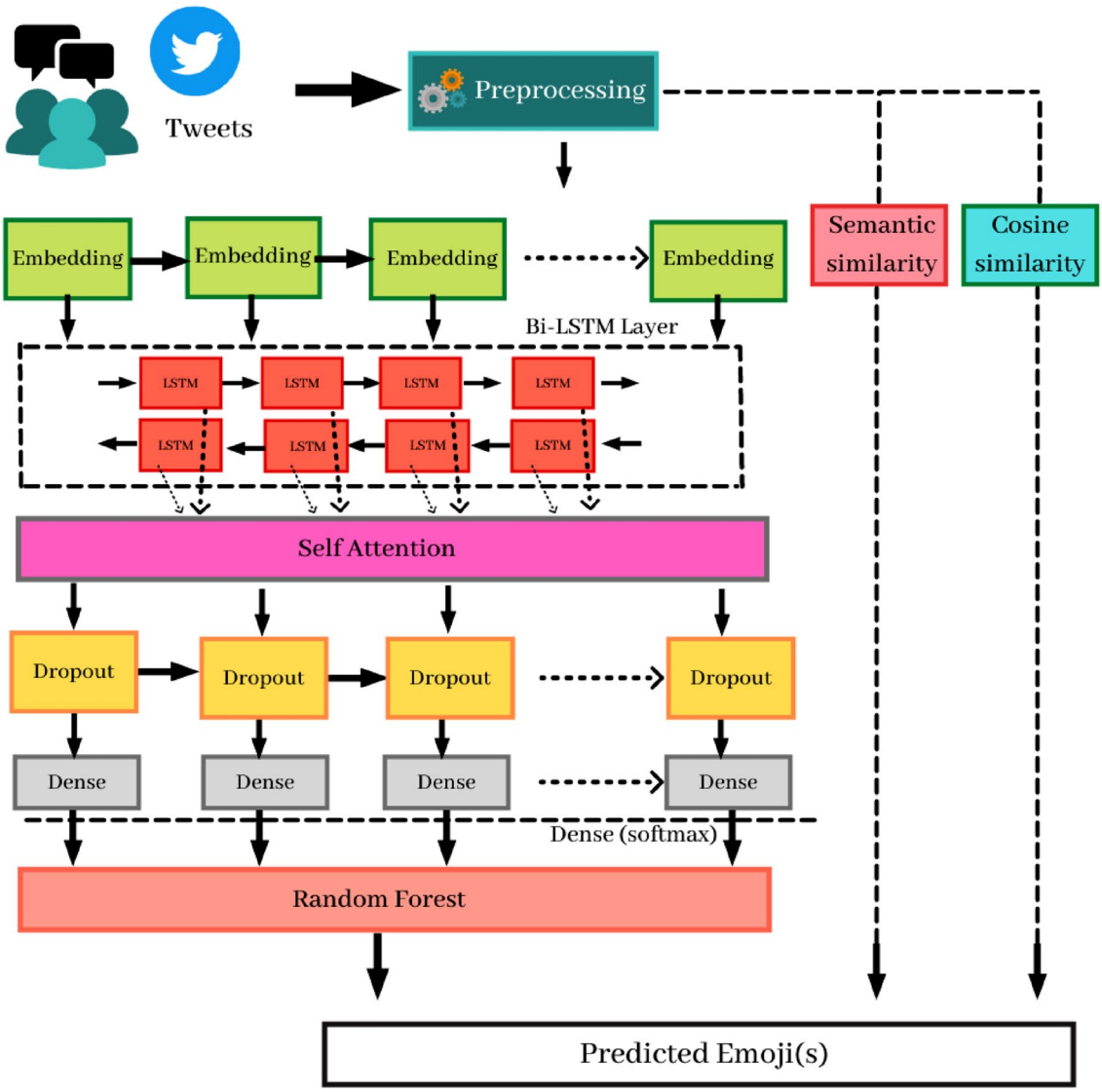



**Fig. 2** Proposed Hinglish emoji prediction model

and construct representations for unusual words common in code-mixed languages.

### 3.3.2 BiLSTM layer

The arrangement of gates in LSTM benefits from sustaining long-term dependencies. BiLSTM has an additional advantage over LSTM in that it can traverse the input data from right to left and vice versa, extracting both past and future contexts. Hence, the model adopts BiLSTM as the first layer to capture all long-term dependencies while considering past and future contexts. This information is subsequently passed on to the attention layer as shown in Fig. 3.

### 3.3.3 Attention layer

The self-attention mechanism is used to capture significant parts of the text. The proposed model uses attention mechanism proposed by Bahdanau et al. (2014). The self-attention mechanism learns the correlation between the current word and the other parts of the sentence. The self-attention mechanism increases the weight of words with significant contributions, which improves prediction accuracy. This layer hence captures the vital part of the sentence. Random forest then classifies these features. It can efficiently deal with the high dimensional data and, at the same time, avoids overfitting as it bags various decision trees. The outputs are then processed by a voting mechanism, finally resulting in the prediction of an emoji.

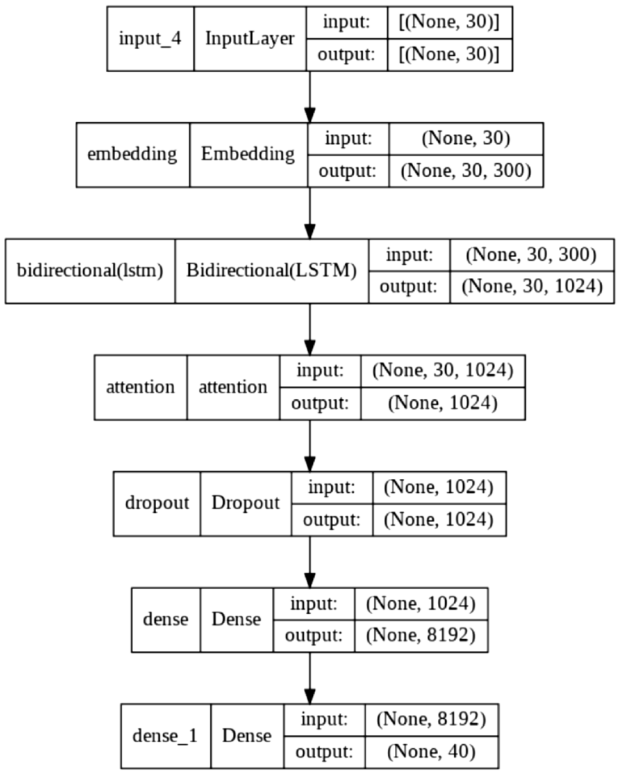



**Fig. 3** BiLSTM + Attention (Partial BARF Model Architecture)

#### 3.3.4 Random forest classifier

The information from the first fully connected layer of the BiLSTM attention model is extracted and sent to the random forest classifier to classify the features into one of the 40 emoji labels.

Random forest is a bagging method that employs the Ensemble Learning approach. It constructs as many trees as it can on the subset of data and then merges the results of all the trees. As a result, it decreases the overfitting problem in decision trees and the variance, thereby improving the accuracy. It is rapid in both model training and assessment, is resistant to outliers, captures complicated nonlinear relationships, deals with class imbalance data, and delivers competitive results for high dimensional data (Hastie et al. 2009), (Han et al. 2021). It has also been demonstrated to deal with problems caused by limited sample sizes (Qi 2012).

#### 3.3.5 Cosine and semantic similarity

Cosine Similarity and Semantic algorithms are employed to predict emojis, which can be easily identified from their meaning like cake, leaf, and fire. Algorithms use an emoji dictionary consisting of 1700 emojis and their description through which emojis can be predicted based on similarity.

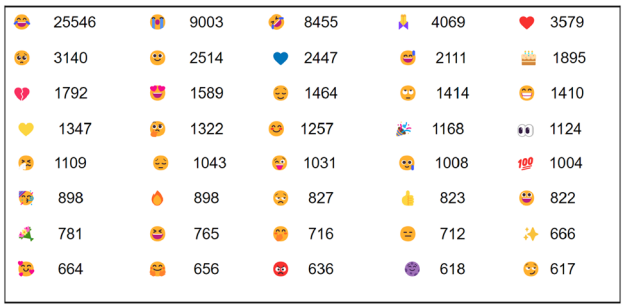



**Fig. 4** Selected emojis and their distribution

## 4 Experiments and results

The proposed BARF model predicts emoji out of the selected 40 most used emojis in the HEP dataset as shown in Fig. 4. For the proposed model, the input word length is set to 30 because the average length of a tweet is around 30 words.

The class weights are applied to the model's loss functions to tackle class imbalances, and misclassification of underrepresented classes is penalized more severely. The model weighs each class according to its inverse frequency in the training set (Baziotis et al. 2018).

**Evaluation metrics** Similar to previous emoji prediction studies, this work uses precision (P), recall (R), average F score (F), and accuracy (A) for evaluating the proposed model, which are defined in Eqs. (1, 2, 3, 4) where TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$F1 = \frac{2 \times P \times R}{P + R} F1 = \frac{2 \times P \times R}{P + R} \tag{3}$$

$$A = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}. \tag{4}$$

This paper tests the validity of the BARF model by evaluating it against the following models: random forest, logistic regression, stochastic gradient descent (SGD), Naive Bayes, XGBoost, gradient boosting, CNN, LSTM, CNN-LSTM, GRU, BiLSTM, BiLSTM+Attention, and combination of

**Table 1** Experimentation of baseline machine learning models on HEP dataset (accuracy in %)

| Model | Embeddings | | | | |
|---|---|---|---|---|---|
| | FastText | One hot encoding | CharLevel TF-IDF | WordLevel TF-IDF | N-Gram TF-IDF |
| Random Forest | 54.63 | 53.0 | 54.21 | 52.28 | 41.63 |
| Logistic Regression | 45.37 | 26.59 | 32.79 | 33.65 | 31.43 |
| Support Vector Machine | 43.82 | 9.67 | 37.64 | 41.86 | 35.82 |
| Naive Bayes | 33.41 | 2.78 | 29.95 | 30.99 | 30.67 |
| XGBoost | 32.30 | 32.33 | 29.39 | 26.25 | 28.25 |
| Gradient Boosting | 28.25 | 31.76 | 30.71 | 30.82 | 30.37 |

**Table 2** Comparison of BARF model with baseline deep learning models

| Model | Accuracy (%) | Precision | Recall | *F*-score |
|---|---|---|---|---|
| CNN | 30.00 | 0.15 | 0.27 | 0.17 |
| GRU | 30.12 | 0.18 | 0.28 | 0.19 |
| LSTM | 31.12 | 0.20 | 0.31 | 0.21 |
| CNN LSTM | 35.82 | 0.23 | 0.37 | 0.23 |
| BiLSTM | 38.80 | 0.27 | 0.32 | 0.27 |
| BiLSTM + Attention | 40.10 | 0.31 | 0.38 | 0.30 |
| CNN + BiLSTM +Attention | 40.75 | 0.31 | 0.38 | 0.31 |
| CNN + BiGRU + Attention | 39.88 | 0.30 | 0.37 | 0.29 |
| BiGRU + Attention + Random Forest | 58.85 | 0.58 | 0.57 | 0.57 |
| **BARF** | **61.14** | **0.66** | **0.59** | **0.59** |

Bold indicates the proposed model's name

them (as indicated in Tables 1 and 2). A brief description of these models is as follows:

*Random Forest (RF)* According to Xu et al. (2012), RF classifiers are well suited for coping with noisy data with a large dimension in text categorization. An RF model comprises a set of decision trees, each trained using a different set of random feature subsets. Given an instance, the RF predicts based on a majority vote of all the trees in the forest.

*Logistic Regression(LR)* Wright (1995) is a probability-based predictive analytic method. The logistic regression classifier sends the weighted combination of input characteristics via a sigmoid function. The sigmoid function can convert any real number between 0 and 1.

*Stochastic Gradient Descent(SGD)* classifier is a linear classifier trained with SGD (Kabir et al. 2015). Pure SGD tends to converge to minima with better generalization performances across various NLP problems.

*Naive Bayes(NB)* classifier is a basic classifier that classifies events based on their probability. It is based on the Bayes theorem, which states that "conditional probability is the likelihood of an outcome occurring, based on a previous outcome occurring."

*XGBoost* is a decision-tree-based ensemble machine learning algorithm that uses a gradient boosting framework (Chen et al. 2015).

*Support Vector Machine (SVM)* is a supervised machine learning algorithm employed to solve linear and nonlinear problems. SVM separates the data into classes by creating a line of the hyperplane.

*GRU* (Cho et al. 2014) is a type of RNN network that overcomes the vanishing gradient problem that is present in a regular recurrent neural network by using an update and reset gate.

*CNNs* recognize patterns across space (Kim 2014). CNNs excel at detecting position-invariant and local patterns. Patterns could be keywords expressing a certain attitude, such as "I hate," or a subject like "Harappa civilization." Thus, CNNs have risen as a paradigm architecture for text classification.

*LSTM* (Sundermeyer et al. 2012) cell helps in preserving the context and recent dependencies in the text. It can give good results even when the dataset is small.

*BiLSTM* (Graves and Schmidhuber 2005) is a bidirectional long-term and short-term memory network that is particularly well suited to modeling sequential data. Applying two LSTMs in opposite directions could better understand the context.

*BiLSTM+Attention* Zhou et al. (2016) utilizes attention mechanism to further enhance BiLSTMs feature learning process. It uses the fact that the model's attention mechanism allows it to focus on and acquire essential information; thus, it delivers good results in text classification.

**Experimental results** This work experiments with a combination of various machine learning models and embeddings as shown in Table 1 to find the best performing model. It uses embeddings, such as pre-trained Word2Vec, Glove, Word2Vec trained on the HEP dataset, and FastText embeddings. It explores machine learning models, namely random forest, logistic regression, support vector machine, Naive Bayes, XGBoost, and gradient boosting. Table 1 presents the results in terms of accuracy (A), where the random forest algorithm performs significantly better than other machine learning models. Random forest algorithm with FastText embeddings outperforms the Logistic Regression machine learning model by 9%, with the second-best results.

This work also performs experiments using standard deep learning models. The primary concept is to use the most effective machine learning and deep learning models for emoji prediction tasks. A deep learning model understands the text's context, dependencies, and essential sections. The RF machine learning model further classifies the text without overfitting. This work evaluates different deep learning models against four metrics: precision, recall, *F* score (F), accuracy (A) as shown in Table 2. The proposed BiLSTM attention random forest (BARF) model significantly outperforms the baseline models and achieves an accuracy of 61.14%, the precision of 0.66, recall of 0.59, and an F1 score of 0.59. BiLSTM attention random forest (BARF) also outperforms the commonly used BiLSTM attention model by 21% on accuracy, 35% on precision, 21% on recall, and 29% on F1 score.

In the ablation study, each component of the BARF model gives the optimal performance. As shown in Table 2, complete BiLSTM attention random forest (BARF) model performs best among all the models. Performance of the BiLSTM model increases by 2% on accuracy, 4% on precision, 6% on recall, and 3% on F1 score after the addition of attention mechanism, whereas the performance increases by 21% on accuracy, 35% on precision, 21% on recall and 29% on F score for the BiLSTM with Attention layer and RF model.

**Hyperparameter optimization** To identify optimal hyperparameter values for the BARF's deep learning model, it uses KerasTuner; and employs Scikit-RandomizedSearchCV for random forest. It uses Adam algorithm (Kingma and Ba 2014) as the optimizer and a dropout of 0.2. It adds the BiLSTM in one layer with 512 neurons and a dense layer with 8192 neurons. The loss function is Categorical Cross-Entropy, and the batch size is 32. BARF uses 200 random forest estimators and sets the number of features to 40. The proposed model obtained 61.14% accuracy on the HEP dataset with these hyperparameters.

## 5 Conclusion and future work

As emojis have become popular and their numbers have increased, there is a huge demand for the emoji prediction model. Several researchers have worked on emoji prediction systems for resource-rich languages such as English. However, there is limited attention for resource-poor and code-mixed languages. This paper offers a new Hinglish Emoji Prediction (HEP) dataset created using Twitter as a corpus and proposes a novel hybrid emoji prediction model for Hinglish, one of the most often used code-mixed languages worldwide. To the best of our knowledge, this is the first attempt to explore emoji prediction for Hinglish. The suggested approach treats the emoji prediction problem as a text classification task.

This study considers the 40 most frequently used emojis in the newly created Hinglish Emoji Prediction (HEP) dataset for the emoji prediction task. The proposed BiLSTM attention random forest (BARF) model uses a BiLSTM network which effectively increases the amount of information available to the network by providing the current and future context and passes this to the attention layer. The attention mechanism increases the weight of words, which helps to improve prediction accuracy. Further, random forest enhances the feature classification. It classifies the features from the attention layer into one of the forty emoji labels. Further, Semantic Similarity and Cosine Similarity Algorithm is used for prediction emoji can be easily identified through the words like cake, leaf, fire. The experimental findings on the HEP dataset show that the proposed BARF model has a better prediction efficacy than the competitive models, and its emoji prediction is more relevant to real life.

Continuing the Hinglish emoji prediction work further, we explored the emoji prediction task as a translation problem instead of a classification problem and used an encoder–decoder-based model to predict multiple emojis for a given plain text input (Himabindu et al. (in press)). In the future, instead of just predicting emoji, we can concentrate on further enhancing the emoji usage experience by introducing sentiment-aware emoji insertion. For given plain text input, the emoji insertion task aims at inserting relevant emojis into the text at proper positions (Lin et al. 2021; Jiang et al. 2020; Kwon et al. 2021).

## Declarations

**Conflict of interest** The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

## References


Aoki S, Uchida O (2011) A method for automatically generating the emotional vectors of emoticons using weblog articles. In: Proceedings 10th WSEAS international conference on applied computer and applied computational science, Stevens Point, Wisconsin, USA, pp 132–136

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473

Bali K, Sharma J, Choudhury M, Vyas Y (2014) "I am borrowing ya mixing?" An analysis of english–hindi code mixing in facebook. In: Proceedings of the first workshop on computational approaches to code switching, pp 116–126

Barbieri F, Ballesteros M, Saggion H (2017) Are emojis predictable? arXiv preprint arXiv:1702.07285

Barbieri F, Camacho-Collados J, Ronzano F, Anke LE, Ballesteros M, Basile V, Patti V, Saggion H (2018) SemEval 2018 task 2: multilingual emoji prediction. In: Proceedings of The 12th international workshop on semantic evaluation, pp 24–33

Barbieri F, Espinosa-Anke L, Camacho-Collados J, Schockaert S, Saggion H (2018) Interpretable emoji prediction via label-wise attention lstms. In: Proceedings of the 2018 conference on empirical methods in natural language processing; 2018 Oct 31–Nov 4; Brussels, Belgium. New York: Association for Computational Linguistics; 2018. ACL (Association for Computational Linguistics)

Barbieri F, Espinosa-Anke L, Saggion H (2016) Revealing patterns of twitter emoji usage in Barcelona and Madrid. In: Artificial intelligence research and development, pp 239–244

Barbieri F, Kruszewski G, Ronzano F, Saggion H (2016) How cosmopolitan are emojis? Exploring emojis usage and meaning over different languages with distributional semantics. In: Proceedings of the 24th ACM international conference on multimedia, pp 531–535

Barbieri F, Ronzano F, Saggion H (2016) What does this emoji mean? A vector space skip-gram model for twitter emojis. In: Proceedings of the Tenth international conference on language resources and evaluation (LREC), pp 3967–3972

Barbieri, Francesco and Espinosa-Anke, Luis and Saggion, Horacio (2016) Revealing Patterns of Twitter Emoji Usage in Barcelona and Madrid. Artificial Intelligence Research and Development IOS Press, pp 239-244

Baziotis C, Athanasiou N, Paraskevopoulos G, Ellinas N, Kolovou A, Potamianos A (2018) Ntua-slp at semeval-2018 task 2: predicting emojis using rnns with context-aware attention. arXiv preprint arXiv:1804.06657

Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H et al (2015) Xgboost. Extreme gradient boosting. R package version 0.4-2. 1(4):1–4

Choudhary N, Singh R, Bindlish I, Shrivastava M (2018) Contrastive learning of emoji-based representations for resource-poor languages. arXiv preprint arXiv:1804.01855

Choudhary N, Singh R, Rao VA, Shrivastava M (2018) Twitter corpus of resource-scarce languages for sentiment analysis and multilingual emoji prediction. In: Proceedings of the 27th international conference on computational linguistics, pp 1570–1577

Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078

Çöltekin Ç, Rama T (2018) Tübingen-oslo at semeval-2018 task 2: Svms perform better than RNNS in emoji prediction. In: Proceedings of the 12th international workshop on semantic Evaluation, pp 34–38

Eisner B, Rocktäschel T, Augenstein I, Bošnjak M, Riedel S (2016) emoji2vec: learning emoji representations from their description. arXiv preprint arXiv:1609.08359

Felbo B, Mislove A, Søgaard A, Rahwan I, Lehmann S (2017) Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv preprint arXiv:1708.00524

Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional lstm networks. In: Proceedings 2005 IEEE international joint conference on neural networks, 2005., vol 4, pp 2047–2052. IEEE

Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional lstm networks. In: Proceedings 2005 IEEE international joint conference on neural networks, 2005., vol 4, pp 2047–2052. IEEE

Guibon G, Ochs M, Bellot P (2018) Emoji recommendation in private instant messages. In: Proceedings of the 33rd Annual Acm symposium on applied computing, pp 1821–1823

Han S, Williamson BD, Fong Y (2021) Improving random forest predictions in small datasets from two-phase sampling designs. BMC Med Inform Decis Mak 21(1):1–9

Hastie T, Tibshirani R, Friedman JH, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York

Himabindu GSSN, Rao R, Sethia D (2022) Encoder-decoder based multi-label emoji prediction for Code-Mixed Language (Hindi+English). In: 2nd International Conference on Intelligent Technologies (CONIT), pp 1–6. https://doi.org/10.1109/CONIT55038.2022.9848356

Jiang H, Guo A, Ma J (2020) Automatic prediction and insertion of multiple emojis in social media text. In: 2020 International conferences on Internet of Things (iThings) and IEEE green computing and communications (GreenCom) and IEEE cyber, physical and social computing (CPSCom) and IEEE smart data (SmartData) and IEEE congress on cybermatics (Cybermatics), pp 505–512. IEEE

Kabir F, Siddique S, Kotwal MRA, Huda MN (2015) Bangla text document categorization using stochastic gradient descent (sgd) classifier. In: 2015 international conference on cognitive computing and information processing (CCIP), pp 1–4 . IEEE

Kim Y (2014) Convolutional neural networks for sentence classification. New York University. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Association for Computational Linguistics

Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980

Kwon J, Kobayashi N, Kamigaito H, Takamura H, Okumura M (2021) Making your tweets more fancy: emoji insertion to texts. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pp 770–779

Liebeskind C, Liebeskind S (2019) Emoji prediction for hebrew political domain. In: Companion proceedings of the 2019 world wide web conference, pp 468–477

Lin F, Song Y, Ma X, Min E, Liu B (2021) Sentiment-aware emoji insertion via sequence tagging. IEEE Multimed 28(2):40–48

Ling W, Luís T, Marujo L, Astudillo RF, Amir S, Dyer C, Black AW, Trancoso I (2015)Finding function in form: compositional character models for open vocabulary word representation. arXiv preprint arXiv:1508.02096

Mathur P, Sawhney R, Ayyar M, Shah R (2018) Did you offend me? Classification of offensive tweets in Hinglish language. In: Proceedings of the 2nd workshop on abusive language online (ALW2). Association for Computational Linguistics, Brussels, Belgium

Parshad RD, Bhowmick S, Chand V, Kumari N, Sinha N (2016) What is India speaking? Exploring the "hinglish'' invasion. Phys A Statist Mech Appl 449:375–389

Peng D, Zhao H (2021) Seq2emoji: a hybrid sequence generation model for short text emoji prediction. Knowl-Based Syst 214:106727

Pohl H, Domin C, Rohs M (2017) Beyond just text: semantic emoji similarity modeling to support expressive communication. ACM Trans Comput-Human Inter (TOCHI) 24(1):1–42

Qi Y (2012) Random forest for bioinformatics. In: Ensemble Machine Learning Springer Science & Business Media, pp 307

Ronzano F, Barbieri F, Wahyu Pamungkas E, Patti V, Chiusaroli F, et al (2018) Overview of the evalita 2018 Italian emoji prediction (itamoji) task. In: 6th evaluation campaign of natural language processing and speech tools for Italian. Final Workshop, EVALITA 2018, vol 2263, pp 1–9 . CEUR-WS

Sundermeyer M, Schlüter R, Ney H (2012) Lstm neural networks for language modeling. In: Thirteenth annual conference of the international speech communication association

Tomihira T, Otsuka A, Yamashita A, Satoh T (2018) What does your tweet emotion mean? Neural emoji prediction for sentiment analysis. In: Proceedings of the 20th international conference on information integration and web-based applications & services, pp 289–296

Tomihira T, Otsuka A, Yamashita A, Satoh T (2020) Multilingual emoji prediction using BERT for sentiment analysis. International Journal of Web Information Systems Emerald Publishing Limited

Vidal L, Ares G, Jaeger SR (2016) Use of emoticon and emoji in tweets for food-related emotional expression. Food Qual Prefer 49:119–128

Vijay D, Bohra A, Singh V, Akhtar SS, Shrivastava M (2018) Corpus creation and emotion prediction for hindi–english code-mixed social media text. In: Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: student research workshop, pp 128–135

Wijeratne S, Balasuriya L, Sheth A, Doran D (2017) A semantics-based measure of emoji similarity. In: Proceedings of the international conference on web intelligence, pp 646–653

Wright RE (1995) Logistic regression reading and understanding multivariate statistics American Psychological Association, pp 217–244

Wu C, Wu F, Wu S, Huang Y, Xie X (2018) Tweet emoji prediction using hierarchical model with attention. In: Proceedings of the 2018 ACM international joint conference and 2018 international symposium on pervasive and ubiquitous computing and wearable computers, pp 1337–1344

Xie R, Liu Z, Yan R, Sun M (2016) Neural emoji recommendation in dialogue systems. arXiv preprint arXiv:1612.04609

Xu B, Guo X, Ye Y, Cheng J (2012) An improved random forest classifier for text categorization. J Comput 7(12):2913–2920

Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 1480–1489

Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, Xu B (2016) Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short Papers), pp 207–212

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

Springer Nature or its licensor 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 Single-Step Subranging Relaxation Oscillator-Based Open-Loop Sigma-Delta ADC

Sudhanva Vasishta, K. R. Raghunandan, Ananth Dodabalapur, *Fellow, IEEE*, and T. R. Viswanathan, *Life Fellow, IEEE*

***Abstract*— This paper presents a novel subranging Sigma-Delta ADC using a relaxation Current-Controlled Oscillator (ICO). The instantaneous frequency of the ICO at the sampling instant determines its time resolution and hence its quantization error, while its average frequency determines its power dissipation. Therefore, running the ICO at nominally lower frequency for most of the sampling period while increasing its frequency close to the sampling instant achieves high resolution and low power consumption. This idea is similar to the strategy employed by athletes in a race where they speed-up close to the finish line to gain a clear lead from others. A prototype ADC is designed and fabricated in TSMC 180nm CMOS technology. It achieves an ENOB of 11.9 bits consuming 10$\mu$W of power from a 1.8V supply and occupies an active area of 0.06 mm$^2$, which corresponds to a Schreier FoM of 156.8 dB and a Walden FoM of 625 fJ/conversion cycle.**

***Index Terms*— Current-controlled oscillator (ICO), single-step subranging, open-loop sigma-delta ADC, continuous time, noise shaping.**

## I. Introduction

DEEP sub-micron technologies use reduced supply-voltage to limit power consumption. Here, signal conversion and processing in voltage domain gets challenging as the Signal-to-Noise ratio (SNR) is limited by the voltage headroom. On the other hand, time resolution has significantly improved due to smaller rise/fall times in short channel nodes owing to smaller parasitics. Hence, analog-to-digital conversion (ADC) in time-domain is more convenient than in voltage domain. Time-based ADCs using voltage and current controlled oscillators (VCO/ICO) have gained a lot of interest recently for various low- to high-bandwidth applications [1], [2], [3], [4], [5], [6], [7], [8]. The key idea here is to generate a time-based signal proportional to the input analog signal through a controlled oscillator and then convert it to digital domain using phase counters and registers. These architectures scale well with technology as they are implemented mostly using circuits like digital logic gates.

However, inverter ring based VCOs have non-linear voltage-to-frequency conversion and need feedback loop and/or power-hungry calibration for linearization [6], [9]. In addition, open-loop configuration [10], [11] mitigates potential instability issues while offering first-order noise shaping similar to conventional Sigma-Delta ($\Sigma$-$\Delta$) ADCs.

The open-loop $\Sigma$-$\Delta$ ADC [12] presented in this paper accepts a continuous-time input-current and combines level-crossing detection, folding and subranging techniques. In this current-controlled oscillator (ICO), a capacitor C integrates the input current $i_{in}(t)$ flowing into it, developing a voltage $v_c(t)$ across C. It ramps up linearly until it reaches $V_{REF}$, when it is reset to create a sawtooth voltage. This kind of relaxation oscillators have superior linearity characteristics and wide tuning range which are suitable for sensor front ends [13], [14], [15], [16]. However, this creates abrupt transitions in voltage at every $2\pi$ crossing that would cause the practical implementation to be susceptible to non-idealities due to capacitive and inductive effects. This is mitigated by modifying the circuit to discharge the capacitor using the input current instead of abruptly resetting it when it reaches $V_{REF}$. This creates a triangular waveform with frequency given by

$$f(t) = \frac{i_{in}(t)}{2CV_{REF}} \tag{1}$$

This signal can be represented by a radius vector rotating alternately in clockwise and anticlockwise direction. Note that the total angle swept by the radius vector, ignoring the sign, is the same as the sawtooth waveform and represents an accumulated phase of $4\pi$ for each period. This phase is captured with a phase-counter obtaining the digital representation of the integral. The derivative is then computed by taking the first difference in the digital domain to recover the signal.

A key feature of this architecture is that the analog input directly feeds the input of an integrator before being sampled and quantized by the quantizer. This provides an intrinsic anti-aliasing operation, where higher frequency contents of the input signal are significantly attenuated and removed by the integrator before being aliased by the quantizer. Therefore, this architecture does not require an explicit anti-aliasing filter with stringent linearity requirements, thereby providing significant power savings at the system level [17], [18].

Manuscript received 13 March 2022; revised 6 August 2022; accepted 27 August 2022. This work was supported by the Ranjit Shastri Trust, Allentown, PA, USA. This article was recommended by Associate Editor K. Waheed. *(Corresponding author: Sudhanva Vasishta.)*

Sudhanva Vasishta, Ananth Dodabalapur, and T. R. Viswanathan are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: sudhanva.vasishta@utexas.edu; ananth.dodabalapur@engr.utexas.edu; trv248@austin.utexas.edu).

K. R. Raghunandan is with Skyworks Solutions, Austin, TX 78701 USA (e-mail: raghunandan.kolarranganathan@skyworksinc.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSI.2022.3204507.

Digital Object Identifier 10.1109/TCSI.2022.3204507

1549-8328 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 21,2022 at 06:38:23 UTC from IEEE Xplore. Restrictions apply.

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



The resolution of this quantizer is enhanced by increasing the gain of ICO, which is achieved by scaling down C and/or $V_{REF}$, at the expense of increased power dissipation. In other words, there is a direct trade-off between power dissipation and resolution. The proposed subranging ICO-based Σ-Δ ADC leverages this trade-off by acting as both coarse (low-power low-resolution) and fine (high-power high-resolution) ADC within a sampling period. The output from the two modes are combined appropriately to realize low-power high-resolution ADC. The term subranging in this context refers to the finer approximation of the analog residue (quantization error) by switching the ADC into high-resolution mode only for a fraction of the sampling period. The idea here is to run the ICO at nominal frequency for most part of the sampling period and switch to a frequency 'm' times higher just before the sampling instant by using a smaller integration capacitor and/or lower reference voltage. This causes the oscillator to cross $2\pi$ radians more often, enabling precise computation of the derivative of the phase-angle in the digital domain at the sampling edges where quantization error is committed. The counter reading corresponding to slow and fast intervals of the ICO are appropriately scaled and combined in the digital domain to obtain the output. This achieves higher SNR while consuming lower power.

The rest of the paper is divided into four sections. Section II provides the details of the proposed system architecture. Section III gives design details of circuit level implementation of various blocks of the ADC. Measurement results of the prototype chip implemented in TSMC 180nm are presented in Section IV and finally Section V concludes the paper.

## II. Architecture Details

To simplify the anti-aliasing filter, oversampling approach was a solution for high resolution and linearity applications. The popular Σ-Δ architecture has instability issues especially in higher-order loops. Here, a new level-crossing open-loop architecture based on a frequency-modulated current-controlled relaxation-oscillator is presented. The two problems in [19] and [20] namely, the signal stuck between two levels and increasing number of levels for higher resolution are eliminated by first integrating the signal and folding it at two precise levels. Since the phase is proportional to the integral of its frequency it is readily converted into digital form using a phase-counter.

Furthermore, the signal is represented by a rotating radius vector by frequency modulating the instantaneous angular velocity $\omega$(t) of the ICO. Thus, its phase $\phi$(t) is linearly related to the integral of the signal and gets quantized at $2\pi$ radians. Hence, the oscillator is chosen to be a relaxation type with a period defined by a linear sawtooth/triangular waveform. The slope (dV/dt) then becomes analogous to the rate of change of phase (d$\phi$/dt) of the oscillator, and the amplitude of the sawtooth/triangle corresponding to $2\pi$ radians is reset at precise voltage-levels ($V_{TH_HIGH}$ and $V_{TH_LOW}$) obtained from a bandgap-reference source. In other words, the integral of the input current into the oscillator resulting in a voltage across a precise on-chip capacitor is reset by folding the integral at $V_{REF}$. This enables integration for long periods of time using short-channel CMOS processes that require reduced supply voltages. The quasi-static assumption is corroborated by adding a quiescent-current to the input signal-current. This provides a high enough free-running frequency to the ICOs oversampling the signal-bandwidth. The digital output read from the phase counter is differentiated in the digital domain to recover the signal.

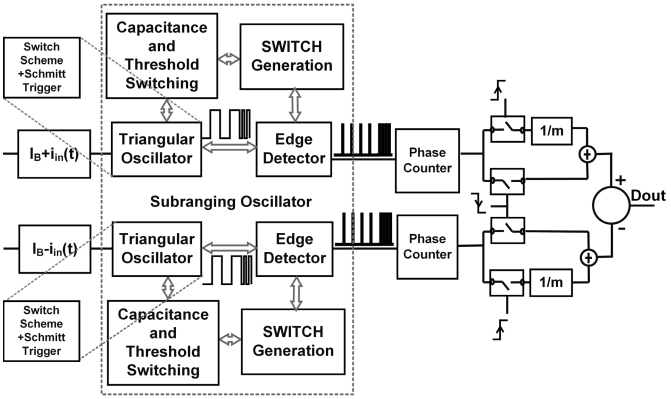



Fig. 1. Block diagram of the proposed subranging ADC architecture.

### A. Proposed Architecture

The block diagram in Fig. 1 presents the proposed continuous-time open-loop ICO-based single-step subranging ADC. It consists of two pseudo-differential oscillating ICOs operating as phase-domain integrators. Since a capacitor is used for integration, a current input with a known DC current ($I_B$) is chosen.

In the ICO, the current flows into the capacitor (C) developing a voltage $V_C$ across it. A threshold-controlled Schmitt trigger is used as a comparator to sense this voltage. As $V_C$ reaches the upper threshold of the comparator $V_{TH_HIGH}$, it changes its state and starts to discharge the current from the capacitor until $V_C$ reaches the lower threshold $V_{TH_LOW}$. (Note: $V_{REF}$=$V_{TH_HIGH}$-$V_{TH_LOW}$. Capacitor charges and discharges with Q=C$V_{REF}$). This linear charging-discharging process develops an oscillating triangular voltage waveform across the capacitor. It is converted to a square wave (ICOPULSE) whose time period is proportional to the amplitude of the triangular wave. An edge detector is used to detect the edges (ICOCLK) of the square wave.

The phase-counter captures the phase of the triangular wave by incrementing on every edge of the square wave. Rising edge indicates completion of ramp-up (0 to $2\pi$ radians) and the falling edge, the ramp-down ($2\pi$ to 0 radians) of the triangular wave. Thus, the integer number (n) in the counter at any clock-time represents $2n\pi$ radians of phase of the ICO. When the 'SWITCH' signal (derived from ICO and sampling clock) is high, a larger capacitor (C) is used for integration. At the falling edge of the sampling clock, the phase counter reading, $N_N$[n] is stored onto a register. The edge of the ICO occurring immediately after the falling edge of the sampling clock switches the integration capacitor to 'm' time smaller

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 21,2022 at 06:38:23 UTC from IEEE Xplore. Restrictions apply.

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



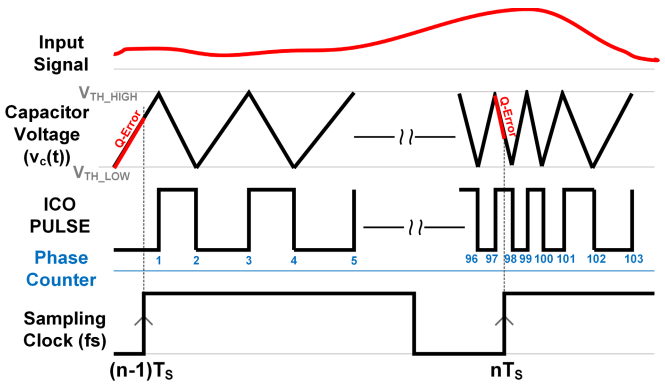



Fig. 2. Quantization error in relaxation oscillator based ADC.

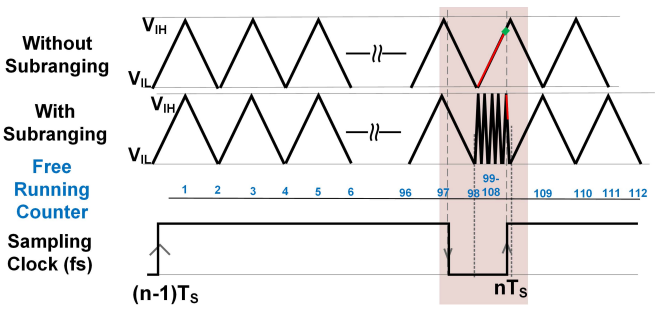



Fig. 3. Concept of subranging in relaxation oscillators.

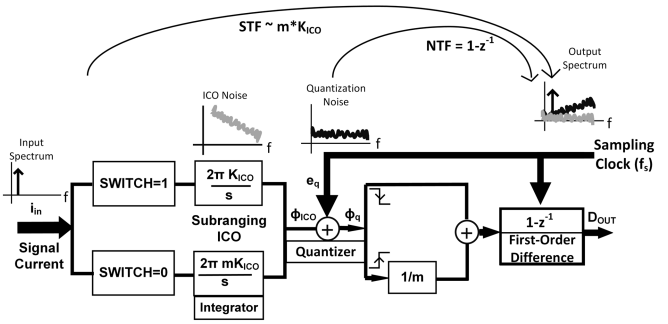



Fig. 4. Ideal model of the proposed ADC.

value (C/m). Accordingly, the conversion gain increases by 'm' times and the ICO runs at 'm' times higher frequency when the sampling clock is low. At the rising edge of the sampling clock, the phase counter reading, $N_P[n]$ is stored onto a separate register. The capacitor is switched back to its original value on the immediate or the successive edge (depending on ICO state during falling edge of sampling clock) of the ICO after the rising edge of the sampling clock.

At the sampling instance, slow count ($N_N[n]$-$N_P[n]$) and fast count ($N_P[n+1]$-$N_N[n]$) corresponding to the slow and fast intervals of the ICO are computed separately. The fast counts are scaled down by 'm' and are added to the slow counts, which effectively reduces the quantization error at each sampling instance by a factor 'm' increasing the SNR by $6\log_2 m$. Thus, by switching capacitors within each sampling period, the ADC achieves resolution akin to the one running at high speed with significantly lower power consumption.

### *B. Quantization Error*

In a controlled oscillator, the signal modulates the frequency of the oscillator. A high $K_{ICO}$ is usually chosen to sharpen the sensitivity of the system. With its usage in an ADC, the oscillator is oversampled with respect to the signal and has a frequency multiple times the Nyquist-rate of the signal bandwidth. Within each sampling period ($T_s=1/f_s$), there are several ICO clock edges ($f_{ICO} \gg f_{signal}$, $f_{ICO} \gg f_s$). The phase counter captures the phase by counting the ICO edges and is sampled periodically using a reference sampling clock.

As the phase counter only captures integral multiples of $2\pi$ radians, the analog residual piece of triangular wave voltage at the sampling edge is not accounted for by the counter. This translates into quantization error as shown in Fig. 2 at each sampling instance. At any instance, $V_{TH_LOW} \leq v_c(t) \leq V_{TH_HIGH}$. Hence, quantization error in voltage domain, $v_q = \frac{v_c(t)}{V_{TH_HIGH}-V_{TH_LOW}}$. It should be noted that the amount of information carried by '$v_q$' depends on the gain of the ICO. The proposed subranging architecture reduces this analog residue by the subranging factor 'm' (shown in green marking) with negligible power overhead by increasing the gain of the ICO by several folds for a small fraction of $T_S$ as shown in Fig 3.

### *C. ADC Model*

The ideal model of the ADC can be represented as shown in Fig. 4. As mentioned earlier, the counter value is registered twice in a given sampling period $T_S$. A 'SWITCH' signal (discussed later) is generated to toggle the ICO between fast and nominal modes. In the ICO, a signal current $i_{in}(t)$ flows into a capacitor C generating voltage

$$v_C(t) = \frac{1}{C}\int_0^t i_{in}(t)dt$$

This integrated voltage charges to $V_{TH_HIGH}$ and discharges to $V_{TH_LOW}$ linearly forming a triangular waveform. The phase counter increments by a count of 1 at either thresholds of the triangle. When SWITCH=1, the ICO has a nominal gain, $K_{ICO}=1/CV_{REF}$. The count value during this time period is given by

$$N_S[T_s] = \frac{1}{CV_{REF}}\int_0^{T_s-t_{sf}} i_{in}(t)\,dt = n_S(T_s)$$

where $t_{sf}$ is the time instant at which the ICO is switched from nominal to fast, $n_S(T_s)$ is the phase counter reading corresponding to the slow frequency of the ICO. Note that there is no quantization error committed during this period since SWITCH is toggled exactly at one of the threshold crossings of the ICO. When SWITCH=0, the gain of the ICO is switched to a 'm' times higher value ($K_{ICO,sub}=m*K_{ICO}$), the frequency increases and counter counts 'm' times faster. The count value in this time is given by

$$N_F[T_s] = \frac{1}{\frac{CV_{REF}}{m}}\int_{T_s-t_{sf}}^{T_s} i_{in}(t)\,dt = n_F(T_s) + \frac{V_q(T_s)}{V_{REF}}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 21,2022 at 06:38:23 UTC from IEEE Xplore. Restrictions apply.

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



where $n_F(T_s)$ is the phase counter reading corresponding to the 'm' times faster ICO frequency and $V_q(T_s)$ is the residual quantization error at the sampling instant. The actual count value in each sampling period $T_s=1/f_s$, is given by

$$N[T_s] = N_S[T_s] + \frac{N_F[T_s]}{m} = \left(n_S(T_s) + \frac{n_F(T_s)}{m}\right) + \frac{V_q(T_s)}{m * V_{REF}}$$

The digital output of the ADC, $D_{OUT}$ is the first difference of the successive phase counter values.

$$D_{OUT}[nT_s] = N[nT_s] - N[(n-1)T_s] = \int_{(n-1)Ts1}^{nTs1} K_{ICO} i_{in}(t)dt + \frac{V_q(nT_s) - V_q(n-1)T_s}{m * V_{REF}}$$

Since this is an oversampled system, the signal $i_{in}(t)$ is assumed to be quasi-static during each sampling period. The Z-transform representation of the final ADC output is

$$D_{OUT}(z) = K_{ICO} T_s z^{-1} I_{IN}(z) + \frac{\left(1 - z^{-1}\right) V_q(z)}{m}$$

The quantization error $V_q(z)$ is first order noise shaped and is reduced by a scaling factor 'm' at each sampling edge. With a sine wave as the input, the maximum signal amplitude at the output is $2^{N-1}$ where N in the number of phase counter bits. The maximum signal power at the output is

$$P_{sig} = \left(\frac{\Delta 2^{N-1}}{\sqrt{2}}\right)^2$$

where $\Delta$ is the resolution of the ADC. Based on the $D_{OUT}$ equation, the quantization noise will have a power at the output given by:

$$P_{Q-noise} = \int_0^{BW} \frac{\Delta^2}{12} \cdot \frac{2}{f_s} \left(\frac{V_{REF}}{m} \frac{j2\pi f}{f_s}\right)^2 df = \frac{\Delta^2}{12} \left(\frac{V_{REF}}{m}\right)^2 \frac{\pi^2}{3} \frac{1}{OSR^3}$$

Maximum attainable

$$SNR = 6.02N - 3.41 + 30log(OSR) + 20log\frac{m}{V_{REF}}$$

In the present pseudo-differential implementation, two ICO-based subranging ADCs (P- and N-side) are arranged as shown in Fig 1. Their digital outputs are independently sampled at $f_s$ and their difference is taken which doubles the dynamic range at the output. Similar mathematical analysis can be followed to come up with the below equations.

$$D_{outp}(z) = K_{ICOP} T_s z^{-1} I_{INP}(z) + K_{ICOP} T_s I_{DC} + \frac{\left(1 - z^{-1}\right) V_{qp}(z)}{m}$$

$$D_{outn}(z) = K_{ICON} T_s z^{-1} I_{INN}(z) + K_{ICON} T_s I_{DC} + \frac{\left(1 - z^{-1}\right) V_{qn}(z)}{m}$$

$$D_{out}(z) = D_{outp}(z) - D_{outn}(z)$$

### D. Effects of Non-Idealities

The effects of various non-idealities that affect the performance of the ADC have been discussed in this section. The model (VerilogA) of the ADC is first simulated without adding any non-ideality. With this as the reference, the degradation of the SNDRs with the addition of various discussed non-idealities are compared in Fig 5.

*1) Effect of Free-Running Frequency:* A known DC bias current ($I_{DC} > \max(i_{in}(t))$ is added to signal current in order to make it differential to remove even harmonic distortions. The ICO oscillates at a constant free-running frequency, $f_{fr}=I_{DC}/K_{ICO}$. Ideally, the count due to the free-running frequency gets canceled upon taking the difference as both P-side and N-side have the same $I_{DC}$ and $K_{ICO}$. However, due to the offset and mismatch variations in the current mirrors and comparator circuits, there may be a small difference in the $I_{DCP}$ and $I_{DCN}$ and their $K_{ICO}$s. Since the difference is constant to the first order approximation, this results in a constant offset in the difference in the counter values and would not affect the dynamic performance of the ADC.

*2) Effect of ICO Non-Linearity:* In a relaxation triangular oscillator, the capacitor voltage (charged using a current source) is constantly compared with a reference voltage and the direction of current is reversed as it reaches the thresholds of the comparator on either side. The comparator takes finite comparison-to-decision delay time ($t_d$) during which the voltage $v_c$ might overshoot depending on the frequency. In this subranging topology, the ICO oscillates at low frequency for a majority of the sampling period ($T_S$). In this mode, $t_d$ (<500ps in TSMC180nm technology) is less than the time-period of the triangular wave ($t_{ICO} > 200$ns). In the subranging mode, the ICO is switched to high frequency for a small fraction of $T_S$ where $t_{ICO}$ gets closer and comparable to $t_d$, the system gets affected by overshoots and non-linearity. With sufficiently high free-running frequency and the calibration technique (discussed in the Circuit design section) used, the ADC becomes mostly tolerant to the non-linearity added by the finite comparator delay. Various techniques [21], [22] can be implemented to mitigate it. In addition, with the advancement of technology to finer process nodes, the comparators have smaller delays and ICO can oscillate at higher frequencies. Signals with higher bandwidths can be processed with better performance.

*3) Effect of Mismatch in Scaled Capacitors:* In the proposed subranging topology, the ICO is switched between low and high frequency modes within a sampling period by switching between large and small integration capacitors. In order to accomplish the best performance, the scaling factor ($m_C=C_{BIG}/C_{SMALL}$) is maximized. With the parasitic capacitance affecting the $C_{SMALL}$ significantly more than the $C_{BIG}$, there will be a difference in the scaling factor obtained and designed for. This will result in inaccurate computation of the final digital output when the fast count ($N_F[nT_s]$) is scaled down by '$m_C$'. To address this, before the actual operation of the ADC, the effective scaling factor ($m_{eff}$) is computed by the digital block using a DC current in the calibration cycle and stored for digital output computation. This way, any

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 21,2022 at 06:38:23 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.



errors created due to capacitor mismatch will not affect the performance of the ADC.

*4) Effect of Setup/Hold Time of the Quantizer:* As established earlier, ICO oscillates asynchronously to the sampling clock. The phase counters increment on ICO and are periodically sampled using the sampling clock. When the edges of the ICO and the sampling clock coincide, the counters will violate either setup or the hold time requirements causing metastability and erroneous sampling. Binary counters face severe degradation in the performance when multiple bits of the counters are toggling exactly at the sampling edge whereas gray counters face only one bit of erroneous sampling [23]. However, in the proposed architecture, one-bit erroneous sampling of slow count is equivalent to '$\log_2$m' bits of erroneous sampling of the ADC. Hence, to make the ADC tolerant, the sampling clock is first double-synchronized with ICO clock and delayed by a safe time period (lesser than the lowest $t_{ICO}$ and several times greater than $t_{setup}$ & $t_{hold}$) and then used to sample the counters.

*5) Effect of Phase Noise / Jitter:* The main sources of noise in the relaxation oscillators are the noise from the input current mirrors, noise from the reference generator and the noise from the devices of the comparator. These can be modeled as equivalent noise current '$i_n$' parallel to the signal current flowing into the capacitor and an equivalent noise voltage '$v_n$' in series with the comparator or the reference voltage. The equation for the frequency of the ICO in the presence of noise sources is given by

$$\begin{aligned} N_{total}[T_s] &= \int_0^{T_s} \frac{i_{in}(t)+i_n(t)}{C(V_{REF}+v_n(t))}dt \\ &= \frac{1}{CV_{REF}}\int_0^{T_s}(i_{in}(t)+i_n(t))\left(1-\frac{v_n(t)}{V_{REF}}\right)dt \\ &= \int_0^{T_s}\frac{i_{in}(t)}{CV_{REF}}+\int_0^{T_s}\frac{i_n(t)}{CV_{REF}} \\ &\quad -\int_0^{T_s}\frac{i_{in}(t)\times v_n(t)}{CV_{REF}^2}-\int_0^{T_s}\frac{i_n(t)\times v_n(t)}{CV_{REF}^2} \\ &\approx \int_0^{T_s}\frac{i_{in}(t)}{CV_{REF}}+\int_0^{T_s}\frac{i_n(t)}{CV_{REF}}-\int_0^{T_s}\frac{i_{in}(t)\times v_n(t)}{CV_{REF}^2} \end{aligned}$$

Considering their power spectral densities,

$$\begin{aligned} &\xrightarrow{PSD}\frac{I_{in}(s)}{sCV_{REF}}+\frac{\mathcal{L}(i_n)}{sCV_{REF}}-\frac{I_{in}(s)*\mathcal{L}(v_n)}{sCV_{REF}} \\ &N_{total}[T_s] = N[T_s]+N_{jitter}[T_s] \\ &D_{out}(z) = K_{ICO}T_s z^{-1}I_{IN}(z)+\frac{\left(1-z^{-1}\right)V_{jitter}(z)}{m} \end{aligned} \tag{2}$$

The second term in Equation. (2) indicates the contribution of noise to the digital output of the phase counter. Note that the noise from the input current sources is indistinguishable with the input and goes through the same transfer function. Therefore, the devices in the current mirrors need to be sized to ensure the flicker noise is below the expected signal level or the input signal needs to be up converted by employing chopping [24] to a frequency where the noise is negligible. The PSD of the voltage noise of the comparator and reference generator, on the other hand, is convolved with the input, implying that care must be taken to ensure noise contribution from them is negligible compared to the signal. The noise from the reference generator is easily minimized by using a low pass filter at its output. The comparator noise, however, is a band limited upto a frequency determined by the delay of the comparator [25] and can be minimized by lowering its delay. To summarize, the analysis presented above suggests that with chopping and reference filtering at the system level, the noise contribution in the proposed architecture reduces with the delay of the comparator, and hence it improves as we migrate to shorter channel technologies.

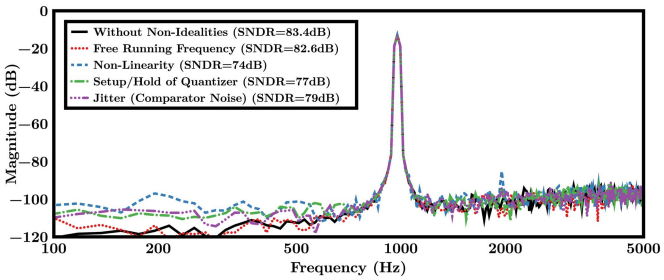



Fig. 5. Comparison of degradation in SNDR due to various added non-linearities: 5% mismatch in free running frequency, 10kΩ Noisy resistor to mimic ICO Jitter, Non-linearity due to 2ns delay in the comparator, setup/hold violation in conventional quantizer.

### *E. Choice of ICO and ADC Architecture*

The proposed ADC consists of a relaxation oscillator designed using a timing capacitor, switches and a Schmitt trigger-based comparator which consists of inverters. As the CMOS technology advances to shorter channel nodes, the availability of faster switches, smaller and accurate capacitors, smaller and faster inverters (hence comparators), offer ICOs which can operate at GigaHertz with superior linearity characteristics occupying a small area. Using such an ICO as a phase-domain integrator in a time-based Sigma-Delta ADC ensures a high oversampling ratio ($f_{ICO} \gg f_{BW}$) to process continuous-time input and relaxes the need for highly linear anti-aliasing filter upfront. Since the capacitor linearly integrates the input signal current to form a triangular voltage, the integral is available partly in digital (number of completed triangles stored in the phase counter as a number) and partly in analog (analog residual voltage at any instant which has not completed the triangle in either direction). Implementing an ADC using this ICO, the quantization error at each sampling instant i.e., the analog residue is available as it is and can be extracted without the need of a digital-to-analog converter (DAC). Several techniques can be implemented orthogonally to further approximate this error to enhance the resolution of the quantizer [26]. Implementing open-loop ADCs also avoids potential instability issues. Higher order open-loop sigma-delta ADCs can also be implemented for higher SNRs [27]. On the digital side of the ADC, counters and shift-registers suffice to process the output of the ICO to generate digital output periodically. The design details will be discussed in future sections.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 21,2022 at 06:38:23 UTC from IEEE Xplore. Restrictions apply.

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



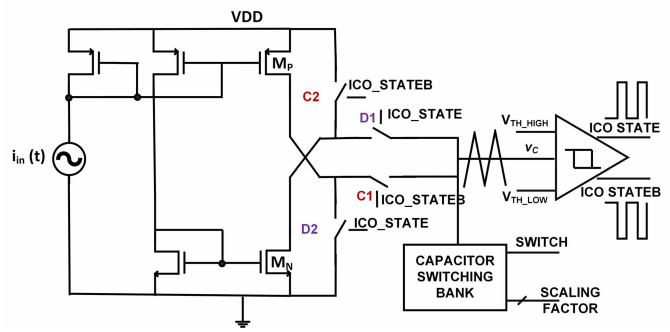



Fig. 6. Circuit diagram of subranging triangular oscillator.

## III. ADC Circuit Design

### A. Subranging Triangular Oscillator

The subranging triangular ICO is the core of this ADC. It consists of a signal current source and sink (simple current mirror shown for simplicity), capacitor switching bank, a Schmitt trigger-based comparator and a charging-discharging switch scheme as shown in Fig. 6.

The capacitor switching block determines the effective value of the integration capacitor. Based on the scaling factor input and the signal 'SWITCH' (generated internally using ICO clock and sampling clock), its value is larger while ICO operates in nominal mode and is switched to a smaller value during the subranging mode.

While the capacitor is charging, the switches C1 & C2 are ON and D1 & D2 are OFF. The output of the Schmitt trigger is LOW. The signal current flows through C1 and accumulates charges on the capacitor ramping up the voltage $v_c$ linearly. As it reaches $V_{TH_HIGH}$ (the upper threshold of the comparator), the Schmitt trigger changes its state. It switches OFF C1 & C2 and switches ON D1 & D2. The accumulated charges on the capacitor get discharged linearly through D1 and the sinking NMOS current source ramping down the voltage $v_c$. When it reaches $V_{TH_LOW}$ (the lower threshold), the Schmitt trigger again changes the state and turns OFF D1 & D2 while switching ON C1 & C2. This charging and discharging process of the capacitor linearly forms an oscillating triangular voltage across it. The rate of charging-discharging, hence the frequency of oscillation, depends linearly on the input current amplitude.

In the present implementation, C1, C2, D1 and D2 are implemented as conventional transmission gate switches with minimum dimensions to avoid parasitic capacitance on node $v_c$. The value of the integration capacitors are 1.12pF (nominal mode) and 35fF (subranging mode).

The reference voltages are set using a replica biasing arrangement (discussed in upcoming subsection). The Schmitt trigger is operated at $V_{TH_HIGH}$=1.2V and $V_{TH_LOW}$=0.6V ($V_{REF}$=1.2-0.6=0.6V). The current-to-frequency conversion gain $K_{ICO}$ are 1.48MHz/$\mu$A and 47.6MHz/$\mu$A. Since the target of this ADC (in TSMC180nm) is to be used in sensor instrumentation systems, the input signal current amplitude is chosen to be less than 3$\mu$A and the DC bias current is chosen to be 3$\mu$A.

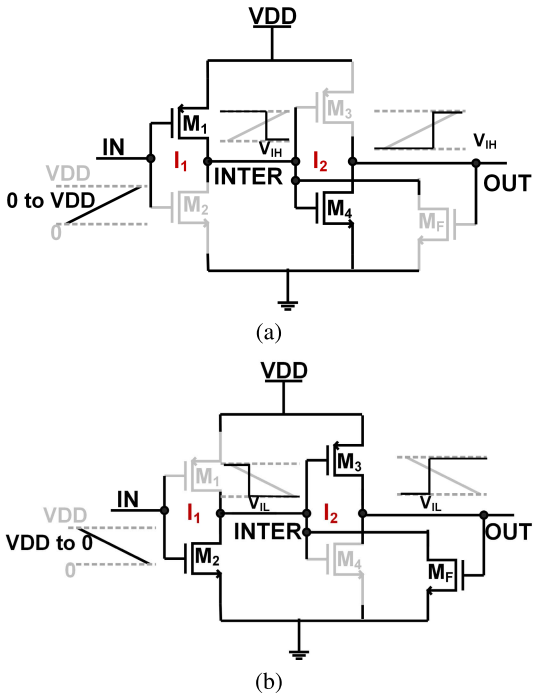



Fig. 7. State of the schmitt trigger when (a) Input goes from 0 to $V_{DD}$ (b) Input goes from $V_{DD}$ to 0.

### B. Schmitt Trigger

The basic Schmitt Trigger circuit consists of two back-to-back inverters ($I_1$ and $I_2$) with controlled positive feedback using $M_F$ around $I_2$ as shown in Fig. 7. When input IN =0, node INTER=$V_{DD}$ and output OUT=0. Hence, $V_{GS}$ of $M_F$=0 and is totally OFF. As IN increases, $M_2$ enters the subthreshold region with the state of the nodes retained until the logical switching threshold ($V_{LT}$) of $I_1$ ($V_{LT1}$) is reached (i.e. IN<$V_{LT1}$; INTER=$V_{DD}$ and OUT=0). $M_F$ remains switched off until $V_{LT1}$ is reached. The whole circuit acts like two back-to-back inverters ($I_1$-$I_2$) which toggles at $V_{LT1}$. At IN =$V_{LT1}$, the $I_1$ goes low, and the positive feedback (formed by $I_2$, and $M_F$) sets OUT to high. Hence, the comparator reference voltage during ramp-up is $V_{LT1}$.

When IN goes from $V_{DD}$ to 0. Until $V_{LTC}$ ($V_{LT}$ of the composite inverter formed by $I_1$ and $M_F$) is reached the states remain the same. When IN = $V_{LTC}$ (<$V_{LT1}$), the output of $I_1$ is pulled high and consequently the OUT goes low. The output changes state from 1 to 0.

*1) Replica Bias Scheme for PVT Invariance:* As established, the comparator reference voltages are determined by the logical thresholds of the inverters. These $V_{LT}$s in turn are determined by the mobility, dimensions, and threshold voltage of the transistors which are highly sensitive to process, supply-voltage, and temperature (PVT) variations. In order to stabilize these reference thresholds for PVT variations and set them to known precise externally controlled reference voltages, a replica-biasing scheme is used. A brief mathematical analysis on stabilizing $V_{LT}$ to precise externally set $V_{TH_HIGH}$ and $V_{TH_LOW}$ has been provided in the appendix. Two reference voltages, $V_{TH_HIGH}$ and $V_{TH_LOW}$ are supplied to the biasing

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 21,2022 at 06:38:23 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.



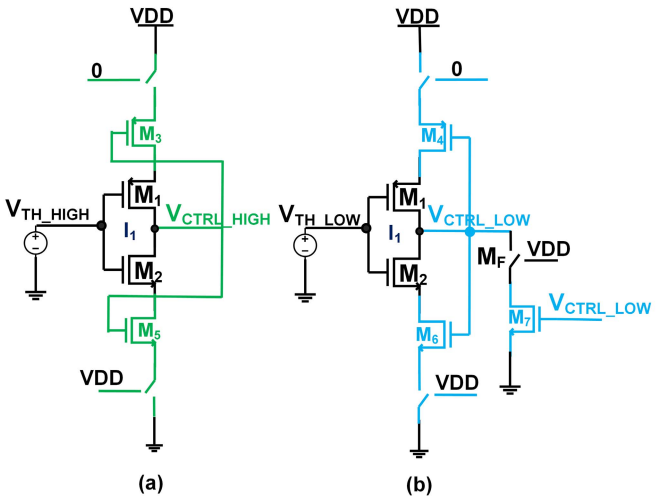



Fig. 8. Replica bias scheme to (a) Stabilize $V_{TH_HIGH}$ (b) Stabilize $V_{TH_LOW}$ for PVT variations.

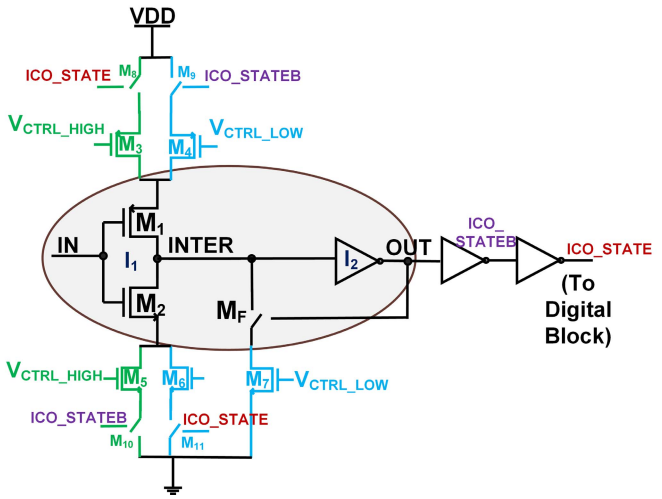



Fig. 9. Complete schematic of the reference controllable schmitt trigger.

circuits to generate two control voltages $V_{CTRL_HIGH}$ and $V_{CTRL_LOW}$ within the Schmitt trigger block as shown in Fig. 8.

The inverters $I_1$ & $I_2$ and transistor $M_F$ are modified by adding series degeneration transistors ($M_3$, $M_4$, $M_5$, $M_6$ and $M_7$) [28] to change their pull-up and pull-down strength as shown in Fig 9. During charging, ICO_STATE=0 and ICO_STATEB=1, switches $M_8$ and $M_{10}$ are ON enabling the replica bias scheme for $V_{TH_HIGH}$ using $V_{CTRL_HIGH}$ to bias $M_3$ and $M_5$. During discharging, ICO_STATE=1 and ICO_STATEB=0, switches $M_9$ and $M_{11}$ are ON enabling the replica bias scheme for $V_{TH_LOW}$ using $V_{CTRL_LOW}$ to bias $M_4$, $M_6$ and $M_7$.

### *C. Capacitor and Threshold Voltage Scaling*

This block consists of effectively two different values of capacitors namely 'C' switched during nominal mode (slower) ICO operation when SWITCH is high and 'C/$m_C$' switched

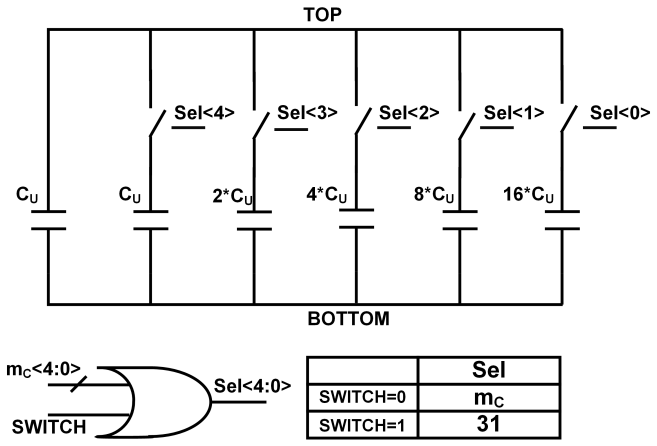



Fig. 10. Capacitor arrangement and selection during switching.

during subranging mode (faster) when SWITCH is low. In the present implementation, $C=32*C_U$ ($C_U$ is chosen to be 35fF MIM capacitor) is fixed for nominal mode ICO. For sub-ranging, scaling factor '$m_C$' (5-bit) is externally set using binary coded capacitors as shown in Fig. 10 which ensures $0<m_C<32$. SWITCH is generated synchronously to ICO.

Referring to the Equation. (1), the ICO conversion gain depends on integration capacitor 'C' and threshold voltage '$V_{REF}$'. In addition to the timing capacitor, $V_{REF}$ can also be switched to a smaller value at the SWITCH signal using a simple analog multiplexer. $V_{REF}=V_{TH_HIGH}-V_{TH_LOW}$; $V_{REF_scaled}=V_{TH_HIGH_fast}-V_{TH_LOW_fast}$. Threshold scaling factor, $m_V= V_{REF}/V_{REF_scaled}$. Total scaling factor; $m=m_C\times m_V$. The analog voltage residue at each sampling instant which contributes to quantization error is high in ADCs with low frequency ICOs. To reduce this error, the ICO frequency is increased by 'm' times, by switching into a smaller integration capacitor just before the arrival of sampling edge. This can be perceived as the same ICO and counter incrementing with a finer granularity around sampling clock edge to better approximate the analog residue as shown in Fig. 3.

### *D. Switch Generation and State Detection*

ICO is signal dependent and is asynchronous to the sampling clock. Therefore, switching a part of the capacitor bank back into the circuit at the sampling instant would create a sudden change in voltage due to charge sharing and hence create an error in the integrated voltage. In order to avoid that, the capacitors are switched at either extreme of the triangle going in the same direction. A custom digital block is designed to accomplish this operation. To begin with, the triangular waveform is converted square wave (ICOPULSE) with extremes of triangle corresponding to the edges of the square wave (maxima → rising edge; minima → falling edge). The edges of the square wave are detected using delay-XOR technique and denoted as ICOCLK.

The circuit consists of a chain of D-flip flops clocked using ICOCLK as shown in Fig. 11 for synchronization and fixed number of clock delays. The main aim is to register the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 21,2022 at 06:38:23 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.



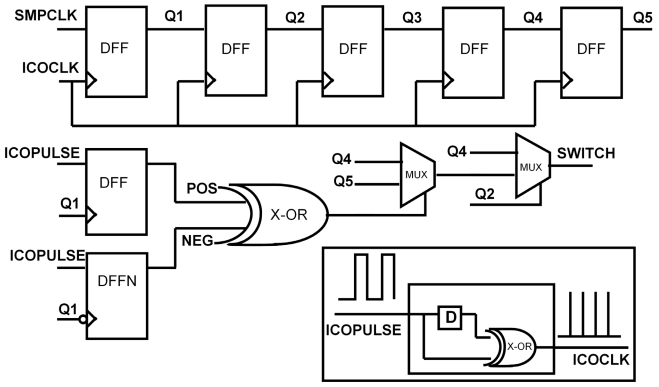



Fig. 11. Custom digital circuit for SWITCH generation.

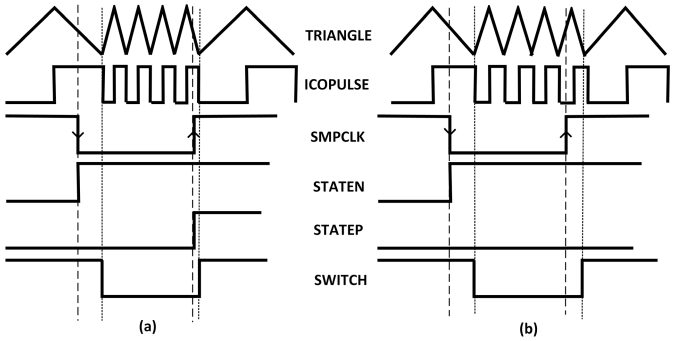



Fig. 12. Two cases of generating SWITCH signal (a) STATEP=STATEN (b) STATEP ≠ STATEN.

state of the ICOPULSE at the falling edge of the sampling clock (STATEN=0 or 1) and toggle the SWITCH at the next ICOCLK edge. At the rising edge of the sampling clock, the state of ICOPULSE (STATEP) is compared with STATEN. If STATEP=STATEN, the SWITCH signal toggles at the upcoming ICOCLK edge. If they are different, the SWITCH toggles after an additional ICOCLK edge. A graphical representation of the two scenarios is presented in Fig. 12. This information is also utilized in computing the final digital output of the ADC. In this manner, the charge imbalance during switching of capacitors is avoided.

### *E. Calibration Cycle and Choice of Scaling Factor*

As established earlier, with the idea of subranging, the quantization error of the ADC reduces by the total scaling factor ($m=m_C \times m_V$). The best performance is accomplished by maximizing the value of 'm'. However, the technology used for implementation poses the upper limit to the total scaling factor of the system. As the frequency of the ICO is increased several folds during the subranging mode, the time period of the oscillation becomes comparable with the delay of the comparator. This causes overshoots in the triangle resulting in the reduction of $f_{ICO}$ adding substantial non-linearity to the system depending on the scaling factor. In addition, in the present implementation MIM capacitors are utilized which add significant routing capacitance when $C_U$ is chosen with a small value. The parasitic capacitance at various nodes also contribute to the increased capacitance. Hence, the externally set scaling factor deviates from the actual ratios of the frequencies.

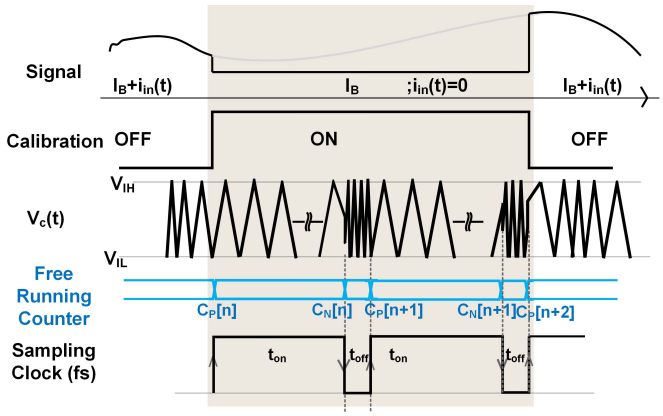



Fig. 13. Illustration of the calibration cycle to determine 'm$_{eff}$'.

A calibration cycle is run to determine the effective scaling factor without any additional hardware or trimming circuits. Before the actual operation of the system, the signal is switched off and provided with a DC quiescent current as the input and the system is run as usual. The counter readings at rising edge ($N_P[n]$) and falling edge ($N_N[n]$) are captured with the sampling clock acting as the SWITCH signal. A typical representation of waveforms is presented in Fig. 13. The duty cycle of the sampling clock is stored in digital format and the effective scaling factor 'm$_{eff}$' is computed using the formula

$$m_{eff} = \frac{t_{off}}{t_{on}} * \frac{N_P[n+1] - N_N[n]}{N_N[n] - N_P[n]}$$

which represents the actual ratio of fast to slow frequencies after the consideration of all the mentioned non-linearities. Since 'm$_{eff}$' depends on the duty cycle of the sampling period, additional averaging of 'm$_{eff}$' is done over various sampling periods to remove any variations.

### *F. Digital Block*

Digital block consists mainly of counters and registers. The counters increment on every ICOCLK edge. The values of the phase counters are registered on both the edges of the sampling clock. Counts of nominal and subranging modes of the ICO are computed separately. The possible one-clock delay of the SWITCH signal due to direction matching is also taken into account while computation. The subranged counts are divided by 'm$_{eff}$' and added to the nominal mode counter value to get the final ADC digital output.

## IV. Measurement Results

The prototype Subranging ICO-based ADC was fabricated in TSMC 180nm technology and occupies an active area of 0.06mm$^2$. The die photograph is presented in Fig 14. When powered by a 1.8V voltage supply, the system consumes a power of 10 $\mu$W at a sampling rate of 10kHz. A sinusoidal signal input current of 1$\mu$Ap-p at a frequency of 1kHz (nominal sensor frequency) with a DC bias current of $I_B$=2$\mu$A is fed into the system using current mirrors.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 21,2022 at 06:38:23 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.



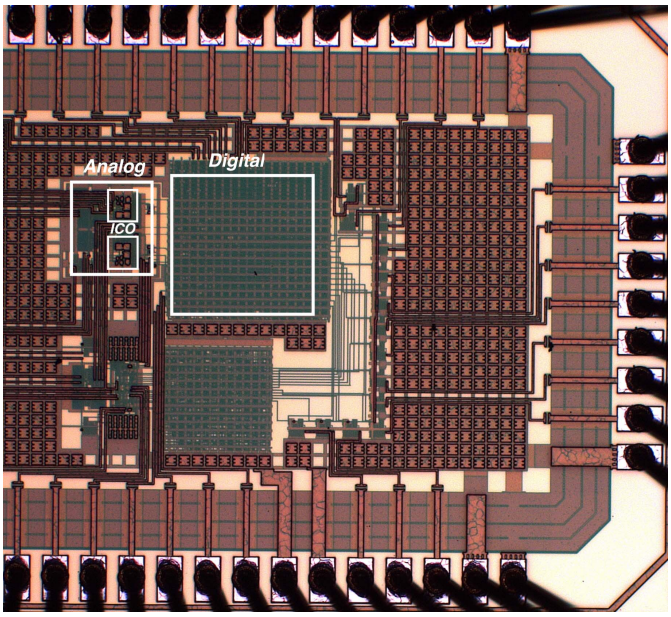



Fig. 14. Chip micrograph.

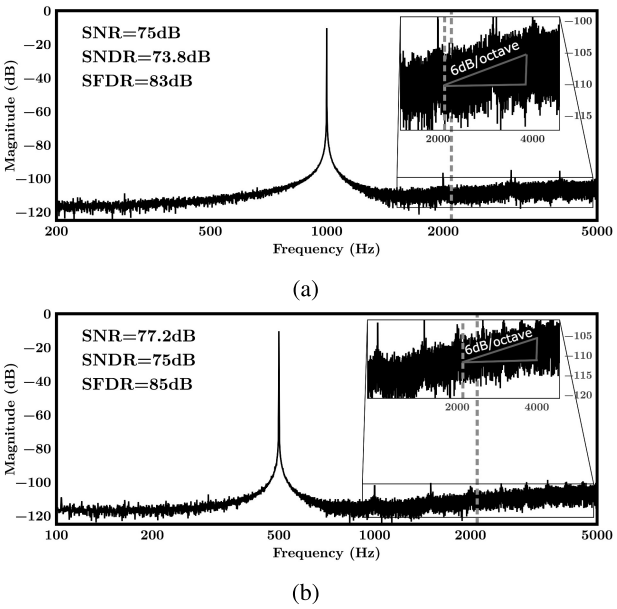



Fig. 15. Measured ADC spectrum for (a) $f_{in}$=1kHz (b) $f_{in}$=500Hz.

TABLE I
MEASURED SNDR VS 'M' FACTOR

| 'm' | 1 | 2 | 4 | 8 | 16 |
|---|---|---|---|---|---|
| $m_{eff}$ | 1 | 2 | 4 | 7 | 12 |
| SNDR (dB) | 56.5 | 62 | 66.4 | 70.5 | 73.8 |

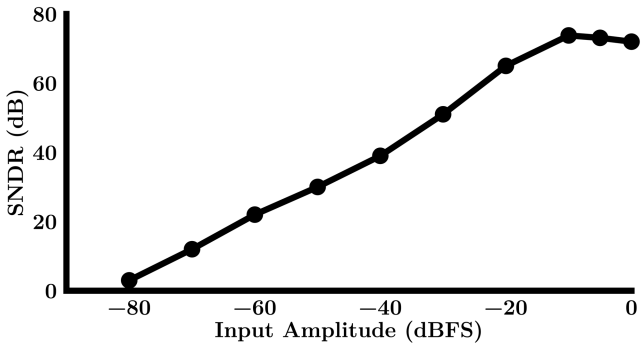



Fig. 16. SNDR vs Input signal amplitude.

The integration capacitor for nominal mode is set to 1.12pF and that for subranging mode is 70pF to obtain an ideal scaling factor of 16. Due to various reasons explained earlier, the accomplished effective scaling factor was found to be 12. The comparator reference voltages were initially set at $V_{TH_HIGH}$=1.2V and $V_{TH_LOW}$=0.6V for both the modes. The sampling clock has a duty cycle of 80% which makes the ADC run in subranging mode for 20% of the sampling period. The measured output spectrum of a 1kHz input signal is shown in Fig 15a where the measured signal to noise and distortion ratio (SNDR) is 73.8dB and spurious free dynamic range (SFDR) is 83 dB in 2kHz bandwidth. Similarly, for a sinusoid signal at a frequency of 500Hz shown in Fig 15b, the measured SNDR and SFDR are 75dB and 85dB respectively.

Even though the system is differential, harmonic spurs are visible due to various mismatches between the P-side and N-side. The quantization noise of the ADC is first order shaped since the integral of the signal is quantized and first difference of counters are taken without resetting them. The quantization error propagates from sample to sample.

The plot of measured SNDR at various signal amplitudes is presented in the Fig 16. A peak SNDR of 73.8 dB is obtained at $-8$ dBFS assuming a full-scale current of $3\mu$A. As the signal amplitude increases, the dynamic range of the counts increases but after a certain amplitude, the range of counts saturates dropping the SNDR due to the non-linearity in the current-to-frequency characteristics of ICO. However, the same architecture implemented in scaled CMOS technologies can achieve higher oscillator frequency there by achieving higher oversampling ratios and SNDR.

The measured frequency of the differential ICO is plotted in Fig 17 as a function of the input current. The corresponding percentage non-linearity is also plotted which is calculated by taking the difference between the obtained current-to-frequency and best-fit curve and normalizing it to the frequency at half full-scale current. The highest SNDR of various chips is studied for spatial variation across the die and plotted in Fig 18.

The 'm' factor decides enhancement in the performance of the ADC. Ideally, for every doubling of 'm', the dynamic range of the digital output should increase by two times, effectively increasing the SNDR by 6dB. The effective scaling factor differs from the set scaling factor due to the addition of parasitic capacitance which affects significantly at higher scaling factor values. In addition, due to the full range input current swing with differential operation, as 'm' increases, the effective scaling factor differs from the instantaneous scaling factor at the extremes of the input. The measured SNDR at various 'm' factors is summarized in Table I.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 21,2022 at 06:38:23 UTC from IEEE Xplore. Restrictions apply.

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



TABLE II
COMPARISON OF THE ADC PERFORMANCE

| **Performance** | **This Work** | JSSC'20 [7] | JSSC'15 [3] | Sensors'19 [29] | JSSC'17 [5] | VLSI'17 [30] | ASSCC'16 [31] |
|---|---|---|---|---|---|---|---|
| Technology | **180nm** | 350nm | 180nm | 180nm | 40nm | 65nm | 65nm |
| Architecture | **ICO** | VCO | VCO | NS SAR | DD VCO | SC | Hybrid CT/DT |
| Input Type | **Current** | Resistance | Current | Voltage | Voltage | Voltage | Voltage |
| Supply Voltage (V) | **1.8** | 1.75 | 5, 1.8 | 1 | 1.2 (A), 0.45 (D) | 1 | 1 |
| Signal Bandwidth (kHz) | **2** | 0.1 | 0.025 | 0.625 | 0.2 | 25 | 25 |
| Sampling Frequency (kS/s) | **10** | 10 | – | 10 | 0.4 | 13000 | 10000 |
| Area ($mm^2$) | **0.06** | 0.435 | 0.36 | 0.12 | 0.135 | 0.1 | 0.03 |
| SNDR (dB) | **73.8** | 79 | 74 | 65 | 74 | 88.2 | 84.2 |
| SFDR (dB) | **83** | 73.8 | 73 | 66.4 | 79 | 106 | 86.2 |
| Total Power Consumption ($\mu W$) | **10** | 0.009 | 340 | 0.09 | 7 | 73.6 | 68 |
| $FOM_W$ (fJ/Step) | **675** | 1180 | 37250000 | 50 | 4270 | 70 | 102.6 |
| $FOM_S$ (dB) | **156.8** | 159.2 | 111 | 163.4 | 157 | 158.5 | 170 |

$$FoM_W = \frac{Power}{2 \times BW \times 2^{ENOB}} \; ; FoM_S = SNDR + 10log_{10}\frac{BW}{Power}$$

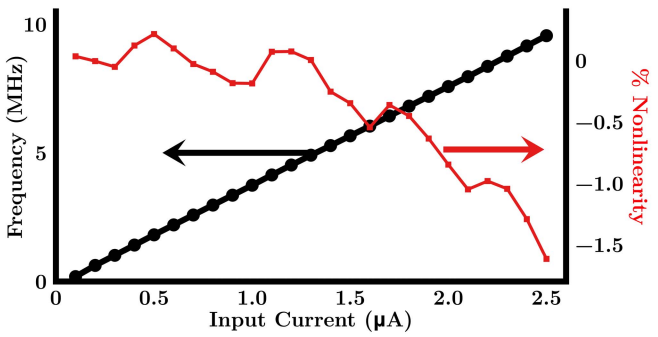



Fig. 17. Measured Non-linearity of the differential ICO.

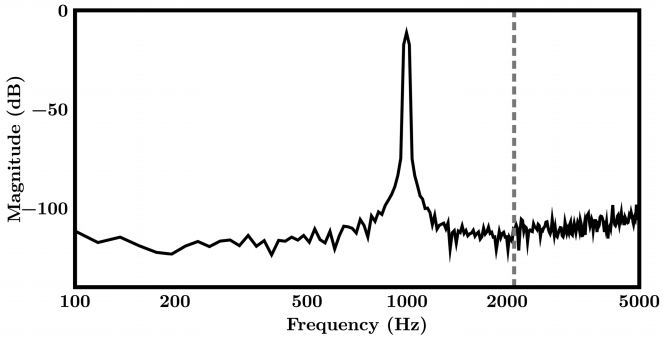



Fig. 19. FFT plot of the ADC output simulated in TSMC40nm for a 1$\mu$A sinusoid at 1kHz frequency.

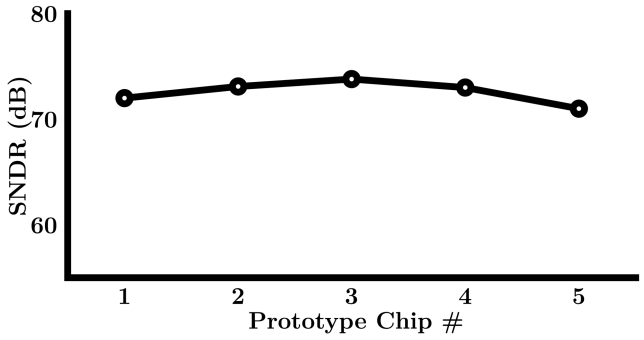



Fig. 18. Measured SNDR across 5 prototype chips.

The proposed prototype ADC achieves Walden FoM of 625 fJ/conversion cycle and a Schreier $FoM_{,DR}$ of 166 dB and $FoM_{,SNDR}$ of 156.8 dB. The Table II compares the complete performance of this prototype chip with the state-of-the-art ADCs for sensor instrumentation and low bandwidth applications. It can be seen that performance of the proposed ADC is either better or on par in terms of the figure of merit with prior art for various sensing applications. It has a significantly smaller area compared to the other ADCs. Although [30] and [31] achieves high resolution and figure of merit, the power consumption is relatively high due to high sampling frequency. These ADCs can be interfaced with sensor with transduced currents directly feeding the input stage thereby providing high sensitivity. In addition, these ADCs can easily handle large amount of DC common mode without causing overflows by simply extending the width of the digital counter. The system also scales well to shorter channel technologies due to the digital nature of its building blocks. For illustration, the same architecture was simulated using TSMC 40nm model files. Based on the simulation results, the expected power consumption is 5$\mu$W from a 1V supply accomplishing 88dB SNDR with $FoM_S$ of 179dB and $FoM_W$ of 40fJ/step. The spectral performance is presented in a basic FFT plot with the simulation results in Fig 19.

## V. CONCLUSION

A novel subranging relaxation ICO-based open-loop continuous-time Sigma-Delta ADC is presented in this article. This architecture leverages dynamic frequency and conversion gain scaling close to the sampling edge where quantization error is committed. A high power-efficiency is accomplished by running ICO at nominally low frequency most of the sampling period and accelerating only for a small fraction. A simple calibration cycle is run to extend the scaling factor to achieve high performance with negligible power overhead. Various techniques like level-detection and multi-stage noise shaping (MASH) can be implemented orthogonally to further reduce the quantization error enhancing the performance. This ADC architecture serves wide range of applications. The present implementation in 180nm is intended for low-power low and medium bandwidth sensor instrumentation. This

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 21,2022 at 06:38:23 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.



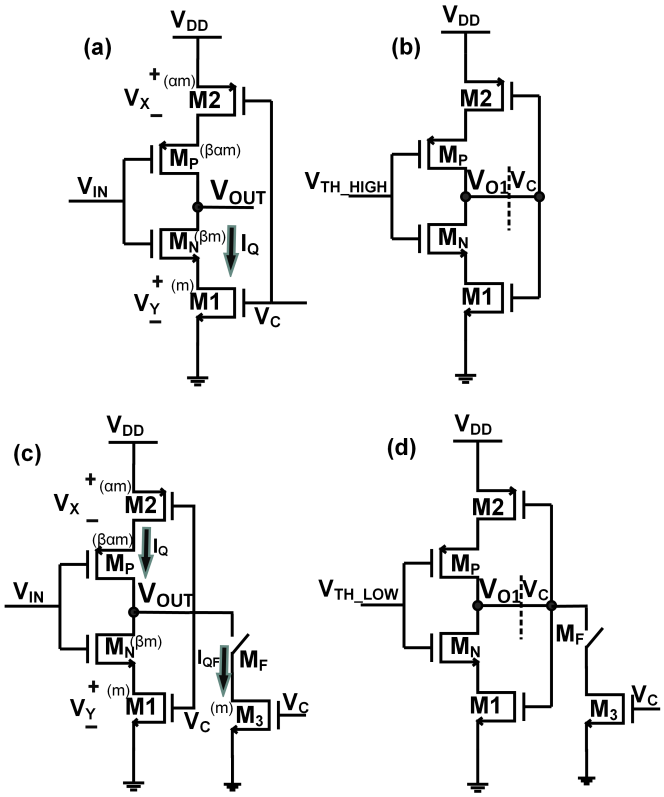



Fig. 20. Simplified Schmitt trigger half circuit to stabilize $V_{TH_HIGH}$ and its replica bias circuit.

architecture is readily scalable to shorter channel nodes, where it can be used for higher bandwidth applications like audio and video processing.

## APPENDIX

A simplified half circuit of the proposed Schmitt trigger and its replica bias to stabilize and set $V_{TH_HIGH}$ are presented in Fig 20a and Fig 20b respectively. The analysis is done by breaking the feedback loop in the replica, evaluating the voltages and equating them.

The quiescent current in the inverter is assumed to be $I_Q$. Consider the inverter in Fig 20a. Assuming $M_1$ and $M_2$ to be in triode and the small signal $v_c$ riding over the quiescent voltage $V_C$, we have

$$\begin{aligned} v_Y &\approx \frac{I_Q}{k_n\left(V_C+v_c-V_{TN}\right)} \\ &\Longrightarrow v_Y \approx \frac{I_Q\left(1+\frac{v_C}{V_C-V_{TN}}\right)}{k_n\left(V_C-V_{TN}\right)} \\ v_X &\approx \frac{I_Q}{k_p\left(V_{DD}-|V_{TP}|-V_C-v_c\right)} \\ &\Longrightarrow v_X \approx \frac{I_Q\left(1+\frac{v_c}{V_{DD}-|V_{TP}|-V_C}\right)}{k_p\left(V_{DD}-|V_{TP}|-V_C\right)} \end{aligned}$$

where $k_p = \mu_P C_{ox} W_P/L_P$ and $k_n = \mu_N C_{ox} W_N/L_N$, $V_{TP}$ and $V_{TN}$ are the threshold voltages of PMOS and NMOS respectively.

$$v_{LT} = V_{LT0} - \frac{\sqrt{k_p}v_X - \sqrt{k_n}v_Y}{\sqrt{k_p}+\sqrt{k_n}} \tag{3}$$

$$\begin{aligned} v_{LT} = V_{LT0} &+ \frac{\sqrt{k_p}v_c}{\left(V_C-V_{TN}\right)^2} + \frac{\sqrt{k_n}v_c}{\left(V_{DD}-|V_{TP}|-V_C\right)^2} \\ &+ \frac{\sqrt{k_p}}{\left(V_C-V_{TN}\right)} \\ &+ \frac{\sqrt{k_n}}{\left(V_{DD}-|V_{TP}|-V_C\right)} \frac{I_Q}{\sqrt{k_p k_n}\left(\sqrt{k_p}+\sqrt{k_n}\right)} \end{aligned}$$

where $V_{LT)}$ is the logical threshold of the basic inverter. By substituting the expressions in the parenthesis with appropriate symbols, the equation is simplified as

$$\begin{aligned} v_{LT} &= V_{LT0} + \Lambda_1\left(\Lambda_2 v_c + \Lambda_3\right) \\ &\Longrightarrow v_c = \frac{v_{LT} - V_{LT0} - \Lambda_1\Lambda_3}{\Lambda_1\Lambda_2} \end{aligned}$$

Now, consider the replica circuit shown in Fig 20b. The voltage $v_{o1}$ is given by:

$$v_{O1} = v_{LT} - A\left(v_{TH_HIGH} - v_{LT}\right)$$

The only factor that changes for the alternate arrangement shown in Fig 20b is the gain A, which increases to mA at the expense of additional delay (and a reduced phase margin) in the loop. Since, the feedback enforces $v_{o1} = v_C = V_C + v_c$, we have:

$$\begin{aligned} (A+1)v_{LT} &\approx A v_{TH_HIGH} + \frac{v_{LT} - V_{LT0} - \Lambda_1\Lambda_3}{\Lambda_1\Lambda_3} \\ &\Longrightarrow v_{LT} \approx \left[\frac{\Lambda_1\Lambda_2 A}{\Lambda_1\Lambda_2(A+1)-1}\right] v_{TH_HIGH} \\ &\quad + \frac{V_{LT0}+\Lambda_1\Lambda_3}{\Lambda_1\Lambda_2(A+1)-1} \end{aligned}$$

If $A \gg 1$,

$$v_{LT} \approx v_{TH_HIGH}$$

Similar simplified half circuit and its replica bias to stabilize and set $V_{TH_LOW}$ are presented in Fig 20c and Fig 20d. The analysis is done by breaking the feedback loop in the replica, evaluating the voltages and equating them.

$$\begin{aligned} v_Y &\approx \frac{I_Q}{k_n\left(V_C+v_c-V_{TN}\right)} \\ &\Longrightarrow v_Y \approx \frac{\left(I_Q - I_{QF}\right)\left(1+\frac{v_C}{V_C-V_{TN}}\right)}{k_n\left(V_C-V_{TN}\right)} \\ &\Longrightarrow v_Y \approx \frac{I_Q\left(1+\frac{v_C}{V_C-V_{TN}}\right) - k_n V_{OUT}\left(\frac{\left(V_C+v_C-V_{TN}\right)^2}{V_C-V_{TN}}\right)}{k_n\left(V_C-V_{TN}\right)} \\ v_X &\approx \frac{I_Q}{k_p\left(V_{DD}-|V_{TP}|-V_C-v_c\right)} \\ &\Longrightarrow v_X \approx \frac{I_Q\left(1+\frac{v_c}{V_{DD}-|V_{TP}|-V_C}\right)}{k_p\left(V_{DD}-|V_{TP}|-V_C\right)} \end{aligned}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 21,2022 at 06:38:23 UTC from IEEE Xplore. Restrictions apply.

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



$$v_{LT} = V_{LT0} + \frac{\sqrt{k_p} v_c}{(V_C - V_{TN})^2} + \frac{\sqrt{k_n} v_c}{(V_{DD} - |V_{TP}| - V_C)^2} + \frac{\sqrt{k_p}}{(V_C - V_{TN})} \frac{I_Q}{\sqrt{k_p k_n}\left(\sqrt{k_p} + \sqrt{k_n}\right)} + \frac{\sqrt{k_n}}{(V_{DD} - |V_{TP}| - V_C)} \frac{I_Q - I_{QF}}{\sqrt{k_p k_n}\left(\sqrt{k_p} + \sqrt{k_n}\right)}$$

Using Equation (3), $v_{LT}$ as shown in the equation at the top of the page.

By substituting the expressions in the parenthesis with appropriate symbols, the equation is simplified as

$$\begin{aligned} v_{LT} &= V_{LT0} + \Lambda_4 (\Lambda_5 v_c + \Lambda_6) \\ \implies v_c &= \frac{v_{LT} - V_{LT0} - \Lambda_4 \Lambda_6}{\Lambda_4 \Lambda_5} \end{aligned}$$

Now, consider the replica circuit shown in Fig 20d. The voltage $v_{o1}$ is given by:

$$v_{O1} = v_{LT} - A\left(v_{TH_LOW} - v_{LT}\right)$$

Following similar analysis of previous circuit for $V_{TH_HIGH}$, we can write

$$\begin{aligned} (A+1)v_{LT} &\approx A v_{TH_LOW} + \frac{v_{LT} - V_{LT0} - \Lambda_4 \Lambda_6}{\Lambda_4 \Lambda_6} \\ \implies v_{LT} &\approx \left[\frac{\Lambda_4 \Lambda_5 A}{\Lambda_4 \Lambda_5 (A+1) - 1}\right] v_{TH_LOW} \\ &\quad + \frac{V_{LT0} + \Lambda_4 \Lambda_6}{\Lambda_4 \Lambda_5 (A+1) - 1} \end{aligned}$$

If $A \gg 1$,

$$v_{LT} \approx v_{TH_LOW}$$

## ACKNOWLEDGMENT

The authors would also like to thank Srinagesh Satyanarayana, B. M. Keshavnandan, and Venkata Suresh Rayudu for their valuable inputs, feedback, and discussions.

## REFERENCES

[1] M. Z. Straayer and M. H. Perrott, "A 12-bit, 10-MHz bandwidth, continuous-time $\Sigma\Delta$ ADC with a 5-bit, 950-MS/s VCO-based quantizer," *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 805–814, Apr. 2008.

[2] M. Park and M. H. Perrott, "A 78 Db SNDR 87 mW 20 MHz bandwidth continuous-time $\Delta\Sigma$ ADC with VCO-based integrator and quantizer implemented in 0.13 $\mu$m CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 12, pp. 3344–3358, Dec. 2009.

[3] P. Prabha *et al.*, "A highly digital VCO-based ADC architecture for current sensing applications," *IEEE J. Solid-State Circuits*, vol. 50, no. 8, pp. 1785–1795, Aug. 2015.

[4] K. Ragab and N. Sun, "A 12-B ENOB 2.5-MHz BW VCO-based 0–1 mash ADC with direct digital background calibration," *IEEE J. Solid-State Circuits*, vol. 52, no. 2, pp. 433–447, Feb. 2017.

[5] W. Jiang, V. Hokhikyan, H. Chandrakumar, V. Karkare, and D. Markovié, "A ±50-mV linear-input-range VCO-based neural-recording front-end with digital nonlinearity correction," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 173–184, Jan. 2017.

[6] G. Taylor and I. Galton, "A mostly-digital variable-rate continuous-time delta-sigma modulator ADC," *IEEE J. Solid-State Circuits*, vol. 45, no. 12, pp. 2634–2646, Dec. 2010.

[7] B. Chatterjee *et al.*, "A wearable real-time CMOS dosimeter with integrated zero-bias floating gate sensor and an 861-nW 18-bit energy-resolution scalable time-based radiation to digital converter," *IEEE J. Solid-State Circuits*, vol. 55, no. 3, pp. 650–665, Mar. 2020.

[8] S. Szczesny, M. Kropidłowski, and M. Naumowicz, "0.50-V ultra-low-power $\Sigma\Delta$ modulator for sub-NA signal sensing in amperometry," *IEEE Sensors J.*, vol. 20, no. 11, pp. 5733–5740, Jun. 2020.

[9] S. Rao, K. Reddy, B. Young, and P. K. Hanumolu, "A determ inistic digital background calibration technique for VCO-based ADCs," *IEEE J. Solid-State Circuits*, vol. 49, no. 4, pp. 950–960, Apr. 2014.

[10] X. Xing, P. Gao, and G. Gielen, "A 40 MHz-BW two-step open-loop VCO-based ADC with 42 fJ/step FoM in 40 nm CMOS," in *Proc. ESSCIRC*, Sep. 2013, pp. 327–330.

[11] V. Nguyen, F. Schembari, and R. B. Staszewski, "A 0.2-V 30-MS/s 11b-ENOB open-loop VCO-based ADC in 28-nm CMOS," *IEEE Solid-State Circuits Lett.*, vol. 1, no. 9, pp. 190–193, Sep. 2018.

[12] S. Vasishta, K. R. Raghunandan, A. Dodabalapur, and T. R. Viswanathan, "A novel low-power single-step subranging open-loop sigma-delta ADC," in *Proc. IEEE Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2021, pp. 116–120.

[13] P. F. J. Geraedts, E. van Tuijl, E. A. M. Klumperink, G. J. M. Wienk, and B. Nauta, "A 90 W 12 MHz relaxation oscillator with a −162d B FOM," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 348–618.

[14] J. Lee, A. K. George, and M. Je, "An ultra-low-noise swing-boosted differential relaxation oscillator in 0.18-$\mu$m CMOS," *IEEE J. Solid-State Circuits*, vol. 55, no. 9, pp. 2489–2497, Sep. 2020.

[15] S.-Y. Lu and Y.-T. Liao, "A low-power, differential relaxation oscillator with the self-threshold-tracking and swing-boosting techniques in 0.18-$\mu$m CMOS," *IEEE J. Solid-State Circuits*, vol. 54, no. 2, pp. 392–402, Feb. 2019.

[16] S. L. J. Gierkink and E. van Tuij, "A coupled sawtooth oscillator combining low jitter with high control linearity," *IEEE J. Solid-State Circuits*, vol. 37, no. 6, pp. 702–710, Jun. 2002.

[17] J. Candy, "A use of double integration in sigma delta modulation," *IEEE Trans. Commun.*, vol. TC-33, no. 3, pp. 249–258, Mar. 1985.

[18] M. Keller, A. Buhmann, F. Gerfers, M. Ortmanns, and Y. Manoli, "On the implicit anti-aliasing feature of continuous-time cascaded sigma–delta modulators," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 54, no. 12, pp. 2639–2645, Dec. 2007.

[19] N. Sayiner, H. V. Sorensen, and T. R. Viswanathan, "A level-crossing sampling scheme for A/D conversion," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 43, no. 4, pp. 335–339, Apr. 1996.

[20] B. Schell and Y. Tsividis, "A continuous-time ADC/DSP/DAC system with no clock and with activity-dependent power dissipation," *IEEE J. Solid-State Circuits*, vol. 43, no. 11, pp. 2472–2481, Nov. 2008.

[21] K. R. Raghunandan, T. L. Viswanathan, and T. R. Viswanathan, "Linear current-controlled oscillator for analog to digital conversion," in *Proc. Custom Integr. Circuits Conf. (CICC)*, Sep. 2014, pp. 1–4.

[22] R. Yadav, K. R. Raghunandan, A. Dodabalapur, T. L. Viswanathan, and T. R. Viswanathan, "Operational current to frequency converter," in *Proc. IEEE 56th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2013, pp. 900–903.

[23] M. Vesterbacka and V. Unnikrishnan, "Ring counters as phase accumulator in VCO-based ADCs," in *Proc. 25th IEEE Int. Conf. Electron., Circuits Syst. (ICECS)*, Bordeaux, France, Dec. 2018, pp. 109–112. [Online]. Available: https://ieeexplore.ieee.org/document/8617848/

[24] C. C. Enz and G. C. Temes, "Circuit techniques for reducing the effects of OP-AMP imperfections: Autozeroing, correlated double sampling, and chopper stabilization," *Proc. IEEE*, vol. 84, no. 11, pp. 1584–1614, Nov. 1996.

[25] A. A. Abidi and R. G. Meyer, "Noise in relaxation oscillators," *IEEE J. Solid-State Circuits*, vol. JSSC-18, no. 6, pp. 794–802, Dec. 1983.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 21,2022 at 06:38:23 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.



[26] S. Vasishta, K. R. Raghunandan, A. Dodabalapur, and T. R. Viswanathan, "Level-crossing detection based low-power sigma-delta ADC for sensor applications," in *Proc. IEEE 63rd Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2020, pp. 663–666.

[27] S. Vasishta, K. R. Raghunandan, and A. Dodabalapur, "A sawtooth relaxation ICO based 1–1 MASH ADC," in *Proc. IEEE 63rd Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2020, pp. 349–352.

[28] K. R. Raghunandan, N. Sun, and T. R. Viswanathan, "Analog signal processing in deep submicron CMOS technologies using inverters," in *Proc. IEEE 57th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2014, pp. 394–397.

[29] J. Hu, D. Li, M. Liu, and Z. Zhu, "A 10-kS/s 625-Hz-bandwidth 65-db SNDR second-order noise-shaping SAR ADC for biomedical sensor applications," *IEEE Sensors J.*, vol. 20, no. 23, pp. 13881–13891, Dec. 2020.

[30] A. Hussain, S.-W. Sin, C.-H. Chan, S.-P. U. Ben, F. Maloberti, and R. P. Martins, "Active-passive $\delta\sigma$ modulator for high-resolution and low-power applications," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 25, no. 1, pp. 364–374, Jan. 2017.

[31] Y. Ren, S.-W. Sin, C.-S. Lam, M.-C. Wong, U. Seng-Pan, and R. P. Martins, "A high DR multi-channel stage-shared hybrid front-end for integrated power electronics controller," in *Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, Nov. 2016, pp. 57–60.

**Sudhanva Vasishta** received the B.E. degree in electronics and telecommunication from the Siddaganga Institute of Technology, Tumkur, India, in 2013, and the M.S. degree in micro devices track from the University of Minnesota, Twin Cities, in 2016. He is currently pursuing the Ph.D. degree in the integrated circuits and systems track with the Department of Electrical and Computer Engineering, The University of Texas at Austin. He was a Research Assistant at the Center for Nano Science and Engineering (CeNSE), Indian Institute of Science, from 2013 to 2014. He has interned at Intel and Cirrus Logic as an Analog IC Designer during the summers of his Ph.D. His research interests include analog/mixed-signal design, low-power time-based ADCs and analog front-end design using novel 2D materials.

**K. R. Raghunandan** received the B.E. degree in electronics and communication engineering from the Delhi College of Engineering, University of Delhi, New Delhi, India, in 2009, and the M.S. and Ph.D. degrees in the integrated circuits and systems track from the Department of Electrical and Computer Engineering, The University of Texas at Austin. He is currently a Design Engineering Manager with the MSS Division, Skyworks Inc, Austin. His research interests include analog and mixed-signal design in short channel technologies with a target application in high speed data communication systems.

**Ananth Dodabalapur** (Fellow, IEEE) received the Ph.D. degree in electrical engineering from The University of Texas at Austin in 1990. From 1990 to 2001, he was at Bell Laboratories, Murray Hill, NJ. He has published more than 250 articles in refereed journals which have resulted in an H Index of 93 (Google Scholar), and has 27 issued U.S. patents, which have been cited nearly 2000 times. Since September 2001, he has been with The University of Texas at Austin and holds the Motorola Regents Chair in Engineering. His current research interests include organic and inorganic thin-film transistors and optoelectronic devices. He is a fellow of NAI. In 2003, he co-founded OrganicID, a company that is investigating using printable polymer electronics to fabricate low-cost RFID tags for the 13.56 MHz frequency. He was the founding Editor-in-Chief of *Flexible and Printed Electronics*.

**T. R. Viswanathan** (Life Fellow, IEEE) received the B.Sc. degree in physics from the University of Madras, India, in 1956, the D.I.I.Sc. degree in electrical communication engineering from the Indian Institute of Science in 1959, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Saskatchewan, Canada, in 1960 and 1964, respectively. He was a Professor of electrical engineering at the IIT Kanpur, India; the University of Michigan, Dearborn; Carnegie Mellon University, Pittsburgh, PA; and the University of Waterloo, ON, Canada. From 1985 to 1995, he was a Technical Manager at AT&T Bell Laboratories and from 1995 to 2001, the Director of Research and Development at Texas Instruments, Inc., Dallas. He was an Adjunct Professor at the University of Pennsylvania, a Visiting Professor at the University of California at Davis, and a Research Professor at The University of Texas at Dallas. Currently, he is a Research Emiritus Professor with the University of Texas at Austin. He was a member of the Administrative Committee of the IEEE Solid-State Circuits Society, ISSCC Executive Committee, and the Board of Governors of the IEEE Circuits and Systems Society. He was a recipient of the CAS Darlington Award and the Kilby Award. He was an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 21,2022 at 06:38:23 UTC from IEEE Xplore. Restrictions apply.

TechRxiv
Powered by IEEE

# Achieving Low-Noise, High-Bandwidth, Nanoscale-Resolution Position Sensing with GMR Sensors

This paper was downloaded from TechRxiv (https://www.techrxiv.org).

LICENSE

CC BY-NC-SA 4.0

SUBMISSION DATE / POSTED DATE

08-09-2022 / 12-09-2022

CITATION

Joseph, Tim John; Kartik, V (2022): Achieving Low-Noise, High-Bandwidth, Nanoscale-Resolution Position Sensing with GMR Sensors. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.21062659.v1

DOI

10.36227/techrxiv.21062659.v1

# Achieving Low-Noise, High-Bandwidth, Nanoscale-Resolution Position Sensing with GMR Sensors

Tim John Joseph, *Student Member, IEEE,* and V. Kartik

***Abstract*—Nanometer-resolution position sensing is a critical requirement in several precision applications such as Atomic Force Microscopy (AFM), especially over large bandwidths. Giant magnetoresistance-based (GMR) sensors have shown great promise for this; however, they exhibit significantly higher 1/f noise even over large bandwidths, which limits the achievable SNR. Therefore, the sensing system needs to be designed at the system-level in order to minimize excess intrinsic noise generation and transmission, as well as external noise pickup. This article, both analytically and experimentally, investigates the noise characteristics of the GMR sensor and its readout circuit components, and their effect on SNR. A general process for designing a low-noise high-bandwidth readout circuit, applicable not just to GMR sensors but also to other high-1/f-noise sensors, is presented. A resolution of 2.5 nm over a bandwidth of 100 kHz is demonstrated on an AFM nanopositioner, a ten-fold improvement over previously reported performance and comparable to other state-of-the-art, but more complex, position sensing schemes. The readout scheme is simple to implement using COTS components and easily extendable to higher bandwidths, unlike other schemes, such as modulation/demodulation, where the requirement for increasingly higher carrier frequencies renders further improvements in bandwidth impractical.**

***Index Terms*—Nanopositioning, giant magnetoresistance, position measurement, circuit noise, 1/f noise.**

## I. Introduction

Position sensing with nanometer resolution is a critical requirement in several high-precison applications such as scanning probe microscopy (SPM) [1], nanofabrication [2], semiconductor metrology [3], and data storage [4]. Advancements in these fields have given rise to the need for *dynamic* nanometer sensing over high bandwidths [5], with the emergence of several promising applications, such as video-rate imaging of fast biological processes [6], high-speed surface inspection [7], and the potential for several others [8].

Achieving nanometer resolution over a wide bandwidth, however, is challenging due to the increase in sensor noise with bandwidth, which is the primary factor limiting the achievable resolution. Although several sensing schemes achieving nanometer resolution have been reported – such as, eddy current-based [9], capacitive [9], interferometric [10], and inductive [11] – the bandwidth is limited to 10-20 kHz. Capacitive sensors have exhibited a resolution of 1 nm over a bandwidth of 100 kHz [12] and have been employed in several high-speed nanopositioners [13], [14]; however, the readout scheme is complex [15], rendering further improvements in bandwidth difficult and impractical [5]. Interferometric sensors achieve both sub-nanometer resolution and bandwidths of several MHz. However, their large form factor and high cost render them prohibitive for several applications, especially when multiple sensors need to be deployed in close proximity to each other. In contrast, giant magnetoresistance (GMR) sensors have been shown to achieve sub-nanometer resolution over a 1 MHz bandwidth, albeit over a short range [16]. With the added advantages of a small form factor and simple readout scheme – which enables collocated dynamic behaviour – these prove to be a remarkable alternative to conventional sensors; their suitability for high-bandwidth multi-axis closed-loop operation has been demonstrated for high-speed atomic force microscopy [17].

While the GMR sensor holds great promise, the main impediment to realizing nanoscale resolution is the inherently higher proportion of flicker or 1/f noise [18], which, unlike thermal noise, is sensitive to the sensor drive current [19]. Therefore incremental enhancements in the drive current produce diminishing improvements in SNR. In fact, at higher bridge imbalance levels, the SNR is found to decrease with an increase in sensor drive current. This 1/f characteristic, otherwise predominantly a low-frequency phenomenon, extends over a bandwidth exceeding 10 kHz and therefore cannot be ignored, as is typically done in high-bandwidth applications. This leads to unique and more stringent requirements placed upon the readout circuitry, including judicious *system-level* component selection. The design of the sensor drive itself is critical, especially due to the sensor's bridge configuration. Previously reported noise estimates for GMR position readout circuits have not factored in the 1/f contribution [20].

In this article, the development of a low-noise high-bandwidth readout circuit for GMR sensors is presented. Noise interactions between the GMR sensor and the other readout circuit components are investigated. Methods to minimize intrinsic noise generation and transmission, and extrinsic noise pickup are discussed. An integrated design process is presented that covers not only the circuit, but also the selection of components, and the design of the precision sensor drive. This process is sufficiently general and can also be applied to other low-noise readout applications, particularly those employing sensors in bridge circuits, such as strain gauges

Tim John Joseph and V. Kartik are with the Department of Mechanical Engineering, Indian Institute of Technology Bombay, Mumbai, 400076 India (e-mail: timjohn@iitb.ac.in; vkartik@iitb.ac.in).

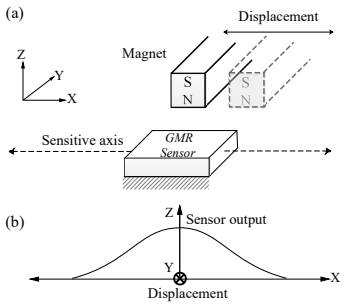



Fig. 1. (a) The GMR sensor-based position sensing method, and (b) the sensor output signal vs. displacement.

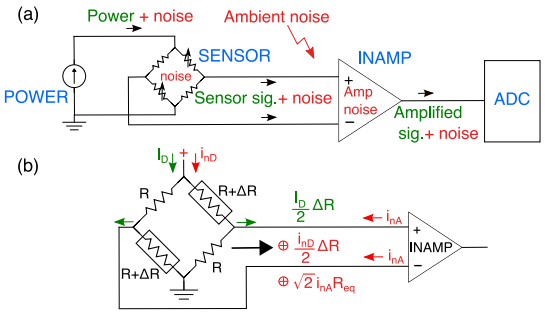



Fig. 2. (a) The sensing system. Ideal noise-free signal is shown in green, and the noise sources and components in red. (b) GMR sensor input and output. ⊕ indicates root-sum-of-squares summation.

and NTC sensors. The GMR sensing system is deployed in a nanopositioner for an atomic force microscope (AFM), and experiments to validate and benchmark its performance are presented. The remainder of the article is organized as follows: in Section II, the GMR sensing method is detailed; Section III presents component-level noise analysis and design details of the circuit; and experiments to validate the design and demonstrate its performance are presented in Section IV.

## II. GMR Sensing Method

A GMR sensor can measure very small changes in magnetic field strength due to its high sensitivity. As compared to other conventional magnetic sensors, its operating principle is the phenomenon of giant magnetoresistance, where the electrical resistance of the sensor's active material changes under the action of an externally-applied magnetic field. The GMR sensor-based position sensing method was first presented in Sahoo *et al.* [16] and Kartik *et al.* [17], and is depicted schematically in Fig. 1. The key principle is to convert a relative displacement into a corresponding change in magnetic field gradient, which is then detected by the GMR sensing elements. This is achieved by attaching a permanent magnet to the moving body while positioning the GMR sensor underneath (or directly above) it. As the body moves over the sensor, the magnetic field 'seen' by it also changes; the corresponding change in sensor voltage is then read out as a displacement. The conversion from displacement to voltage is given as

$$\Delta V = \frac{dV}{dR}\frac{dR}{dB}\frac{dB}{dx}\Delta x + \Delta N \qquad (1)$$

where $\frac{dB}{dx}$ is the change in the incident magnetic field with displacement, $\frac{dR}{dB}$ is the corresponding change in the resistance of the sensor, $\frac{dV}{dR}$ is the change in the sensor's output voltage with a change in its resistance, $\Delta x$ is the displacement to be measured, and $\Delta N$ is unavoidable measurement noise. While under ideal conditions, any infinitesimally small $\Delta x$ should be detectable, $\Delta N$ limits the achievable resolution, and the ratio of $\Delta V$ to $\Delta N$ (SNR) becomes the resolution-limiting factor. Therefore, the design problem is to maximize $\Delta V/\Delta N$, and the challenge in our application is that, due to the inherently small displacements, the corresponding $\Delta V$ is also very small. Thus, the demands placed on the SNR are also extreme.

Increasing the SNR entails either increasing the signal, or decreasing the noise, or both. Increasing the signal implies increasing one or more of the three ratios in (1). $\frac{dR}{dB}$, which is the GMR effect, is an inherent characteristic of the particular sensor, and selecting a more sensitive sensor is the only solution to improving this ratio. Since that possibility is limited by advances in materials, $\frac{dB}{dx}$ is the primary design parameter that can be improved upon. It depends primarily upon geometric factors, such as the shape, size, orientation and region of operation of the permanent magnet used, and the separation distance between the sensor and the magnet. The choice of the magnet's material also affects this ratio. A study by these authors on these geometrical aspects was presented in [21], while [20] provided a study on the magnetics involved. $\frac{dV}{dR}$ depends upon the excitation applied to the sensor, and increasing the excitation amplitude, as much as allowable, would appear to be the apparent solution. However, due to the particular noise characteristics of the GMR sensor, the noise is also seen to increase, albeit at a lower rate, as excitation increases. This article primarily investigates methods to characterize and reduce the total noise in the system. Of particular interest are the current-dependent 1/f noise characteristics of the GMR sensor, which, we show, critically limit the achievable resolution. Judicious design and component selection are critical in realizing the required performance.

## III. Component-level Intrinsic Noise Analysis and Design

We begin by first estimating the flow and build-up of noise within the system, comprising the GMR sensor and the readout circuitry. The primary requirements for the readout circuit are a stable power source to drive the sensor, and an amplifier for the sensor output signal prior to A/D conversion. Fig. 2(a) schematically depicts the system's components along with the various noise sources, forming a signal chain. As depicted in the figure, there are two broad categories of noise: (a) the ambient noise that couples into the system (*extrinsic* noise), and (b) the noise that arises within each component (*intrinsic* noise). These noise components propagate through the chain and are transformed at each step. The objective, then, is to (1) limit the coupling of extrinsic noise, and (2) reduce intrinsic noise generation while attenuating noise propagation and limiting noise amplification. This section focuses on the intrinsic noise, while methods to limit extrinsic noise coupling are discussed in Section IV.

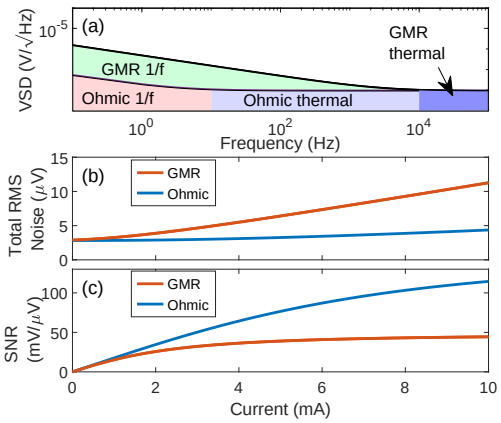



Fig. 3. Simulated noise for the GMR sensor and Ohmic device. (a) The noise voltage spectral density, (b) RMS noise vs. current, and (c) SNR vs. current.

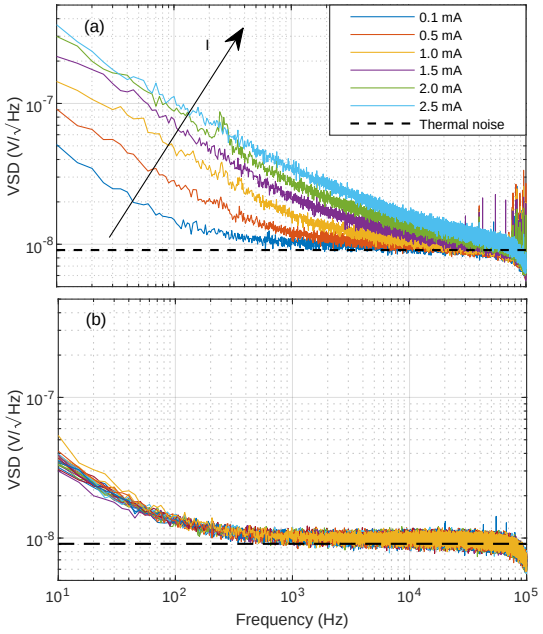



Fig. 4. Measured noise voltage spectral density for different drive currents for (a) the GMR sensor and (b) the Ohmic device. The thermal noise floor is indicated by the black dashed type line.

To be able to compare the noise contributions from each component, the noise levels are referred to a single point, namely the output of the sensor, or, equivalently, the input to the amplifier. This also allows a direct comparison of the total noise with the sensor's output signal. We now examine the noise characteristics of each component, beginning with the sensor itself.

### A. GMR Sensor Noise

We consider a commercially-available GMR sensor (Model AA002-02, NVE Corporation, Eden Prairie, MN, USA); this sensor is configured as a Wheatstone bridge (with two active elements, i.e., as a half-bridge). It, therefore, can easily interface with any standard instrumentation amplifier ('inamp'). The sensor's output voltage is directly proportional to the excitation amplitude due to the bridge configuration. Consequently, any noise from the power source also appears at the sensor's output. The inamp too interacts with the sensor and gives rise to additional noise at the sensor output. Fig. 2(b) schematically depicts the ideal input to and output from the GMR sensor and the additional noise components. Each arm of the sensor has one active and one passive element with a total resistance of $2R + \Delta R$ per arm, where $R$ is the nominal element resistance, and $\Delta R$ is the change in resistance for the active elements. Therefore, the drive current $I_D$ splits equally between the two arms, and the same is the case with the drive current noise $i_{nD}$. As the noise currents flowing in each arm have the same source, the two noise currents are correlated. Thus, the resulting noise voltage components – $\frac{1}{2} i_{nD} R$ and $\frac{1}{2} i_{nD}(R + \Delta R)$ – at the inamp's input subtract algebraically to result in total noise of $\frac{1}{2} i_{nD} \Delta R$. The current noise $i_{nA}$ from the inamp flows through each sensor arm separately. In this case, however, the noise currents in the two arms are independent of each other and are therefore uncorrelated; the two resulting noise voltage components at the inamp input will add as their root sum-of-squares. The total output of the sensor, excluding its intrinsic noise, will then be

$$V_{op} = \sqrt{\left(\frac{I_D}{2}\Delta R\right)^2 + \left(\frac{1}{2} i_{nD} \Delta R\right)^2 + \left(\sqrt{2} i_{nA} R_{eq}\right)^2} \quad (2)$$

where $R_{eq}$ is the Thevenin equivalent resistance at each inamp input.

The intrinsic noise of the GMR sensor also contributes considerably to the total noise. Due to its inverse frequency characteristic, the 1/f noise is predominantly a low-frequency phenomenon, with a minor contribution to the total noise, and, so, is usually neglected, especially in high-bandwidth applications. GMR materials, however, produce a significant amount of 1/f noise [18], which is typically an order of magnitude higher than the noise ordinarily observed for thin-film resistors [22]. As a result, the 1/f noise extends over a much wider frequency range and, so, its contribution cannot be neglected and needs to be included. The key issue with 1/f noise, unlike thermal noise, is its dependence upon drive current. The power spectral density is proportional to the square of the current [19]; equivalently, the voltage spectral density is proportional to the current.

To determine the implications of this 1/f characteristic, we simulate the noise in the GMR sensor. The 1/f noise spectral density is obtained from the noise plot given in the manufacturer's datasheet [22]. The derived expression for the 1/f noise spectral density $e_{nf}$ ($\mathrm{nV}/\sqrt{\mathrm{Hz}}$) at a frequency $f$ as a function of $I_D$ is

$$e_{nf} = (179.04 I_D + 7.6)\frac{1}{\sqrt{f}} \quad (3)$$

The thermal noise at room temperature and resistance $R$ is given as $4.07\sqrt{\mathrm{R}}$ $\mathrm{nV}/\sqrt{\mathrm{Hz}}$ which, for the AA002-02 with a nominal $R = 5$ $k\Omega$, is 9.1 $\mathrm{nV}/\sqrt{\mathrm{Hz}}$. The total RMS noise $V_{nS}$ is the root sum-of-squares of the two noise components and is given as a function of $I_D$ as

$$V_{nS} = \sqrt{(179.04 I_D + 7.6)^2 \ln\left(\frac{f_{BW}}{0.1}\right) + 4.07^2 R f_{BW}} \quad (4)$$

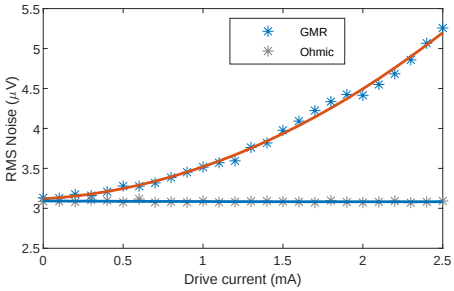



Fig. 5. The measured RMS noise vs. current characteristics of the GMR sensor and the Ohmic device. Trend lines are shown in solid line-type.

where $f_{BW}$ is the bandwidth. For comparison, we also simulate the noise in an equivalent resistor bridge, i.e., a bridge comprising resistors with the same nominal resistance as that of the GMR sensor (henceforth referred to as the 'Ohmic device'). The resistor noise characteristics are taken from previously published data [23]. Fig. 3(a) shows the noise spectral density of the GMR sensor and the Ohmic device together, with the 1/f and thermal noise-dominated regions highlighted. The 1/f region is much more pronounced in the case of the GMR sensor (marked in magenta). Fig. 3(b) shows the total RMS noise vs. $I_D$. The noise grows more rapidly in the case of the GMR sensor due to its much higher current-dependent 1/f noise. The effect of this on the achievable SNR as a function of $I_D$ is shown in Fig. 3(c). For the GMR sensor, the SNR quickly reaches a ceiling once the noise increases at the same rate as the signal. This is an inherent limitation of the 1/f noise-dominated GMR sensor.

To verify the above, we measure the noise from the GMR sensor and the Ohmic device for different drive currents. Thin film 0.1 % tolerance resistors were used to construct the Ohmic device. Figure 4(a) and (b) depict the obtained noise spectral densities. As expected, the 1/f characteristic is much more prominent in the case of the GMR sensor, extending even beyond 10 kHz at higher currents. The current dependence of the 1/f characteristic is also clearly seen, the effect of which can be seen in Fig. 5, which depicts the measured RMS noise for the two devices for different currents. The cause of the increase in RMS noise with current can be understood from the rising spectral density of Fig. 4(a) and the resulting increase in area under the curve; this area equals the total RMS noise. This increase in noise with drive current limits the incremental improvement in SNR typically seen when the sensor drive current is increased.

### B. Sensor Drive Design

As seen in Section III-A, a stable low-noise drive is required, which can be of either constant voltage or constant current type. For a constant voltage source, the relationship between the (half-) bridge's output and $\Delta R$ is non-linear, while for a current source, the relationship is linear [24]. Further, a circuit with a current source is immune to voltage drops caused by the wiring resistance of the connecting cables. In the case of a voltage source, however, these voltage drops can be significant, leading to a reduced drive voltage at the sensor input. Methods such as Kelvin four-wire sensing can be employed to ensure that the desired drive voltage is available at the sensor; this, however, increases the complexity of the system. Thus, a current source enables accurate driving of the sensor, with the added advantages of simpler cabling and a smaller footprint at the point of sensing. Current source signals are also more immune to inductively-coupled noise [25]. The temperature stability of the sensor output itself greatly improves when driven by a current source as the output depends only upon $\Delta R$, while for a voltage source, the sensor output also depends upon $R$, which changes with temperature. For all of the above reasons, we select a constant current drive for our investigation. The noise at the GMR sensor's output due to the noise from the current source $i_{nD}$ is $e_{nD} = i_{nD}\Delta R/2$.

TABLE I
KEY SPECIFICATIONS: REF102 [26], LT3092 [27], AND REF200 [28].

| | | REF102 [26] | LT3092 [27] | REF200 [28] |
|---|---|---|---|---|
| Given | Current range | 0 - 10 mA | 0.5 - 200 mA | 100 - 400 μA |
| | Initial accuracy (%) | ±0.025 | 1 | ±0.25 |
| | Drift (ppm/°C) | 2.5 | Not Available | 25 |
| | Noise 0.1 Hz - 10 Hz | 5 $\mu V_{pp}$ | Not Available | 1 $nA_{pp}$ |
| | Noise Broadband | 200 $\mu V_{pp}$ 0 Hz - 1 MHz | 0.7 nA RMS 10 Hz - 100 kHz | 20 pA/$\sqrt{\text{Hz}}$ at 10 kHz |
| Calculated | Noise density (Normalized RMS) | 20 pA/mA/$\sqrt{\text{Hz}}$ | 2.21 pA/$\sqrt{\text{Hz}}$ | – |
| | Total noise ($e_{nD}$) (pA/$\sqrt{\text{Hz}}$, I (mA) | $\sqrt{9.18\text{I}^2 + 1.66\text{I}}$ | $\sqrt{82.81\text{I} + 5.71}$ | 49I |

A constant current source can be implemented using several methods. Monolithic devices, such as the REF200 (Texas Instruments) [28] or the LT3092 (Analog Devices) [27] enable simple implementation of a constant current source, while the REF102 (Texas Instruments) [26] precision voltage reference-based circuit provides superior performance, albeit requiring additional components [29]. Table I summarizes the key specifications of the three implementations. The REF200 has the advantage of being a fully-integrated device with no external components required. However, it can only supply a limited number of fixed current values, the highest being 400 μA. The REF102-based circuit requires a single external resistor $R_{SET}$ to set the current and an additional precision opamp to ensure accuracy; any current up to 10 mA can be generated. The main advantage of this implementation is the superior performance as compared to the REF200. As seen in Table I, the drift and initial accuracy are an order of magnitude better, while the noise is significantly lower. The LT3092 requires a pair of external resistors to set the current, and any current between 0.5 mA and 200 mA is possible. The initial accuracy, however, is lower. Data on the drift are not available from the manufacturer. Table I also summarizes for comparison the noise performance of each device normalized with respect to the RMS current spectral density ($\text{pA}/\sqrt{\text{Hz}}$). For the REF102, the current spectral density is related to the voltage spectral density by $R_{SET}$, which in turn is determined by $I_D$. Thus, the spectral noise, in this case, is represented as a function of $I_D$. The thermal noise from $R_{SET}$ used in the REF102 and LT3092 circuits also contributes to the total noise. As the noise depends upon the value of $R_{SET}$, this noise component will effectively depend upon $I_D$. The last row gives the total noise spectral density $e_{nD}$ for each device as a function of $I_D$.

TABLE II
NOISE COMPARISON OF THE INA217 [30] AND AD8221 [31] INAMPS.

| | Voltage noise (μV) | Current noise (μV) | Total noise (μV) |
|---|---|---|---|
| AD8221 [31] | 2.54 | 0.05 | 2.54 |
| INA 217 [30] | 0.50 | 0.84 | 0.98 |

Based upon the above observations, we see that the REF102 circuit has the lowest noise (as well as a higher initial accuracy). For currents greater than 9 mA, however, the LT3092 exhibits lower noise. As no information on the drift for the LT3092 is available, the REF102 circuit is taken to be more reliable, with a known performance limit. The current required for our application is also well within 5 mA. Hence, we use the REF102-based implementation. In order to eliminate any noise from the AC power source being injected into the circuit, we use lead-acid batteries to power the circuit.

It is important to note that, despite the excellent performance specifications of the REF102, the actual performance realized is limited by the specifications of the $R_{SET}$ resistor used. In particular, the actual drift is limited by the temperature coefficient ('tempco') of the resistor. For a tempco of 50 ppm/°C, the drift seen is 47.5 ppm/°C, which is worse than the 25 ppm/°C of the REF200. For tempco of 1 ppm/°C, the drift reduces to 1.5 ppm/°C. It is also essential to use resistors with minimal excess noise. Thin film or metal film resistors strike a balance between cost and performance. We use a 10 ppm/°C, 0.1 % tolerance thin film resistor. Finally, note that in the above analysis, the implicit assumption is that the noise from the current source is transmitted entirely to the sensor. However, due to the voltage feedback scheme in the REF102-based circuit, some of the noise is expected to cancel out, with only part of it being transmitted, though it is difficult to predict in what proportion. Therefore, the subsequent noise analysis represents a *worst-case* scenario leading to a *conservative* estimate for the noise.

### C. Instrumentation Amplifier Selection

Typical characteristics of interest are the input offset voltage, the input bias current, the common-mode rejection ratio (CMRR), the gain-bandwidth product (GBW), and the noise. While the CMRR and GBW must be considered, the input offset voltage and bias current, however, are not as important in our application as their main effect is to produce an output offset voltage, which can be accounted for adequately through calibration. The noise is the primary factor of interest, and it has two components: current noise and voltage noise. The current noise flows through the source impedance to produce a corresponding voltage noise, as shown in Sec. III-A and in (2). The relative contribution of the current and voltage noise, thus, depends upon the source and is critical to identify, as a typical inamp is a compromise between the two.

We analyze three inamps (two with adjustable gain and one with fixed gain) with different noise specifications and estimate their total noise. The AD8221 (Analog Devices) [31] has lower output voltage noise, while the INA217 (Texas Instruments) [30] has a lower input voltage noise level of 1.3 $\mathrm{nV}/\sqrt{\mathrm{Hz}}$, compared to the 8 $\mathrm{nV}/\sqrt{\mathrm{Hz}}$ for the AD8221. However, the

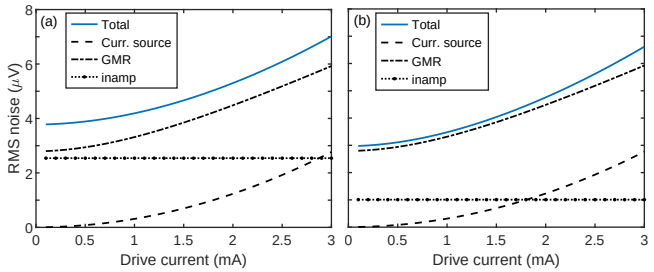



Fig. 6. Individual noise components and the total noise as a function of the drive current $I_D$ for the (a) INA217 and (b) AD8221.

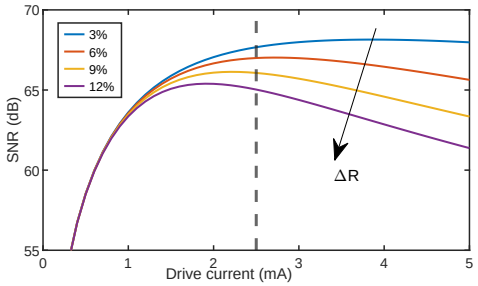
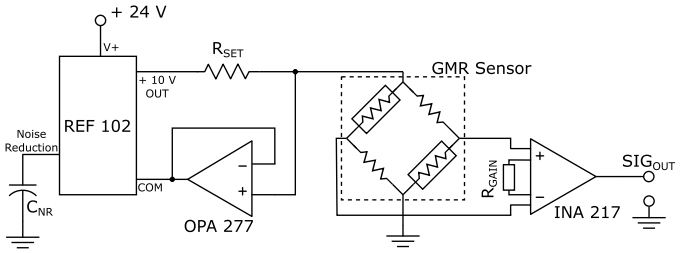



Fig. 7. SNR vs. drive current for different sensor imbalance to nominal element resistance ($\Delta R/R$) values. The gray dashed line-type indicates the selected drive current of 2.5 mA.

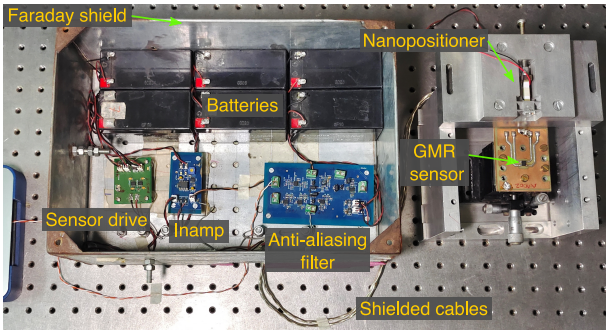
current noise for the AD8221 is just 40 $\mathrm{fA}/\sqrt{\mathrm{Hz}}$, 20 times lower than the 800 $\mathrm{fA}/\sqrt{\mathrm{Hz}}$ for the INA217. The 1/f noise, for both voltage and current, is higher in the case of the INA217. The fixed gain AD8228 (Analog Devices) has a noise performance very similar to that of the AD8221. The CMRR for each device is close to 100 dB. Apparently, the AD8221 is a significantly better inamp.

For this analysis, we select a gain of 100 and an operating bandwidth of 100 kHz. The total referred to the input (RTI) RMS noise for the INA217 and AD8221 inamps are estimated using standard methods as given in [32] and are summarized in Table II. The noise performance for the AD8228 is similar to that of the AD8221. For the gain of 100, the 8 $\mathrm{nV}/\sqrt{\mathrm{Hz}}$ noise density for the AD8221 translates to 800 $\mathrm{nV}/\sqrt{\mathrm{Hz}}$, as compared to 130 $\mathrm{nV}/\sqrt{\mathrm{Hz}}$ for the INA217. The higher voltage output noise only marginally affects the INA217, with an overall voltage noise that is still lower than that of the AD8221 by a factor of five. Thus, even though the current noise exhibited by the AD8221 is close to 17 times lower, the overall (total) noise is still 2.5 times higher. The GMR sensor's relatively lower resistance limits the benefits of the much lower current noise of the AD8221 over the INA217. For a higher source resistance, however, the AD8221 would perform better. For example, for a negative temperature coefficient (NTC) thermistor, with a resistance of 25 kΩ, the RTI noise would be 2.55 and 4.58 $\mu$V RMS for the AD8221 and the INA217, respectively. If the bandwidth were reduced, as is typically the case with an NTC sensor, the difference in exhibited noise increases several-fold. The need for device-specific selection of inamp has been highlighted earlier [23], though no selection criteria were given. The analysis given above can be used to select the best available amplifier for any choice of source and is not limited only to GMR sensors.

Based upon the above discussion, we select the INA217




(a)




(b)

Fig. 8. The implemented GMR readout circuit; (a) circuit diagram (decoupling capacitors have been omitted for brevity), and (b) photograph of the readout circuit, which is placed in a Faraday shield for protection from electromagnetic interference.

for this application. We use a 100 Ω low-noise thin film resistor with a temperature coefficient of 10 ppm/°C. This is comparable to the 3 ppm/°C temperature coefficient of the internal laser-trimmed resistors of the inamp. This results in a gain of 101 with a gain drift no greater than 0.001/°C. In terms of displacement, this translates to a maximum drift of 1 pm/nm/°C.

### D. Total (Combined) Noise and SNR

The individual noise from each of the three devices are combined to obtain the total RTI noise of the inamp. As each noise source is uncorrelated to the other, the noise sources sum orthogonally or as root-sum-of-squares, i.e., the total noise $V_n = \sqrt{V_{nD}^2 + V_{nS}^2 + V_{nA}^2}$, where $V_{nD}$, $V_{nS}$ and $V_{nA}$ are the noise components of the current source, sensor and inamp respectively. Fig. 6 shows the variation with $I_D$ of these individual noise components, as well as the total RTI noise, for the INA217 and AD8221 inamps. In both cases, the inamp noise is relatively unaffected by any increase in $I_D$; the GMR and current source noise, however, vary considerably with $I_D$. In the case of the INA217, the inamp noise is significantly lower than that of the other two components. The dominant noise component is the GMR noise, with the total noise curve almost coinciding with the GMR noise curve at lower current levels. Only at higher $I_D$ does the effect of the current source noise become considerable. In the case of the AD8221, however, the inamp noise's contribution is considerable throughout, with the current source noise not significantly adding more noise. Therefore, the total noise is consistently higher, with a higher proportion of noise contribution from the readout circuit than when the INA217 is employed.

Another parameter affecting the current source noise component is the sensor bridge imbalance $\Delta R$, as seen in (2). $\Delta R$, and consequently the current source noise, increases with the applied magnetic field bias $B$. However, over the linear operating region of the sensor, the magnetic gradient $\frac{dB}{dx}$ does not vary significantly with $B$. As a result, the SNR reduces with increasing $B$ and, equivalently, with increasing $\Delta R$. Fig. 7 shows the SNR as a function of $I_D$ for different $\Delta R$. For higher values of $\Delta R$ the SNR can actually reduce with increasing $I_D$. As the sensing action depends only upon $\frac{dB}{dx}$, and not $B$, the sensor should be operated at the lowest $B$ within the linear region in order to achieve the best SNR. This corresponds to the orange curve in Fig. 7. As the peak SNR is reached near 2.5 mA, we select this as the sensor drive current. The estimated worst-case total RMS noise is 0.56 mV, while that of the GMR sensor alone is 0.52 mV. It is noteworthy that as a result of the presented design process, the additional noise due to the readout circuit is minimal. Also, since this is a worst-case estimate, the actual exhibited noise can possibly be lower.

## IV. Extrinsic Noise and Experimental Validation

Next, the circuit design expounded upon in the previous section is implemented. The drive current is set to 2.5 mA, which enables maximal SNR while also limiting voltage swings within the circuit. Fig. 8(a) shows the complete circuit diagram and Fig. 8(b) a photograph of the setup. The circuit is enclosed in a Faraday shield to minimize external noise pickup. It also protects from air currents that cause temperature fluctuations and affect the performance of the inamp [30]. Shielded twisted-pair cables are used to interface the circuit and the sensor. Lead-acid batteries are used to power the circuit in order to minimize input noise. Fig. 9(a) shows the voltage spectral density of the noise measured at the output of the circuit. The 1/f trend is clearly seen up to nearly 10 kHz, and from thereon, it tends towards the thermal noise floor. The measured RMS noise is 0.53 mV, over a bandwidth of 100 kHz, which is close to but slightly lower than the 0.56 mV estimated in Section III-D. As noted there, that estimate was for a worst-case scenario, where the noise from the sensor drive source is assumed to be wholly transmitted to the sensor. Thus, the readout circuit negligibly degrades the sensor's output, and the sensor's inherent resolution is almost entirely utilized.

The sensor is calibrated using a laser Doppler vibrometer (LDV) (NLV-2500, Polytec GmbH, Waldbronn, Germany). The sensitivity obtained is 220 µV/µm; thus, the resolution of the system is 2.46 nm over a bandwidth of 100 kHz. This sensitivity is comparable to that earlier demonstrated by Kartik *et al.* [17], while the bandwidth has now improved by a factor of 10. The readout circuit's performance is remarkable given the high bandwidth and the sensor's high intrinsic noise, especially its higher 1/f component. Typically, modulation/demodulation

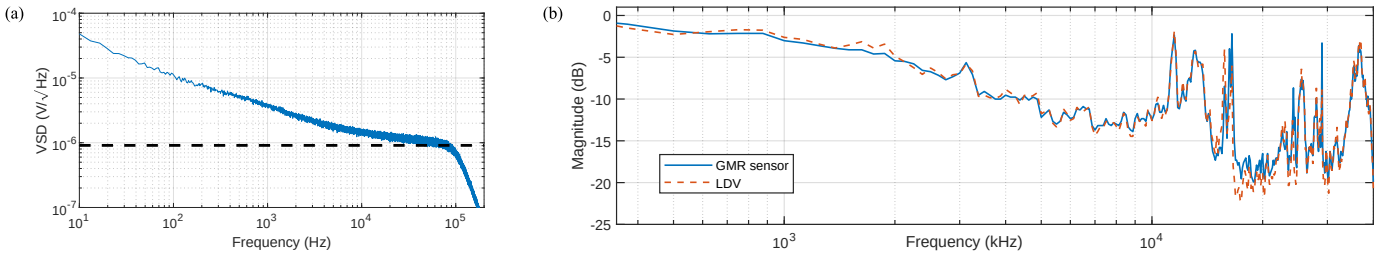



Fig. 9. (a) Voltage spectral density of the noise measured at the output of the system. The black dotted line-type indicates the thermal noise floor. (b) Frequency response of the nanopositioner to a white noise input signal measured simultaneously using the GMR sensor and LDV.

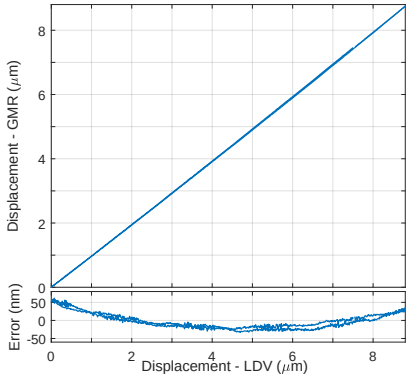



Fig. 10. Hysteresis measured over the nanopositioner's scanning range.

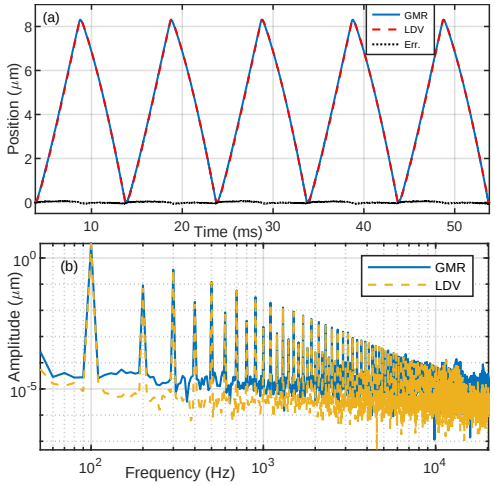



Fig. 11. Position tracking of the nanopositioning scanner. The (a) displacement, and (b) the frequency spectrum of the measured trajectories.

schemes or a lock-in amplifier are employed to negate the effect of 1/f noise and achieve a high resolution. Not only are these schemes complex, but their requirement for increasingly higher carrier frequencies with bandwidth renders them impractical for high-bandwidth sensing. The readout circuit presented here is simple to implement and employs commercially available off-the-shelf (COTS) components, even while exhibiting performance comparable to state-of-the-art sensors such as capacitive sensors. Note that the above resolution is achieved solely by limiting the measurement noise, and no attempt has been made to maximize the sensitivity. A significant increase in the field gradient, and therefore the sensitivity, can be achieved by reducing the sensor-magnet separation [17], [21] and by the use of smaller magnets [20]. The increased sensitivity can directly result in increased resolution, thereby enabling even sub-nanometer resolution. Additionally, the increased sensitivity can be used to offset the increase in noise with bandwidth, thus enabling nanometer resolution over the sensor's full bandwidth of 1 MHz.

The sensor is deployed in an in-house developed high-speed Atomic Force Microscopy (AFM) nanopositioning scanner. To demonstrate the bandwidth, the frequency response of the scanner to white noise excitation is simultaneously measured using both the GMR sensor and the LDV for reference, and is shown in Fig. 9(b). The measured responses match very well over the entire operating bandwidth. The scanner has several eigenmodes with eigenfrequencies above 10 kHz, which are detected by the GMR sensor.

The GMR sensor is inherently prone to hysteresis. To measure this hysteresis, the scanner is repeatedly displaced forward and backward, and the measured displacement is compared with that obtained using the LDV. Fig. 10 shows the measured displacement versus the reference LDV displacement measurement, as well as the error. The sensor exhibits hysteresis of just under 1 % of the scan range, which is similar to previously reported levels [20]. Hysteresis being a systematic error – unlike noise which is stochastic in nature – is relatively more tractable, and several corrective measures have been reported. Compensation schemes based upon hysteresis modeling [33] seem to be the most promising, requiring minimal alteration to the readout circuitry, in contrast to schemes involving alternative or additional biasing measures [34].

Finally, the sensor's position sensing performance is demonstrated by tracking the scanner's motion when actuated using a triangular signal at 100 Hz. This is typical of actuation signals employed in high-speed scanning applications. Fig. 11(a) depicts the quality of the GMR position sensing with reference to the LDV measurement; no visible distortion is discernible over the entire scanning range. This is confirmed from Fig. 11(b), which shows the frequency spectra of the two measurements. The absence of harmonic distortion further confirms the high bandwidth of the system.

## V. Summary

A low-noise high-bandwidth GMR sensor-readout circuit for nanometer resolution position sensing has been presented. The noise interactions between the GMR sensor and the other readout circuit components have been investigated, and methods to minimize both intrinsic and extrinsic noise have been discussed. The GMR sensor's 1/f noise was shown to critically affect the achievable SNR, even at high bandwidths;

its dependence on drive current leads to diminishing SNR improvements as drive current is increased. Selection of the inamp based on the source (sensor) impedance was shown to critically impact SNR, based on which a method to judiciously select the inamp was detailed. A system-level design process for the readout circuit has been presented, which is sufficiently general to be applicable to other low-noise readout applications, particularly those employing bridge sensors. A resolution of 2.5 nm over a bandwidth of 100 kHz has been demonstrated, a ten-fold improvement over previously-reported performance and comparable to other, state-of-the-art – but more complex – position sensing schemes. The readout scheme is simple to implement and easily extendable to higher bandwidths, unlike other high-performance schemes, such as modulation/demodulation, where the requirement for increasingly higher carrier frequencies renders further improvements in bandwidth impractical. The sensor is deployed in a high-speed AFM nanopositioner and is shown to sense position and measure its resonant dynamics remarkably well. The sensor's sensitivity can be further improved by enhancing the magnetic field gradient, as discussed, rendering nanometer to sub-nanometer resolution over the sensor's full bandwidth of 1 MHz achievable.

## REFERENCES


[1] M. V. Salapaka and S. M. Salapaka, "Scanning probe microscopy," *IEEE Control Syst. Mag.*, vol. 28, no. 2, pp. 65–83, 2008.

[2] A. A. Tseng, A. Notargiacomo, and T. P. Chen, "Nanofabrication by scanning probe microscope lithography: A review," *J. Vac. Sci. Technol. B*, vol. 23, no. 3, p. 877 – 894, 2005.

[3] A. D. L. Humphris, B. Zhao, D. Catto, J. P. Howard-Knight, P. Kohli, and J. K. Hobbs, "High speed nano-metrology," *Review of Scientific Instruments*, vol. 82, no. 4, p. 043710, 2011.

[4] V. Kartik, A. Pantazi, and M. A. Lantz, "Track-following high frequency lateral motion of flexible magnetic media with sub-100 nm positioning error," *IEEE Trans. Magn.*, vol. 47, no. 7, pp. 1868–1873, 2011.

[5] A. J. Fleming, "A review of nanometer resolution position sensors: Operation and performance," *Sens. Actuator A Phys.*, vol. 190, pp. 106–126, 2013.

[6] T. Ando, T. Uchihashi, and S. Scheuring, "Filming biomolecular processes by high-speed atomic force microscopy," *Chem. Rev.*, vol. 114, no. 6, pp. 3120–3188, 2014.

[7] O. D. Payton, L. Picco, and T. B. Scott, "High-speed atomic force microscopy for materials science," *Int. Mater. Rev.*, vol. 61, no. 8, pp. 473–494, 2016.

[8] S. Ito, M. Poik, J. Schlarp, and G. Schitter, "Atomic force microscopy breaking through the vertical range-bandwidth tradeoff," *IEEE Trans. Ind. Electron.*, vol. 68, no. 1, pp. 786–795, 2021.

[9] S. Nihtianov, "Measuring in the subnanometer range: Capacitive and eddy current nanodisplacement sensors," *IEEE Ind. Electron. Mag.*, vol. 8, no. 1, pp. 6–15, 2014.

[10] S. Yang and G. Zhang, "A review of interferometry for geometric measurement," *Meas. Sci. Technol.*, vol. 29, no. 10, p. 102001, 2018.

[11] G. Chen, B. Zhang, P. Liu, and H. Ding, "An adaptive analog circuit for LVDT's nanometer measurement without losing sensitivity and range," *IEEE Sensors J.*, vol. 15, no. 4, pp. 2248–2254, 2015.

[12] "Microsense 6810 ultra high resolution capacitive position sensor," Microsense LLC, Lowell, MA, USA, 2013.

[13] Y. K. Yong, S. S. Aphale, and S. O. R. Moheimani, "Design, identification, and control of a flexure-based xy stage for fast nanoscale positioning," *IEEE Trans. Nanotechnol.*, vol. 8, no. 1, pp. 46–54, 2009.

[14] X. Yang and W.-L. Zhu, "Design, analysis, and control of an xy parallel nanomanipulator with multiple actuation modes," *IEEE Trans. Ind. Electron.*, vol. 67, no. 9, pp. 7639–7648, 2020.

[15] B. George, Z. Tan, and S. Nihtianov, "Advances in capacitive, eddy current, and magnetic displacement sensors and corresponding interfaces," *IEEE Trans. Ind. Electron.*, vol. 64, no. 12, pp. 9595–9607, 2017.

[16] D. R. Sahoo, A. Sebastian, W. Häberle, H. Pozidis, and E. Eleftheriou, "Scanning probe microscopy based on magnetoresistive sensing," *Nanotechnology*, vol. 22, no. 14, p. 145501, Feb. 2011.

[17] V. Kartik, A. Sebastian, T. Tuma, A. Pantazi, H. Pozidis, and D. R. Sahoo, "High-bandwidth nanopositioner with magnetoresistance based position sensing," *Mechatronics*, vol. 22, no. 3, pp. 295 – 301, 2012.

[18] N. A. Stutzke, S. E. Russek, D. P. Pappas, and M. Tondra, "Low-frequency noise measurements on commercial magnetoresistive magnetic field sensors," *J. Appl. Phys.*, vol. 97, no. 10, p. 10Q107, 2005.

[19] H. T. Hardner, S. S. P. Parkin, M. B. Weissman, M. B. Salamon, and E. Kita, "1/f noise in giant magnetoresistive materials," *J. Appl. Phys.*, vol. 75, no. 10, pp. 6531–6533, 1994.

[20] T. Tuma, A. Pantazi, D. R. Sahoo, P. Eib, G. Salis, H. Pozidis, and A. Sebastian, "A high-bandwidth spintronic position sensor," *Nanotechnology*, vol. 25, no. 37, p. 375501, Aug. 2014.

[21] T. J. Joseph and V. Kartik, "Nanometer resolution dynamic displacement measurement in industrial/high noise environments," in *Proc. IEEE Int. Conf. on Ind. Technol. (ICIT)*, Mar. 2016, pp. 673–678.

[22] "GMR sensor catalog," NVE Corporation, Eden Prairie, MN, USA, p. 105, Feb. 2012.

[23] F. Seifert, "Resistor current noise measurements," LIGO, Apr. 2009. [Online]. Available: https://dcc.ligo.org/public/0002/T0900200/001/current_noise.pdf [Accessed Jan. 18, 2021]

[24] W. Jung, *Op Amp Applications Handbook*. Amsterdam, The Netherlands: Elsevier, 2005.

[25] S. J. Shah, "Field wiring and noise considerations for analog signals - Application Note 025," National Instruments, Austin, TX, USA, 1992. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.607.6492 [Accessed Oct. 15, 2021]

[26] "10V Precision Voltage Reference," Texas Instruments, Dallas, TX, USA, Datasheet, Sept. 2000 [Revised Jun. 2009].

[27] "200mA 2-Terminal Programmable Current Source," Analog Devices, Norwood, MA, USA, Datasheet, 2009 [Revised 2020].

[28] "REF200 Dual Current Source and Sink," Texas Instruments, Dallas, TX, USA, Datasheet, Sept. 2000 [Revised Feb. 2020].

[29] R. M. Stitt, "Make a precision current source or current sink," Burr-Brown, Tucson, AZ, USA, 1994.

[30] "INA217 Low-Noise, Low-Distortion Instrumentation Amplifier Replacement for SSM2017," Texas Instruments, Dallas, TX, USA, Jun. 2002 [Revised Nov. 2015].

[31] "Precision Instrumentation Amplifier AD8221," Analog Devices, Norwood, MA, USA, Datasheet, 2011.

[32] A. Kay, *Operational Amplifier Noise*. Boston, MA, USA: Newnes, 2012.

[33] I. Jedlicska, R. Weiss, and R. Weigel, "Linearizing the output characteristic of gmr current sensors through hysteresis modeling," *IEEE Trans. Ind. Electron.*, vol. 57, no. 5, pp. 1728–1734, 2010.

[34] F. Xie, R. Weiss, and R. Weigel, "Hysteresis compensation method for magnetoresistive sensors based on single polar controlled magnetic field pulses," *IEEE Trans. Ind. Electron.*, vol. 64, no. 1, pp. 710–716, 2017.

**Tim John Joseph** (S'16) received the B.Technology degree in mechanical engineering from Mahatma Gandhi University, Kottayam, India, in 2009, and the M.Technology degree in mechanical engineering from IIT Bombay, Mumbai, India, in 2012. He is currently pursuing the Ph.D. degree in mechanical engineering at IIT Bombay.

His research interests include precision dynamics, low-noise electronics, high-speed nanoscale positioning and sensing, and vibration control.

**V. Kartik** received a B.Engineering from Delhi College of Engineering, New Delhi, India, an M.S. from The Ohio State University, Columbus, OH, USA, and a Ph.D. from Carnegie Mellon University, Pittsburgh, PA, USA, in mechanical engineering. After receiving his Ph.D., he was with the IBM Zurich Research Laboratory in Switzerland. He joined IIT Bombay in 2011, where he is currently an Associate Professor of Mechanical Engineering. His research interests include atomic force microscopy, non-linear dynamics, sensors and actuators, and high-speed dynamic position control systems.

arXiv:2209.10966v1 [cs.CL] 22 Sep 2022

Anmol Bansal
Arjun Choudhry
Anubhav Sharma
Seba Susan

# ADAPTATION OF DOMAIN-SPECIFIC TRANSFORMER MODELS WITH TEXT OVERSAMPLING FOR SENTIMENT ANALYSIS OF SOCIAL MEDIA POSTS ON COVID-19 VACCINES

**Abstract** *Covid-19 has spread across the world and several vaccines have been developed to counter its surge. To identify the correct sentiments associated with the vaccines from social media posts, we fine-tune various state-of-the-art pre-trained transformer models on tweets associated with Covid-19 vaccines. Specifically, we use the recently introduced state-of-the-art pre-trained transformer models RoBERTa, XLNet and BERT, and the domain-specific transformer models CT-BERT and BERTweet that are pre-trained on Covid-19 tweets. We further explore the option of text augmentation by oversampling using Language Model based Oversampling Technique (LMOTE) to improve the accuracies of these models, specifically, for small sample datasets where there is an imbalanced class distribution among the positive, negative and neutral sentiment classes. Our results summarize our findings on the suitability of text oversampling for imbalanced small sample datasets that are used to fine-tune state-of-the-art pre-trained transformer models, and the utility of domain-specific transformer models for the classification task.*

**Keywords** Covid-19, Vaccine, Transformer, Twitter, BERTweet, CT-BERT, BERT, XLNet, RoBERTa, Text oversampling, LMOTE, Class imbalance, Small sample dataset

**Note: The paper has been accepted for publication in Computer Science journal: http://journals.agh.edu.pl/csci**

## 1. Introduction

Covid-19 vaccines were developed in response to the widespread devastation caused by the Covid-19 pandemic which started in December 2019 and has been on the rise ever since. These vaccines have elicited a mixed response among the general public, and these reviews help us to understand just how these vaccines have affected people emotionally. Social media is the primary platform for finding solutions to health-based queries related to the Covid-19 pandemic [2]. One of the best ways to determine public opinion is to survey Twitter which has millions of users every day and receives lots of tweets on vaccines from all over the world [35]. Most of the researchers who have used machine learning to analyze Covid-19 vaccine-related tweets have adopted unsupervised techniques to determine the sentiments of tweets [21]. A majority of researchers use the rule-based Valence Aware Dictionary for Sentiment Reasoning (VADER) [15] which assigns a sentiment score to each tweet in an unsupervised manner [21, 23, 39]. Other unsupervised techniques used for Twitter sentiment analysis include AFINN [28] and TextBlob [27]. However, the unsupervised techniques are dependent on predefined rules meant for general sentiment analysis that may not work out for text data related to the ongoing pandemic. Supervised learning methodologies overcome the drawback of unsupervised techniques by learning appropriate text patterns that distinguish between positive, negative and neutral sentiments [45, 32, 43]. Supervised learning for text classification is achieved in the present times using transformers that are now popularly replacing Long Short-Term Memory (LSTM) [26] and Convolutional Neural Network (CNN) [11] in various Natural Language Processing (NLP) tasks [13, 33]. Transformer models can be used for sequence modeling to predict the next word in a sentence, and are usually trained on a large corpus such as Wikipedia or Brown Corpus [41, 9]. These pre-trained models are generalized, and are usually fine-tuned for downstream tasks such as classification and text generation.

One of the challenges faced in supervised machine learning is the small size of the dataset which leads to less accurate models. Such datasets are called small sample datasets (SSD) [16]. The situation is complicated when the class distribution is imbalanced with the majority class samples outnumbering the minority class samples [35]. SSD provides fewer examples for the model to identify patterns from, and therefore results in less accurate models. An imbalanced dataset can also be detrimental to the model as it results in biased results towards the majority class and therefore results in wrong predictions. This paper aims to explore solutions to the class imbalance problem associated with the sentiment categories of Covid-19 related tweets when the size of the dataset is small. We explore the viability of text-based oversampling as a possible solution. The synthetic tweets generated are from classes having a lower population in order to balance the population of all classes while increasing the number of training samples at the same time. We explore different pre-trained transformer

models for supervised learning from an imbalanced, small sample dataset containing tweets on Covid-19 vaccine related discussions, and compare their performance with that of domain-specific transformer models for the classification task. Specifically, we investigate the performance of the state-of-the-art transformer models RoBERTa, BERT and XLNet as compared to the domain-specific pre-trained transformer models CT-BERT and BERTweet for sentiment analysis of Covid-19 vaccine related tweets. CT-BERT and BERTweet are pre-trained transformer models obtained after intensive training on English tweets related to the Covid-19 pandemic. On the other hand, RoBERTa, XLNet and BERT provide input embeddings for sentences written in English, and are used for generalized natural language processing. CT-BERT and BERTweet models have the advantage of being familiar with text patterns emanating from Covid-19 related discussions such as "Covid positive" which may not be understood well by the RoBERTa, BERT and XLNet models.

The organization of this paper is as follows. Section 2 describes some related work on text oversampling of Covid-19 related text datasets, and also reviews the LMOTE algorithm used for text oversampling in the current work, Section 3 outlines several state-of-the-art pre-trained transformer models including domain-specific transformer models, Section 4 presents the methodology, Section 5 contains a detailed analysis of the results, and Section 6 concludes the paper.

## 2. Preliminaries of text oversampling

### 2.1. Text oversampling of Covid-19 related text

The Small Sample Size (SSS) problem refers to the availability of a small number of training samples in high dimensional datasets [25]; this leads to inadequate training rendering supervised learning a challenging task. This paper uses a small subset of annotated Covid-19 tweets to observe the effects of using a smaller text dataset for fine-tuning pre-trained transformer models that are the state of the art for implementing various NLP tasks. The situation is complicated when the class distribution is uneven and the number of majority samples outnumber the number of minority samples [35]. In literature, there are several examples of text oversampling being applied to Covid-19 related social media posts due to the apparent scarcity of text samples belonging to some of the minority classes. We discuss some of these works next. In [22], Liu et al. proved that oversampling the term-frequency inverse document frequency (TF-IDF) features using Synthetic Minority Oversampling Technique (SMOTE) [10] improved the results of Covid-19 vaccine hesitancy prediction. Support Vector Machine (SVM) was used for the classification. SMOTE was also used in [4] to oversample word embeddings for sentiment analysis of Arabic tweets related to Covid-19 conspiracy theories. A recent work [36] investigated ensemble models for the classification of Covid-19 infodemic tweets oversampled using SMOTE. Mohsen et al. [31] recommended text oversampling using SMOTe Edited Nearest Neighbor (SMOTEENN) for the sentiment analysis of Arabic tweets related to Covid-19 quarantine.

Random oversampling was performed in [5] for detecting Covid-19 misinformation on Twitter.

### 2.2. A review of the LMOTE algorithm for text oversampling

Language Model-based Oversampling Technique (LMOTE), proposed by Leekha et al. in 2020 [20] is a language modelling based synthetic datapoints generation approach for tackling the problem of class imbalance in natural language processing tasks. Previous synthetic datapoints generation approaches for tackling class imbalance, like SMOTE and its variants [42], lack the ability to allow for proper qualitative analysis of the generated synthetic data points since the synthetic samples were generated in the Euclidean space. This made it difficult to concretely judge the semantic and contextual validity of the generated synthetic data points. Unlike SMOTE and its variants, LMOTE works specifically on textual data, and the generated synthetic data points using LMOTE allow for more concrete and intuitive balancing of the dataset.

In our current work on Covid-19 vaccine sentiment analysis, there are three classes of sentiments: positive, negative and neutral. The neutral tweets are large in number since a lot of people tweet about generic information regarding vaccines without expressing any sentiments. Hence neutral sentiment is the majority class in our problem. The algorithm for text oversampling of the tweets belonging to the minority class (positive and negative tweets in our case) using LMOTE is given below.

**Algorithm** Text oversampling of the minority class using LMOTE

**Input**: minority class (positive/ negative tweets)

**Output**: Balanced dataset with augmented minority class

1: Compile the tweets in the minority class into one text corpus
2: Find the most frequently occurring 5-grams in this text corpus (select top-100)
3: Train a Bidirectional LSTM language model on the selected 5-grams
4: Use the trained language model to generate tweets one word at a time
5: REPEAT for all minority classes.

## 3. Pre-trained transformer models for domain-specific tasks

Transformers introduced by Vaswani et al. in 2017 [44] rely on the concept of self-attention that involves computation of intra-attention between positions in the input sequence. Transformer is rapidly replacing LSTM and CNN in encoder-decoder models that incorporate an attention mechanism between the encoder and decoder [7, 21]. Google's BERT [12] and XLNet [46] are bi-directional transformer models used for learning representations for various NLP tasks, with XLNet outperforming BERT on several tasks that involve learning from long sequences [3, 46]. BERT is a transformer-based model based on masked language modeling. BERT and its advanced versions such as RoBERTa [24] and ALBERT [17] are trained on the English Wikipedia and the Brown Corpus. Pre-trained BERT models generate word embeddings that can

be used for text understanding and classification. BERT can also be fine-tuned to adapt to specific tasks. XLNet is an autoregressive (AR) language model which uses permutation language modeling during the pre-training phase. Though it is similar in architecture to BERT, it differs in its pre-training objective due to which it surpasses BERT in various NLP tasks. XLNet is pre-trained using only a subset of output tokens as a target. Like BERT, pre-trained XLNet models can also be used for many other downstream tasks while also increasing the limits for sequences. In a recent work, the XLNet transformer was used successfully for sentiment analysis of unlabeled Covid-19 tweets by transfer learning [8]. The XLNet transformer was pre-trained on the US Airlines tweets dataset that was unconnected with Covid-19. Several language-specific models have been developed such as CamemBERT for French [29] and GottBERT for the German language [40]. These specific language models give better results than the BERT multi-language model. The generalized transformer models can be further trained on downstream tasks to create separate models for different tasks and different languages. We discuss some of these domain-specific models here. Each of them is trained on a specialized corpus, relevant to the topic at hand, and is more effective in that domain.

1. SciBERT (biomedical and computer science literature corpus): It is a BERT-based language model for performing scientific tasks. It was introduced by Beltagy et al. in 2019 [9].

2. FinBERT (financial services corpus): This is a pre-trained NLP model proposed by Araci in 2019 [6] for analyzing the sentiments of financial statements, and is trained using a large financial corpus.

3. BioBERT (biomedical literature corpus): This NLP model pre-trained on biomedical corpora outperformed BERT and various state-of-the-art models in a variety of biomedical text mining tasks. It was introduced by Lee et al. in 2019 [19].

4. ClinicalBERT (clinical notes corpus): This model focuses on clinical notes and its representations using bidirectional transformers and uncovers the relationship between medical concepts and humans as discussed by Huang et al. in 2019 [14].

5. mBERT (corpora from multiple languages): mBERT is a single BERT model proposed by Pires et al. in 2019 that is trained on 104 different languages [37]. The languages with fewer data were oversampled and those with surplus of data were undersampled to balance the corpus.

6. patentBERT (patent corpus): The patentBERT model is a fine-tuned pre-trained BERT model for patent classification proposed by Lee et al. in 2019 [18]. The fine-tuning was done using over 2 million patents and CNN with word embeddings.

7. RoBERTa (Optimized pre-training approach for BERT): RoBERTa is a robustly optimized pre-trained model based on BERT [24]. It is implemented on PyTorch and modifies key hyperparameters of BERT, including the removal of BERT's next-sentence pre-training objective and is trained with much larger mini-batches and learning rates. This helps it to achieve better downstream performance in the masked language modeling approach of BERT.

8. COVID-Twitter-BERT or CT-BERT (Covid-19 tweets): This is a domain-specific transformer-based model which is pre-trained on 160 million Twitter messages specifically related to Covid-19 [30]. The aim is to understand the content of social media posts related to the Covid-19 pandemic. Muller et al. proposed this model in 2020 [30] and applied it for five different classification tasks. The model gave an improvement over BERT on COVID-19 datasets but needed more pre-training to achieve similar performance on out-of-domain contents.

9. BERTweet (English Tweets): This model is a large-scale pre-trained language model for English Tweets and has the same architecture as the base BERT model [34]. Experiments have proved that this model outperforms strong baselines RoBERTa-base and XLM-R-base on various NLP tasks. BERTweet is the first public large-scale model pre-trained on English Tweets. It is trained using the RoBERTa pre-training procedure. This model is trained using a corpus of 850 million English Tweets comprising of 845 million tweets streamed from January 2012 to August 2019 and 5 million tweets belonging to the COVID-19 pandemic.

## 4. Implementation details

In our work, we test the suitability of text data augmentation for Covid-19 vaccine-related tweets applied for fine-tuning pre-trained transformer models. Data augmentation of the minority class is a popular remedy for class imbalance [38]. Due to the significant class-imbalance among positive, negative and neutral tweets, we oversample the positive and negative tweets only (minority class), concatenating the synthetic data points to the original dataset to generate a more balanced dataset. We adapt the LMOTE model for augmenting the data that is given as input to the pre-trained transformer models.

The dataset used is a small subset of the larger set of Covid-19 tweets presented by Gabriel Preda [1] which is annotated for positive, negative, and neutral sentiments by FullMoonDataScience. The dataset contains 6000 tweets with 3680 tweets belonging to the neutral class, 1900 tweets to the positive class, and 420 tweets to the negative class. The dataset is thus both small in size and highly imbalanced. The data format is shown in Table 1.

**Table 1**
Data format

| column | type |
|---|---|
| tweetID | integer |
| label | 1, 2, 3 |
| text | string |

The input text sequences from Covid-19 tweets are tokenized. We use the hugging-face library in Python for the implementation of the transformer models.

The text in the tweets is pre-processed by removing hashtags, links, emails, punctuations and extra spaces using regex. The sequential pre-processing steps are shown for an example tweet in Table 2. In addition to the steps shown, tabs and extra spaces are also removed.

**Table 2**
Sequential pre-processing steps for an example tweet

| **Pre-processing step** | **Output** |
|---|---|
| Original tweet | "More #GoodNewsfrom @bopanc & @DovLieber! #PfizerBioNTech's #COVID #vaccine is highly effective after just 1 dose & can be stored in ordinary freezers for up to 2 weeks, according to new data https://t.co/ ZWwi00rIU via @WSJ https://t.co/7TMIPCkkBa" |
| Removing @ mentions | "More #GoodNews from & ! #PfizerBioNTech's #COVID #vaccine is highly effective after just 1 dose & can be stored in ordinary freezers for up to 2 weeks, according to new data https://t.co/QZWwi00rIU via https://t.co/7TMIPCkkBa" |
| Removing Hash tags | "More from & ! 's is highly effective after just 1 dose & can be stored in ordinary freezers for up to 2 weeks, according to new data https://t.co/QZWwi00rIU via https://t.co/7TMIPCkkBa" |
| Removing websites | "More from & ! 's is highly effective after just 1 dose & can be stored in ordinary freezers for up to 2 weeks, according to new data via " |
| Removing Every Punctuation Mark except -!?. | "More from amp ! 's is highly effective after just 1 dose amp can be stored in ordinary freezers for up to 2 weeks according to new data via " |
| Removing numbers | "More from amp ! 's is highly effective after just dose amp can be stored in ordinary freezers for up to weeks according to new data via " |

A three-fold cross-validation is performed for all the models in our experimentation:- RoBERTa, XLNet and BERT, and the domain-specific CT-BERT and BERTweet. The cross-entropy loss function and Adam Optimizer are used for training the models with a learning rate of 2e-5, and five epochs. Hyperparameter settings are chosen as per the guidelines in the original papers [25]. Fig. 1 shows the process flow pipeline showing the training and testing phases.

The tokenized text is subject to text oversampling, where the (minority classes) positive and negative tweets are oversampled using LMOTE such that the population of all three classes positive, negative and neutral are balanced. Some of the generated tweets are shown in Table 3 for reference. Neutral tweets were not oversampled since they constitute the majority class. Negative tweets were highly augmented using synthetic samples to match the population of the neutral class.

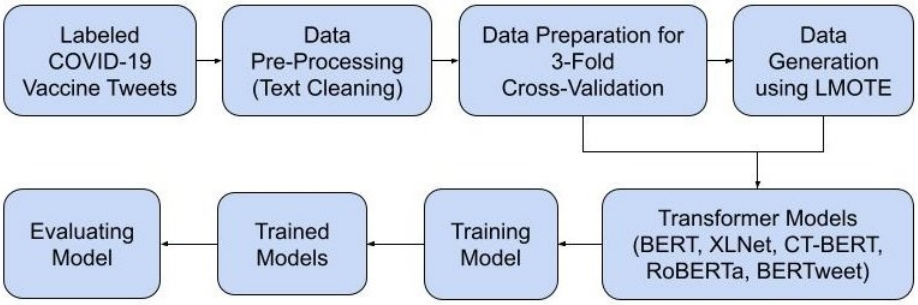



**Figure 1.** Process flow

**Table 3**
Some instances of tweets generated by text oversampling using LMOTE

| text generated | label |
|---|---|
| reporting cases of new cases in toronto closed business church no family no hope canada get out of syria effective and tech | positive |
| the commission has secured million additional doses of vaccine bringing the total number of doses secured to billion europeans will have had the | positive |
| home no family no business no church lent no hope until jab unproven vaccine order russia get out of syria that we have no idea of the | negative |
| company trial participants have died had received amp s s now not not worry to show for any but i know that experienced after the first dose of my | negative |

The augmented and balanced dataset is applied for fine-tuning of the pre-trained models RoBERTa[1], BERT[2] and XLNet[3], and the domain-specific pre-trained models BERTweet[4] and CT-BERT[5].

[1]https://github.com/facebookresearch/fairseq/blob/main/examples/roberta/README.md
[2]https://github.com/google-research/bert
[3]https://github.com/zihangdai/xlnet
[4]https://github.com/VinAIResearch/BERTweet
[5]https://github.com/digitalepidemiologylab/covid-twitter-bert

## 5. Results

Our experiments were performed in Python version 3.8 on a 2.8 GHz Intel core PC. We have made our source code available online[6] for research purposes. The Covid-19 tweets were pre-processed and tokenized as per the procedure outlined in Section 4. In order to explore the effects of text oversampling on our small sample dataset, we performed the experiment twice, one with text oversampling and one without. The text data was used to fine-tune the pre-trained models RoBERTa, BERT and XLNet, and the domain-specific pre-trained models CT-BERT and BERTweet. The transformer models were trained using three-fold cross-validation on three different 80-20 splits of the dataset, and then we obtained the mean of the performance metrics over the three runs.

### 5.1. Results without text oversampling

The performance metrics - test accuracy, F1-score and Mathew's correlation coefficient (MCC) are summarized in Table 4 for the five transformer models - RoBERTa, BERT, XLNet, CT-BERT and BERTweet in the absence of text oversampling. The corresponding receiver operating characteristic (ROC) curves (with Area Under Curve (AUC) readings) are plotted for the three sentiment classes in Fig. 2 (a, b, c).

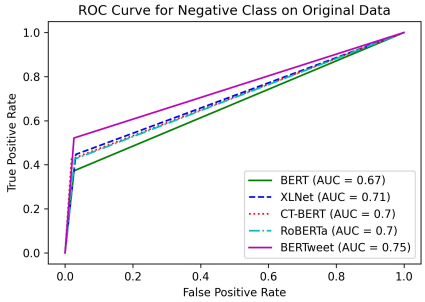
**Table 4**
Test accuracy, F1-Score and Mathew's correlation coefficient (MCC) for the five transformer models (w/o text oversampling)

| Model | Accuracy | F1-score | MCC |
|---|---|---|---|
| BERT | 75.06% | 0.63 | 0.51 |
| XLNet | 75.64% | 0.67 | 0.53 |
| RoBERTa | 76.44% | 0.67 | 0.55 |
| CT-BERT | 77.70% | 0.70 | 0.57 |
| BERTweet | 77.25% | 0.70 | 0.56 |

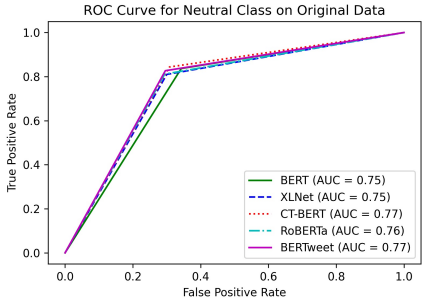
It is observed from Table 4, that the domain-specific CT-BERT and BERTweet models show a better performance as compared to the RoBERTa, BERT and XLNet models in terms of test accuracy, F1-score and MCC. RoBERTa is observed to be better than BERT and XLNet, while BERTweet proved to be almost as good as CT-BERT and better than RoBERTa. The reason for the better performance of CT-BERT and BERTweet models is that they are specifically trained on COVID-19 tweets, and are hence familiar with text patterns related to the pandemic-related discussions.

The detailed class-wise precision, recall, F1-score and MCC readings are presented in Table 5 for the five transformer models. There are three sentiment classes-neutral which is the majority class, and positive and negative which are minority

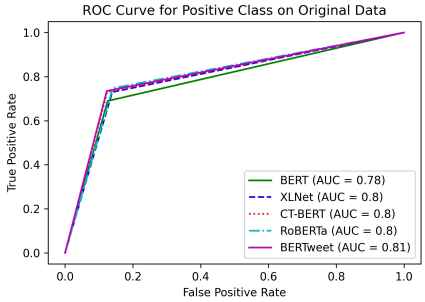
[6]https://github.com/Ace117MC/transformer-models-covid




(a) Negative class




(b) Neutral class




(c) Positive class

**Figure 2.** ROC curves and AUC values for different models (w/o text oversampling)

classes, with the negative tweets being very small in number. As expected, the results are biased, with the neutral and positive classes performing better than the negative class, as observed from both Table 5 (F1-score, MCC) and Fig. 2 (AUC). The negative tweets being very low in number are highly mis-classified as evident from the poor performance of the negative class. CT-BERT is the best performer out of all the five models, followed by BERTweet, in terms of test accuracy, F1-score and MCC, followed by RoBERTa and XLNet. The accuracies of CT-BERT and

**Table 5**
Class-wise performance analysis of the transformer models for positive, negative and neutral sentiments (w/o text oversampling)

| Model | Category | Precision | Recall | F1-score | MCC |
|---|---|---|---|---|---|
| BERT | neutral | 0.79 | 0.83 | 0.81 | 0.49 |
| BERT | positive | 0.72 | 0.69 | 0.70 | 0.57 |
| BERT | negative | 0.44 | 0.34 | 0.38 | 0.34 |
| XLNet | neutral | 0.82 | 0.80 | 0.81 | 0.51 |
| XLNet | positive | 0.70 | 0.74 | 0.72 | 0.58 |
| XLNet | negative | 0.50 | 0.48 | 0.49 | 0.45 |
| RoBERTa | neutral | 0.82 | 0.81 | 0.81 | 0.53 |
| RoBERTa | positive | 0.71 | 0.75 | 0.73 | 0.6 |
| RoBERTa | negative | 0.51 | 0.45 | 0.48 | 0.44 |
| CT-BERT | neutral | 0.81 | 0.84 | 0.82 | 0.55 |
| CT-BERT | positive | 0.75 | 0.70 | 0.72 | 0.61 |
| CT-BERT | negative | 0.58 | 0.54 | 0.56 | 0.53 |
| BERTweet | neutral | 0.81 | 0.82 | 0.81 | 0.53 |
| BERTweet | positive | 0.73 | 0.73 | 0.73 | 0.6 |
| BERTweet | negative | 0.56 | 0.52 | 0.54 | 0.51 |

BERTweet for the minority class (negative sentiment) are found significantly higher than the other models in Table 5, verifying that domain-specific transformer models mitigate the effect of the class imbalance to a certain extent, even in the absence of text augmentation.

## 5.2. Results with text oversampling

We next demonstrate the effects of text oversampling (of the positive and negative sentiment classes) using LMOTE to investigate the suitability of text augmentation prior to the training phase. The test accuracy, F1-score and MCC values are summarized in Table 6 for the five transformer models - RoBERTa, BERT, XLNet, CT-BERT and BERTweet when text oversampling is performed using LMOTE.

**Table 6**
Test accuracy, F1-Score and Mathew's correlation coefficient (MCC) for the five transformer models on dataset augmented using LMOTE

| Model | Accuracy | F1-score | MCC |
|---|---|---|---|
| BERT | 74.25% | 0.62 | 0.49 |
| XLNet | 75.80% | 0.65 | 0.53 |
| RoBERTa | 76.69% | 0.67 | 0.55 |
| CT-BERT | 76.14% | 0.67 | 0.52 |
| BERTweet | 77.78% | 0.68 | 0.56 |

A scrutiny of the results in Table 6 reveal a slight dip in the performance scores after text oversampling as compared to the readings in Table 4 (w/o text oversampling). This indicates that text oversampling of the minority classes for a small sample dataset will not improve the classification accuracy. For the augmented dataset, BERTweet performed best followed by RoBERTa and CT-BERT.

We also compare the performance of LMOTE with SMOTE; the performance scores for the five transformer models when performing text oversampling using SMOTE are compiled in Table 7. On comparing the scores of SMOTE in Table 7 with the results of LMOTE in Table 6, we note that the performance of LMOTE is found to be significantly higher than that of SMOTE. This proves that text generation by language modeling is a better option for augmentation of text corpora than the resampling strategies prevalent in data mining.

**Table 7**

Test accuracy, F1-Score and Mathew's correlation coefficient (MCC) for the five transformer models on dataset augmented using SMOTE

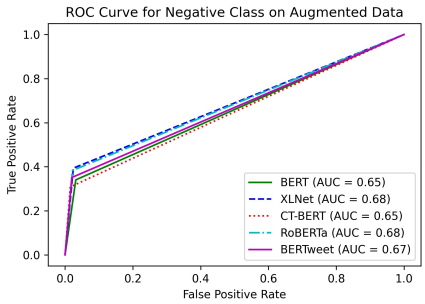
| **Model** | **Accuracy** | **F1-score** | **MCC** |
|---|---|---|---|
| BERT | 61.43% | 0.54 | 0.37 |
| XLNet | 56% | 0.51 | 0.36 |
| RoBERTa | 57.97% | 0.54 | 0.38 |
| CT-BERT | 69.58% | 0.63 | 0.47 |
| BERTweet | 69.02% | 0.6 | 0.45 |

The detailed class-wise accuracies are presented in Table 8 for all the five transformer models, in case of text augmentation using LMOTE. The corresponding ROC curves (with AUC readings) are plotted for the three sentiment classes in Fig. 3 (a, b, c).

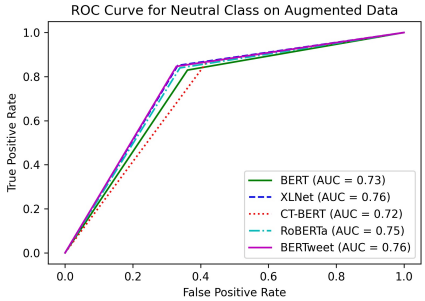
Both Table 8 and Fig. 3 indicate a decrease in the performance scores (after text oversampling) of the neutral class (majority class) as compared to the results of the original dataset in Table 5 and Fig. 2. The positive class accuracies remain more or less the same with a slight increase noted for some models. However, the negative class scores have distinctly increased for all models which can be attributed to text oversampling.

## 5.3. Discussion

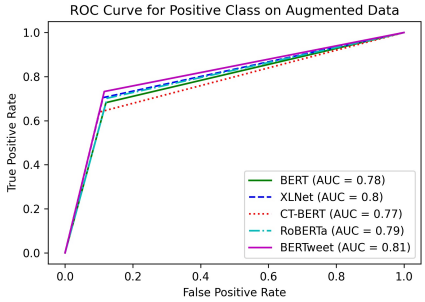
In our work, we investigate the utility of domain-specific pre-trained transformer models and text oversampling for the sentiment analysis of Covid-19 vaccine related tweets from an imbalanced small sample dataset. As observed from the performance scores in Tables 4-8, the domain-specific pre-trained transformer models CT-BERT and BERTweet significantly outperform the pre-trained transformer models - RoBERTa, BERT and XLNet. An instance of a Covid-19 tweet that is classified correctly by the domain-specific CT-BERT and BERTweet, but incorrectly by all other transformer




(a) Negative class




(b) Neutral class




(c) Positive class

**Figure 3.** ROC curves and AUC values for different models (with text oversampling)

models, is shown in Table 9, along with another instance of a tweet which is misclassified by all the models including CT-BERT and BERTweet. The latter tweet is a mixture of positive and negative news, though the human annotation labelled it as negative. Both CT-BERT and BERTweet labeled the tweet as positive due to the phrase "raised no safety concerns".

**Table 8**

Class-wise performance analysis of the transformer models for positive, negative and neutral sentiments (for dataset augmented using LMOTE)

| Model | Category | Precision | Recall | F1-score | MCC |
|---|---|---|---|---|---|
| BERT | neutral | 0.78 | 0.83 | 0.80 | 0.47 |
| BERT | positive | 0.72 | 0.67 | 0.69 | 0.56 |
| BERT | negative | 0.45 | 0.34 | 0.39 | 0.35 |
| XLNet | neutral | 0.80 | 0.82 | 0.81 | 0.51 |
| XLNet | positive | 0.72 | 0.73 | 0.72 | 0.59 |
| XLNet | negative | 0.48 | 0.38 | 0.42 | 0.39 |
| RoBERTa | neutral | 0.81 | 0.82 | 0.81 | 0.53 |
| RoBERTa | positive | 0.73 | 0.74 | 0.73 | 0.6 |
| RoBERTa | negative | 0.52 | 0.43 | 0.47 | 0.43 |
| CT-BERT | neutral | 0.81 | 0.83 | 0.82 | 0.49 |
| CT-BERT | positive | 0.73 | 0.71 | 0.72 | 0.56 |
| CT-BERT | negative | 0.55 | 0.47 | 0.51 | 0.46 |
| BERTweet | neutral | 0.81 | 0.84 | 0.82 | 0.54 |
| BERTweet | positive | 0.75 | 0.73 | 0.74 | 0.62 |
| BERTweet | negative | 0.54 | 0.41 | 0.47 | 0.43 |

**Table 9**

Examples of tweet classification by the pre-trained transformer models

| Tweet | ground truth | BERT | XLNet | RoBERTa | CT-BERT | BERTweet |
|---|---|---|---|---|---|---|
| i will not be taking amp j or any other vaccine. it's clear now that governments have no idea of the safety profile short medium long term of these experimental vaccines | negative | neutral | neutral | neutral | negative | negative |
| there were six deaths during the late stage trials but the fda says this raised no safety concerns | negative | neutral | neutral | neutral | positive | positive |

The following observations were made from the performance scores of the five transformer models before and after text oversampling by LMOTE (augmentation of the minority classes only).

1. The domain-specific Covid-Twitter-BERT (CT-BERT) model performed significantly better than the pre-trained models- RoBERTa, XLNet and BERT for the original dataset (Table 4) since it was pre-trained on a Twitter dataset consisting of only Covid-19 tweets. The CT-BERT results also outperformed the other domain-specific model BERTweet for the original non-augmented dataset.

2. BERTweet, which has BERT as the base model, and is trained on English Tweets, gives consistently better results, and outperformed CT-BERT for the augmented dataset. The results of CT-BERT and BERTweet were overall better than that of RoBERTa, XLNet and BERT, which is expected since they both are trained on domain-specific information (tweets related to the Covid-19 pandemic).

3. Both CT-BERT and BERTweet performed well for the minority class containing negative sentiments even without text augmentation (Table 5), indicating that pre-training with domain-specific information helps to mitigate the effects of an imbalanced class distribution.

3. RoBERTa model was proved to be a better fit for the task at hand as compared to BERT and XLNet as observed from the accuracy, F1-score, MCC scores in Tables 4 and 6.

4. Training the models using synthetically generated textual data yielded worse results for the neutral class, and a marginal increase for the positive class, while the scores of the negative class significantly increased as observed from Tables 5 and 8.

5. It was concluded that LMOTE works poorly in multi-class settings, often degrading performance for the majority classes. LMOTE was used to augment and balance the dataset, in our case, so that training can be performed equally for all classes.

6. Thus text oversampling is not an advisable choice in the case of an imbalanced small sample multi-class dataset since it can downgrade the precision and/or recall rate for the majority class, as observed from the drop in the performance scores of the neutral class in Table 8. Though oversampling does improve results in data mining, the results of oversampling of Covid-19 tweets using LMOTE and SMOTE (Tables 6 and 7) were not encouraging, since the synthetically generated text for the low population negative class was not of high quality and degraded the performance of the neutral class (majority class).

## 6. Conclusion

In this paper, we explore the effectiveness of domain-specific pre-trained transformer models and text oversampling for learning from small sample datasets with an imbalanced class distribution. We consider the specific task of sentiment analysis of Covid-19 vaccine-related tweets. The majority class is the neutral sentiment, while the positive and negative sentiments form the minority classes. The performance scores of the negative sentiment class are the lowest which occurred due to the small number of training samples in this class. In this scenario, the domain-specific pre-trained transformer models CT-BERT and BERTweet outperform RoBERTa, BERT

and XLNet transformer models that are state-of-the-art pre-trained transformer models popularly used for text classification tasks. Thus we conclude that domain-specific transformer models are able to mitigate the class imbalance to a certain extent. Text oversampling of the minority class Covid-19 tweets was found to deteriorate the overall performance, with BERTweet performing better than the other models on the augmented dataset. Hence synthetic tweet generation via text oversampling for the minority classes is not advisable for imbalanced small sample text datasets. We propose to adapt domain-specific transformer models for the classification of Covid-19 related documents in digital repositories in our future works. Since both CT-BERT and BERTweet models are based on the BERT transformer model, we would like to explore, in future, the pre-training of more recent transformer versions such as XLNet using Covid-19 related tweets.

## 7. References

[1] URL https://www.kaggle.com/gpreda/all-covid19-vaccines-tweets. Last accessed on 11th Feb 2022.

[2] Adeyemi I., Esan A.: Covid-19-Related Health Information Needs and Seeking Behavior among Lagos State Inhabitants of Nigeria. In: International Journal of Information Science and Management (IJISM), vol. 20(1), 2022.

[3] Adoma A., Henry N., Chen W.: Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In: 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), p. 117–121. IEEE, 2020.

[4] Al-Hashedi A., Al-Fuhaidi B., Mohsen A.M., Ali Y., Gamal Al-Kaf H.A., Al-Sorori W., Maqtary N.: Ensemble classifiers for Arabic sentiment analysis of social network (Twitter data) towards COVID-19-related conspiracy theories. In: Applied Computational Intelligence and Soft Computing, vol. 2022, 2022.

[5] Alenezi M.N., Alqenaei Z.M.: Machine learning in detecting COVID-19 misinformation on twitter. In: Future Internet, vol. 13(10), p. 244, 2021.

[6] Araci D.: Finbert: Financial sentiment analysis with pre-trained language models, 2019. ArXiv preprint arXiv:1908.10063.

[7] Bahdanau D., Cho K., Bengio Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR. 2015.

[8] Bansal A., Susan S., Choudhry A., Sharma A.: Covid-19 Vaccine Sentiment Analysis During Second Wave in India by Transfer Learning Using XLNet. In: International Conference on Pattern Recognition and Artificial Intelligence, pp. 443–454. Springer, 2022.

[9] Beltagy I., Lo K., Cohan A.: SciBERT: A Pretrained Language Model for Scientific Text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), p. 3615–3620. 2019.

[10] Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P.: SMOTE: synthetic minority over-sampling technique. In: Journal of artificial intelligence research, vol. 16, pp. 321–357, 2002.

[11] Dastgheib M., Koleini S., Rasti F.: The application of deep learning in persian documents sentiment analysis. In: International Journal of Information Science and Management (IJISM), vol. 18(1), p. 1–15, 2020.

[12] Devlin J., Chang M., Lee K., Toutanova K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, p. 4171–4186. 2019.

[13] Goel R., Susan S., Vashisht S., Dhanda A.: Emotion-Aware Transformer Encoder for Empathetic Dialogue Generation. In: 2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), p. 1–6. IEEE, 2021.

[14] Huang K., Altosaar J., Ranganath R.: Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2019. ArXiv preprint arXiv:1904.05342.

[15] Hutto C., Gilbert E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the international AAAI conference on web and social media, vol. 8(1), p. 216–225, 2014.

[16] Kou G., Yang P., Peng Y., Xiao F., Chen Y., Alsaadi F.: Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. In: Applied Soft Computing, vol. 86, p. 105836, 2020.

[17] Lan Z., Chen M., Goodman S., Gimpel K., Sharma P., Soricut R.: Albert: A lite bert for self-supervised learning of language representations, 2019. ArXiv preprint arXiv:1909.11942.

[18] Lee J., Hsiang J.: Patentbert: Patent classification with fine-tuning a pre-trained bert model, 2019. ArXiv preprint arXiv:1906.02124.

[19] Lee J., Yoon W., Kim S., Kim D., Kim S., So C., Kang J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. In: Bioinformatics, vol. 36(4), p. 1234–1240, 2020.

[20] Leekha M., Goswami M., Jain M.: A multi-task approach to open domain suggestion mining using language model for text over-sampling. In: European Conference on Information Retrieval, p. 223–229. Springer, Cham, 2020.

[21] Liew T., Lee C.: Examining the Utility of Social Media in COVID-19 Vaccination: Unsupervised Learning of 672,133 Twitter Posts. In: JMIR public health and surveillance, vol. 7(11), p. 29789, 2021.

[22] Liu J., Lu S., Lu C.: Exploring and Monitoring the Reasons for Hesitation with COVID-19 Vaccine Based on Social-Platform Text and Classification Algorithms. In: Healthcare, vol. 9, p. 1353. MDPI), 2021.

[23] Liu S., Liu J.: Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis. In: Vaccine, vol. 39(39), p. 5499–5505, 2021.

[24] Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Stoyanov V.: 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[25] Lu J., Plataniotis K., Venetsanopoulos A.: Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition. In: Pattern recognition letters, vol. 26(2), p. 181–191, 2005.

[26] Mallick R., Susan S., Agrawal V., Garg R., Rawal P.: Context-and sequence-aware convolutional recurrent encoder for neural machine translation. In: Proceedings of the 36th Annual ACM Symposium on Applied Computing, p. 853–856. 2021.

[27] Manguri K., Ramadhan R., Amin P.: Twitter sentiment analysis on worldwide COVID-19 outbreaks. In: Kurdistan Journal of Applied Research, p. 54–65, 2020.

[28] Marcec R., Likic R.: Using twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. In: Postgraduate Medical Journal, 2021.

[29] Martin L., Muller B., Su´arez P., Dupont Y., Romary L., De La Clergerie V., Sagot B.: CamemBERT: a Tasty French Language Model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, p. 7203–7219. 2020.

[30] M¨uller M., Salath´e M., Kummervold P.: Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter, 2020. ArXiv preprint arXiv:2005.07503.

[31] Mohsen A., Ali Y., Al-Sorori W., Maqtary N.A., Al-Fuhaidi B., Altabeeb A.M.: A performance comparison of machine learning classifiers for Covid-19 Arabic Quarantine tweets sentiment analysis. In: 2021 1st International Conference on Emerging Smart Technologies and Applications (eSmarTA), pp. 1–8. IEEE, 2021.

[32] Naseem U., Razzak I., Khushi M., Eklund P., Kim J.: COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis. In: IEEE Transactions on Computational Social Systems, vol. 8(4), p. 1003–1015, 2021.

[33] Naseem U., Razzak I., Musial K., Imran M.: Transformer based deep intelligent contextual embedding for twitter sentiment analysis. In: Future Generation Computer Systems, vol. 113, p. 58–69, 2020.

[34] Nguyen D., Vu T., Nguyen A.: BERTweet: A pre-trained language model for English Tweets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, p. 9–14. 2020.

[35] Nowak S., Chen C., Parker A., Gidengil C., Matthews L.: Comparing covariation among vaccine hesitancy and broader beliefs within Twitter and survey data. In: PloS one, vol. 15(10), p. 0239826, 2020.

[36] Olaleye T., Abayomi-Alli A., Adesemowo K., Arogundade O.T., Misra S., Kose U.: SCLAVOEM: hyper parameter optimization approach to predictive modelling of COVID-19 infodemic tweets using smote and classifier vote ensemble. In: Soft Computing, pp. 1–20, 2022.

[37] Pires T., Schlinger E., Garrette D.: How Multilingual is Multilingual BERT? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, p. 4996–5001. 2019.

[38] Saini M., Susan S.: Data augmentation of minority class with transfer learning for classification of imbalanced breast cancer dataset using inception-V3. In: Iberian Conference on Pattern Recognition and Image Analysis, p. 409–420. Springer, Cham, 2019.

[39] Sattar N., Arifuzzaman S.: COVID-19 Vaccination awareness and aftermath: Public sentiment analysis on Twitter data and vaccinated population prediction in the USA. In: Applied Sciences, vol. 11(13), p. 6128, 2021.

[40] Scheible R., Thomczyk F., Tippmann P., Jaravine V., Boeker M.: Gottbert: a pure german language model, 2020. ArXiv preprint arXiv:2012.02110.

[41] Somasundaran S.: Two-level transformer and auxiliary coherence modeling for improved text segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34(05), p. 7797–7804, 2020.

[42] Susan S., Kumar A.: The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art. In: Engineering Reports, vol. 3(4), p. 12298, 2021.

[43] Vashishtha S., Susan S.: Inferring sentiments from supervised classification of text and speech cues using fuzzy rules. In: Procedia Computer Science, vol. 167, p. 1370–1379, 2020.

[44] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Polosukhin I.: Attention is all you need. In: Advances in neural information processing systems, vol. 30, 2017.

[45] Wang T., Lu K., Chow K., Zhu Q.: COVID-19 Sensing: Negative sentiment analysis on social media in China via Bert Model. In: Ieee Access, vol. 8, p. 138162–138169, 2020.

[46] Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R., Le Q.: Xlnet: Generalized autoregressive pretraining for language understanding, 2019. Advances in neural information processing systems, 32.

## Affiliations

**Anmol Bansal**
Delhi Technological University, e-mail: anmolbansal_2k18it025@dtu.ac.in

**Arjun Choudhry**
Delhi Technological University, e-mail: arjunchoudhry_2k18it031@dtu.ac.in

**Anubhav Sharma**
Delhi Technological University, e-mail: anubhavsharma_2k18it029@dtu.ac.in

**Seba Susan**
Delhi Technological University, e-mail: seba_406@yahoo.in

**Received**: ???
**Revised**: ???
**Accepted**: ???

Materials Chemistry and Physics 291 (2022) 126720

Contents lists available at ScienceDirect

Materials Chemistry and Physics

journal homepage: www.elsevier.com/locate/matchemphys

# Adsorption of cytosine on prebiotic siliceous clay surface induced with metal dications: Relevance to origin of life

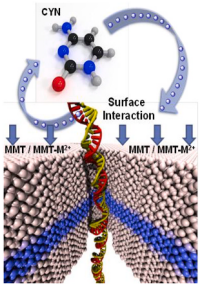
Preeti Bhatt [a,*], Chandra Kala Pant [a,**], Pramod Pandey [a,b,1], Yogesh Pandey [a], Satish Chandra Sati [a], Mohan Singh Mehata [c]

[a] Department of Chemistry, DSB Campus, Kumaun University (KU), Nainital, 263002, Uttarakhand, India
[b] State Key Laboratory for Molecular Reaction Dynamics, Dalian Institute of Chemical Physics (DICP), Chinese Academy of Sciences (CAS), 457 Zhongshan Road, Dalian, 116023, Liaoning, China
[c] Department of Applied Physics, Delhi Technological University (DTU), Bawana Road, Delhi, 110042, India

## HIGHLIGHTS

- Adsorption studies of cytosine on montmorillonite clay surface with and without incorporated metal ions ($Mg^{2+}$, $Ca^{2+}$, $Fe^{2+}$ and $Cu^{2+}$).
- The study signifies that metal ions (specifically transition metal ion) enhance the adsorption capability of clay minerals.
- Langmuir type adsorption was observed with monolayer formation.
- Present study suggests a simple pathway in support of chemical evolution on primitive Eath.
- Physico-chemical parameters might be helpful to explore the plausibility of life beyond Earth.

## GRAPHICAL ABSTRACT




## ARTICLE INFO

*Keywords:*
Adsorption
Cytosine
Montmorillonite clay
Metal ions
Langmuir adsorption parameters
Prebiotic chemical evolution

## ABSTRACT

Adsorption of protobiomolecules on primigenial mineral surfaces transacts a very specific and selective role in prebiotic chemical evolution. Cytosine (CYN) is a core constituent of genetic macromolecules i.e., DNA and RNA. A prebiotic simulation experiment following adsorption behaviour of a nucleic acid base CYN, as a function of pH, concentration, time and temperature on the surface of Montmorillonite (MMT), a prominent siliceous clay of smectite group having specific surface area, was studied systematically and methodically using UV, FTIR, SEM and XRD. Biologically significant metal dication [$M^{2+}$ = $Mg^{2+}$, $Ca^{2+}$, $Fe^{2+}$ and $Cu^{2+}$] embeded surfaces of MMT (MMT-$M^{2+}$) were prepared separately by cation-exchange approach. Results observed through UV spectral data reveal a very significant role of metal ions on the quantity of CYN adsorbed on different solid surfaces. The adsorption behaviour fitted well with Langmuir isotherm model and the adsorption isotherm shows monolayer formation of CYN (as adsorbate) on the surface of MMT and MMT-$M^{2+}$(as adsorbent) showing Langmuir type adsorption. The calculated Langmuir adsorption parameters ($K_L$& $X_m$) delineate appreciable interaction of CYN

* Corresponding author.
** Corresponding author.
*E-mail addresses:* preetibhatt00@gmail.com, preeti_bhatt@kunainital.ac.in (P. Bhatt), ckp_ku@yahoo.co.in (C.K. Pant).
[1] Current address: Applied Science Department (Chemistry), Bipin Tripathi Kumaun Institute of Technology (BTKIT), Dwarahat, Almora 263653, Uttarakhand, India.

https://doi.org/10.1016/j.matchemphys.2022.126720
Received 8 January 2022; Received in revised form 26 August 2022; Accepted 27 August 2022
Available online 9 September 2022
0254-0584/© 2022 Elsevier B.V. All rights reserved.

on MMT-$Fe^{2+}$ surface following MMT-$Cu^{2+}$ as compared to MMT alone and to other metal induced MMTs.In terms of % binding and $X_m$ data, the effectiveness of various adsorbents in adsorption of CYN on MMT/MMT-$M^{2+}$ at pH 6.6 and temperature 298K was found in the order; MMT-$Fe^{2+}$> MMT-$Cu^{2+}$> MMT > MMT-$Ca^{2+}$> MMT-$Mg^{2+}$. The results were further confirmed by SEM, XRD and FTIR which also support the outcomes. The present work throws light on the affinity of MMT clay (having multifarious applications in biological systems) with and without biogenic metal diions towards the surface interaction of nucleic acid base mimicking primeval sea shore/sea bed or hydrothermal vent conditions deciphering the phenomena that how biomonomers could have been protected from extreme hazardous conditions of primitive Earth and stabilized during chemical evolution and availed them for consecutive molecular evolution. Results further embark better adsorption capability of transition metal ion incorporated MMTs than the clay without metal ion and exchanged with alkaline Earth dications.

## 1. Introduction

Life can be seen as a result of chemical activity of biologically significant and water soluble bionanoassemblies such as proteins, nucleic acid bases, sugars and phosphate molecules. In early periods of the Earth, Life is supposed to be originated from lifeless environment through chemical evolutionary processes. Proteins and nucleic acids constitute the main matrix of living systems. How did these complex biomolecules, capable of self-replication and self-organisation, appear abiotically on primitive Earth from their monomeric building blocks is a question of paramount importance [1–3]. The fundamental pathways from non-living matter to life forming molecules, starting with the build-up of nucleobases, over the synthesis of nucleotides and finally to the formation of RNA and DNA, are still shrouded in mystery.

Bernal in 1951 speculated and stressed the potential importance of clays as solid mineral surfaces for concentration of biomolecules via adsorption [4]. Clay minerals have the capacity to concentrate, protect and catalyse the polymerization of enormous organic molecules due to their fine particle size, large specific surface area and charge characteristics [5–8]. Recently, layered clay minerals are reported to be advantageous for adsorption [9–11] and phyllosilicates containing Fe–Mg strongly bind the nucleotides by exchanging their ligands and thereby preserving the genetic information inside the inorganic templets. In succession the polymer formation needs excessive assemblance of nucleotides under favoured conditions of thermodynamics [12,13].

Montmorillonite (MMT) a prominent siliceous clay with confined metal dications especially known to have particular affinities for organic molecules by adsorbing them on and in between their silicate layers [14–16]and further stabilize them against degradation [17]. Metal induced smectite clays provide good adsorbent sites for the frequent concentration of biologically active molecules [14,18–22] and also can act as catalyst for their polymerization as it might have happened on contemporaneous primeval Earth [23–25]. *In silico* surface interaction studies of nucleobases, nucleotides and their polymers with MMT and boron nitride nanotubes were also reported [26,27] along with electronic properties of metal semiconductor structures with cytosine biopolymer [28] and *in silico*quantum mechanical studies of role of mineral surfaces in prebiotic chemical evolution [29].

An extensive contemplation of the affinities of clay minerals towards surface interaction of biomolecules has fascinated the minds of great thinkers in past few decades [30–35]. Currently, several inquisitive researchers are engaged in exploring various aspects of MMT nanoclay as a drug carrier system with multitudinous applications and its implant biocompatibility outcomes [36,37]. Among all four nucleic acid bases, cytosine (CYN), a potent biomolecule and a viable functional biogenic material [38], seems to be significant from surface interaction point of view during the course of formation of genetic macromolecules (DNA and RNA) and their monomers [39,40].Currently, experimental and theoretical study of cytosine with cations present in artificial seawater has also been proposed to emphasize the reactivity, tautomerism and role in prebiotic chemistry [41]. Therefore, CYN is taken as model for adsorption studies on the prominent smectite clay MMT, with and without entrapped metal di-cations [$Mg^{2+}$, $Ca^{2+}$, $Fe^{2+}$ and $Cu^{2+}$], which may be relevant to possible catalytic role of aluminosilicate and cation exchanged clays in abiotic formation of biomolecules, their subsequent stabilization and protection through mineralization via surface interaction and further polymerization under varied prebiotic conditions of pH, concentration, temperature and time available near lithosphere-hydrosphere boundary of primitive sea shore/sea bed or within the submarine alkaline hydrothermal vents existed in prebiotic scenario of Earth.

Solution pH is an important parameter in adsorption studies, as the net charge on mineral surface changes as a function of pH. It is generally accepted that initial oceanic pH was acidic (~4.0) in nature due to higher dissolved $CO_2$ concentration, which is further supported by A. W. Schwartz's calculations assuming if all the available $CO_2$ was dissolved in ocean during prebiotic era of Earth [42]. This acidity of ocean would have turned towards higher pH (i.e., ~5–6) after its reactions with rocks [43].Therefore, in this study a wide range of solution pH from 4.0 to 9.3 is considered. Life itself is originated from small organic molecules which are highly susceptible to degradation at higher temperatures [44], so most of the biomolecular interaction studies have been investigated at lower temperature simulating the prebiotic conditions of Earth where temperature started to decrease gradually through natural processes and thus, provide all the necessary conditions for the formation, concentration and polymerization of the available biomonomers leading the chemical evolution of life [45]. Earlier, Kamaluddin et al. [45] have shown the effect of high concentration of electrolytes such as alkalis and halides on the adsorption studies and reported that the presence of electrolytes have minimal effect in such surface interactions.

The main objective of the present work is to explore the conditions of maximum adsorption of the pyrimidine base CYN on MMT clay, and the role of biogenic divalent metal cations entrapped in MMT clay. Metal ions have been associated with biological systems for billions of years, through the interplay of biogeochemical weathering and metabolic pathways involving the products of that weathering. A number of metal ions play a vital role in a vast number of widely differing biological processes essential for the activity of enzymes [46]. All life forms have an absolute requirement for metals, as metals play critical roles in fundamental processes, including osmotic regulation, catalysis, metabolism, biomineralization, and signalling [47]. In a critical review Cleaves et al. [33] mentioned that various transition metals might have played a key catalytic role in prebiotic organic synthesis of biologically significant molecules. It is increasingly recognized that metals are involved in cellular and subcellular functions and also used as inorganic drugs for many diseases [48]. The multifaceted role of clay minerals such as feasibility of different clays with zero valent iron for nitrate reduction in groundwater [49], adsorption studies of organic compound on granular activated carbons [50], use of montmorillonite clay to sequester cationic dyes [51] and uptake and biodegradation of contaminants from aqueous phase [52]has also studied by many workers.

The present study throws light upon adsorption of CYN on prebiotically available siliceous MMT clay and its biogenic metal dications induced solid surfaces deciphering the inevitability of adsorption phenomenon, clay minerals and metal ions in successive molecular evolution in primitive era of the Earth as well as its biological implications.

## 2. Materials and methods

### 2.1. Chemicals and reagents

Analytical grade reagents were used during all experimentation. Montmorillonite (MMT K10) clay was procured from Sigma Aldrich, Germany; Cytosine (CYN) from Loba Chemie, India; $CaCl_2$ from CDH, India and $H_2O_2$, $MgCl_2$, $FeCl_2$, $CuCl_2$ from Qualigens, India. All the solutions were prepared freshly using triple distilled deionised water ($_{td}H_2O$ = pH 6.8 ± 0.2) and borosilicate glasswares were used throughout the work.

### 2.2. Preparation and purification of sensitizers/adsorbents

Preparation and purification of MMT and MMT-$M^{2+}$ i.e., metal [$Mg^{2+}$, $Ca^{2+}$, $Fe^{2+}$ and $Cu^{2+}$] entrapped MMT clay surfaces was carried out as described (Scheme 1) in our earlier work and reported elsewhere [14,18,34,53]. After drying in vacuum desiccators the adsorbents were grounded, sieved to 100 mesh size for further experimentation (Scheme I).A brief description for the preparation of stock solution and various adsorbents is also provided in supplementary information (SI 1–2).

### 2.3. Experimental procedure

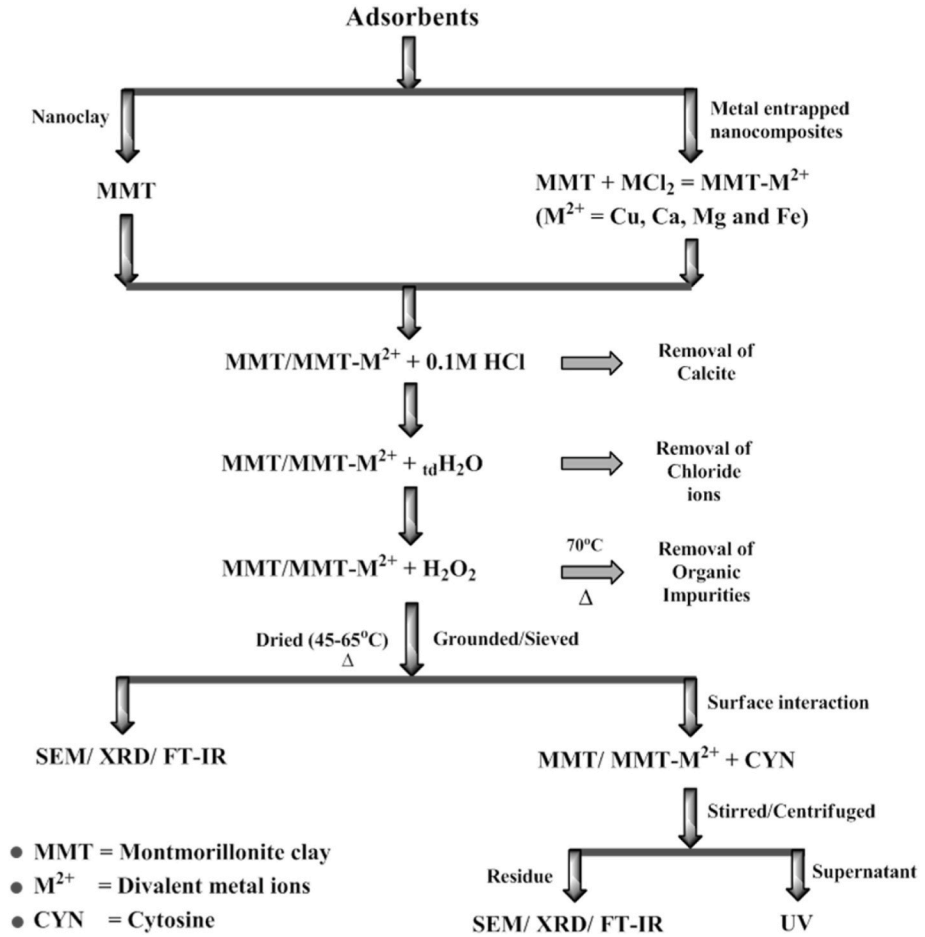
The adsorption studies of CYN (Fig. 1) separately on MMT and MMT induced with biologically significant metal dications [$Mg^{2+}$, $Ca^{2+}$, $Fe^{2+}$ and $Cu^{2+}$] were carried out in aqueous environment as a function of pH (4.0–9.3) and concentrations ($1.0 \times 10^{-5}$ M – $9.0 \times 10^{-5}$ M) of adsorbate at room temperatureunderoptimized condition of time in order to obtain saturation point. Buffer solution (100 μl each) ofacetic acid-sodium acetate (pH 4.0–6.0) or boric acid-borax (6.6–9.3)was addedto the adsorbate solution (3 ml each) containing different adsorbents MMT/MMT-$M^{2+}$ [$M^{2+}$ = $Mg^{2+}$, $Ca^{2+}$, $Fe^{2+}$, $Cu^{2+}$] (30 mg each), assuming that buffer should be a very poor ligand and unable to form stable complex with the minerals. This was confirmed by using different buffer concentration on CYN solution. Buffer solution of CYN at different pH values was added to the adsorbent in separate conical flasks (10 ml), stirred mechanically for 15 min and allowed to stand at room temperature (298K) to find out the adsorption capability of the given adsorbents.

In each experiment the adsorption of CYN was measured after a period of 8hrs (equilibrium reached after this time period). The suspensions were then centrifuged at 3500 rpm for 30min and the supernatant solutions were decanted for subsequent quantitative estimation, leaving the adsorption adduct as residue and the pH of the solution was again noted which was found to be almost unaffected. Adducts were separated, washed and then dried under vacuum at ambient temperature and used for further analytical investigations. The complete procedure was triplicated and the formation of degradation product after adsorption (if any) was checked by chromatographic techniques (PC and HPLC) [14]. Other necessary experimental conditions are provided as supplementary information (SI3)

### 2.4. Analysis and characterization

#### 2.4.1. Properties of adsorbent

The colour of MMT K10 used as adsorbent was found 10R8/1 (White) on Munsell Notations with water holding capacity (W.H.C.) 95%, pH 3.6




**Scheme 1.** Schematic presentation of various steps involved in preparation/purification of adsorbents and their characterization before and after adsorption.

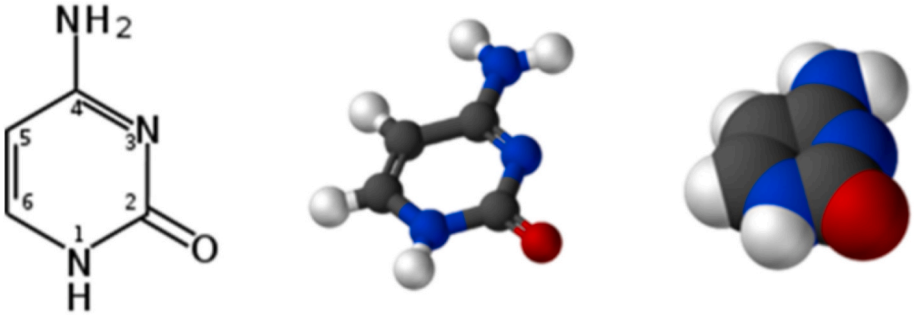



**Fig. 1.** Chemical structure (2D & 3D) of cytosine.

and cation exchange capacity (CEC) 200.17meq/100g.

*2.4.2. Absorbance measurements*

The different concentrations of CYN in supernatant were determined by recording their absorbance's at $\lambda_{max}$ = 263 nm by using Systronics double beam UV/Vis spectrophotometer 2203$^{smart}$ and quartz cuvette with 1 cm path length. pH was recorded on Hanna pH meter (model pHep) with an accuracy of ±0.1 pH unit.The amount of CYN adsorbed under optimized conditions was calculated from the difference between initial concentration and concentration after adsorption in each case. Fig. 2 depicts the optical absorbance of cytosine solution before and after adsorption. The equilibrium concentration of CYN and the amount adsorbed was used to obtain adsorption isotherms [14].

*2.4.3. Morphological observation*

Scanning electron microscopy (SEM) was used to observe the surface morphology of adsorbate, adsorbents with or without divalent metal ions and their respective adducts using a Hitachi S–3700 N instrument with an acceleration voltage 15.0 kV, take-off angle 44.1° and ZEISS EHT 20 kV.

*2.4.4. Relative intensity and inter-planar spacing identification*

X-ray diffraction (XRD) studies were performed to check the crystalline nature as well as to confirm purity and to know the relative intensity and inter-planar spacing. A Thermal X-TRA ARL diffractometer equipped with a 'Peltier' detector, using Cu Kα radiation (wavelength = 1.5406 Å) operated at 30 mA and 40 kV with graphite monochromator detection between the range 10–80° 2θ angle was used for powder diffraction analysis of adsorbents and adsorption adducts.

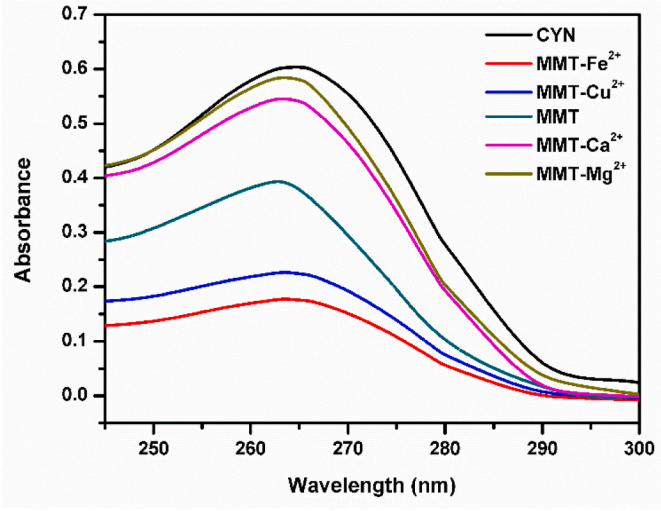



**Fig. 2.** Optical absorbance of cytosine solution before and after adsorption.

*2.4.5. Vibrational studies*

PerkinElmer DX-II FT-IR spectrophotometer equipped with DTGS detector was used to determine the IR absorption frequencies. The data were collected at 4 cm$^{-1}$ resolution and 100 scan with frequency range 400-4000 cm$^{-1}$.

## 3. Results and discussion

*3.1. Adsorption (steady state observations)*

The steady state observations are based on UV spectral data. Surface interaction studies of CYN on MMT and on MMT-M$^{2+}$ [M$^{2+}$ = Mg$^{2+}$, Ca$^{2+}$, Fe$^{2+}$, Cu$^{2+}$] have been carried out under varied conditions of pH (4.0–9.3) and concentration (10μM–90μM) at room temperature (298K) and optimized time with a view to figure out the possible affinity of the clay and its metal induced solid surfaces with CYN. Initially, based on our previous lab experimentsa widerange of pH (from acidic to basic), concentration (from different milimolar to micromolar solutions), temperature (from 20 °C to 35 °C) and time (from 1min. to 24hrs)was studied separately to figure out the best conditions for a better adsorption of the nucleobase on different solid surfaces and only those physicochemical conditions were taken for the study which showed maximum adsorption (optimized conditions) and are mentioned in the present work. The quantity of adsorbate, the amount of adsorbent and the concentration of buffer solution used in the study was also optimized after a range of variations studied separately.Blank solution was also taken for study and showed no observations. Hence, under optimized conditions 3 ml solution of adsorbate has been treated with 30 mg of adsorbent using 100 μl buffer solutions. No effect of buffer on the optical density has been observed before and after adsorption.

The maximum adsorption of CYN was observed at pH 6.6, concentration 7 μM at room temperature after 8hrs and after that saturation was attained. It may be attributed to ionic interaction between negatively charged surface of adsorbents (MMT/MMT-M$^{2+}$) and cationic and anionic forms of CYN in solution phase till all the interactive sites of clay were consumed. Results which are recorded in Tables (1 and 2) and illustrated in Figures (Figs. 2–8 and 11) have also been elaborated further in detail.

Primarily, the standard concentration curve was plotted between optical absorbance versus molar concentration of various CYN solutions monitored at their absorption maxima ($\lambda_{max}$ = 263 nm) which showed the amount of CYN adsorbed during adsorption process (Fig. 3). Linear relationship between absorbance and concentration of CYN has been observed showing $r^2$ = 0.99. Preliminary studies showed that the amount of nucleic acid base adsorbed is dependent on pH of the solutionand was found maximum at pH 6.6 (Fig. 4). As shown in Fig. 4, graph between quantity adsorbed vs pH is showing maximum quantity of cytosine adsorbed on clay was at 6.6, therefore, subsequent studies were carried out at pH 6.6 and the temperature choosen wastheroom temperature(298K).

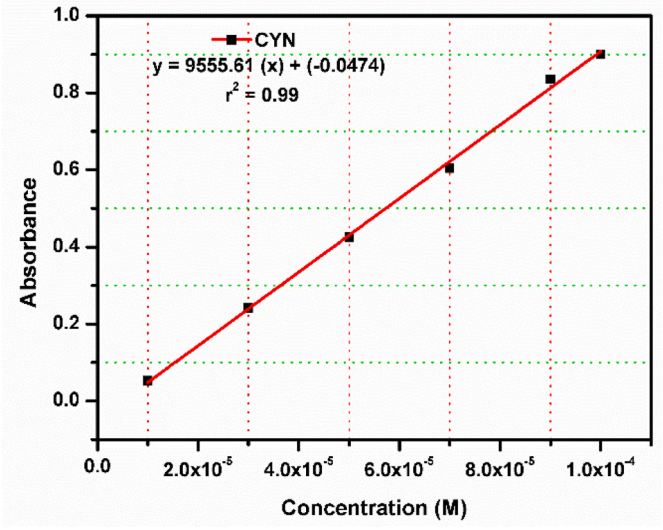



**Fig. 3.** Standard concentration curve (optical absorbance v/s molar concentration) of cytosine (CYN).

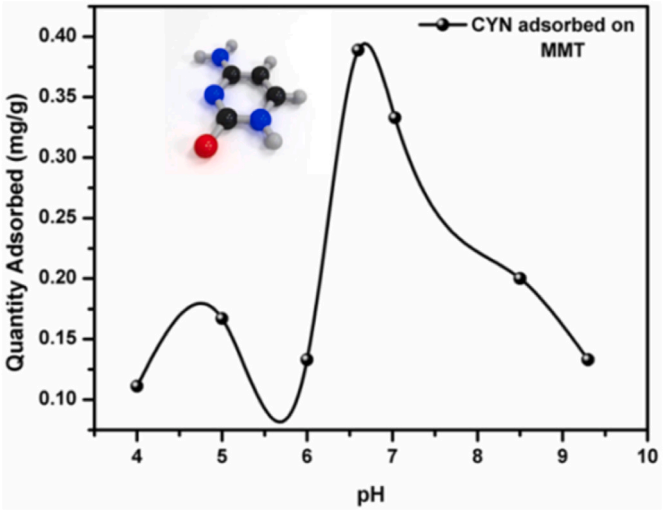



**Fig. 4.** Quantity (mg $g^{-1}$) of CYN adsorbed on MMT at different pH.

### 3.2. Adsorption isotherm

An adsorption isotherm is a relation between the extent of adsorption (i.e., the amount of adsorbate adsorbed per unit mass of an adsorbent) and the equilibrium concentration at a constant temperature.Actually, it is a curve obtained by plotting quantity adsorbed(extent of adsorption) against equilibrium molar concentration at at constant temperature.In the present work, the adsorption isotherm indicate the amount of CYN adsorbed on MMT/MMT-$M^{2+}$increases gradually as their equilibrium concentration increases in solution up to saturation point, above which adsorption become independent of adsorbate concentration and attained equilibrium thereafter (Fig. 5). Initially, plots show linear relationship where the regression coefficient, i.e., $r^2$ and the value 1.0 were taken as limit of significance. The nature of the curve is asymptotic which means approaching a value or a curve arbitrarily closely.Langmuir model and Freundlich model were used to fit the adsorption isotherm data and the model parameters were calculated which were listed in Table 1.The asymptotic nature of curve (Fig. 5) suggests Langmuir type adsorption [54] and calculated by the following equation (Eq. (1)) [55,56];

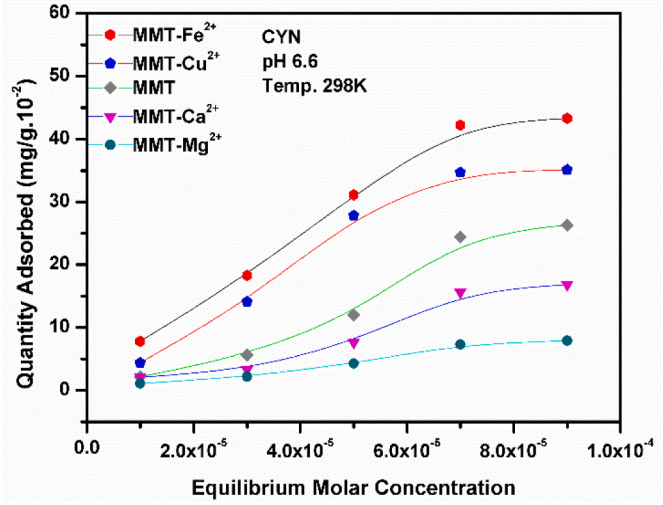



**Fig. 5.** Graph showing adsorption behavior of CYN as a function of quantity adsorbed versus equilibrium molar concentration.

$$\left|\frac{C_{eq}}{X_e}\right| = \left|\frac{1}{K_L \cdot X_m}\right| + \left|\frac{C_{eq}}{X_m}\right|. \quad (1)$$

Where, $C_{eq}$ is equilibrium concentration of CYN, $K_L$ is a constant related to enthalpy (ΔH) of adsorption ($K_L \alpha\ e^{-\Delta H/RT}$), $X_e$ the amount (mg) of solute adsorbed per gram weight of adsorbent and $X_m$ the amount (mg) of solute required per gram weight of adsorbent for complete monolayer formation. The value of $X_m$ is a measure of accessibility of adsorption sites and $K_L$ is a measure of affinity of adsorbenttowards the adsorbate. $K_L$ is a constant which will be fitted in variant of Clausius Clapeyron equation and the final data will be heat of adsorption/enthalpy of adsorption ($\Delta H_{ads}$). It is also known as adsorption capacity or sorption energy andhigh values of $K_L$ implies strong binding or the extent of interaction between adsorbate (CYN) and the surface of adsorbent MMT/MMT-$M^{2+}$ (Table 1).

The Langmuir adsorption parameters $X_m$ and $K_L$ were calculated from the slope and the intercept obtained from the graph between $C_{eq}/X_e$ vs $C_{eq}$ (Fig. 6). $X_e$ can be obtained asymptotically from Fig. 5 on extrapolating the adsorption curve towards Y-axis when saturation phenomenon occurs. It was observed that the adsorption trend (% binding and Langmuir constants) of adsorbate adsorbed on MMT with or without entrapped dications largely depends on adsorbate as well as the nature of adsorbent.

The Freundlich isotherm model [57], which is commonly represented by equation (2), wasapplied to study adsorption from solution.

$$|\log X_e| = |\log K_f| + \left|\frac{1}{n}\log C_{eq}\right|. \quad (2)$$

Where, $C_{eq}$ is the equilibrium concentration of CYN, $X_e$ (mg) is the amount of solute adsorbed per gram weight of adsorbent (difference between initial nucleic acid base concentration and concentration after the equilibration was assumed to represent the amount adsorbed on the clay), $K_f$ is the Freundlich constant or maximum adsorption capacity and n is the adsorption intensity. The parameters $K_f$ and n can be obtained from the graph of the linear relationship between log $X_e$ andlog $C_{eq}$, known as the Freundlich plot [58]. The $r^2$ values for Langmuir and Freundlich isotherm are shown in Table 1.

Experimental data found to fit well for both the isotherms showing higher $r^2$ values (~1.0) for Langmuir isotherm than Freundlich isotherm, indicating monolayer formation of CYN on MMT/MMT-$M^{2+}$ surfaces. Hence, the adsorption data of CYN adsorption on MMT/MMT-$M^{2+}$ surface fit to Langmuir isotherm better than Freundlich isotherm and the $K_L$ values are greater than zero indicates a favourable adsorption

**Table 1**
Langmuir and Freundlich adsorption parameters and (%) binding of CYN on the surface of various adsorbents.

| Adsorbents | Conc. | pH | Temp. | T i m e | % Binding | Langmuir Isotherm Parameters | | | Freundlich Isotherm Parameters | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $X_m$ ($mgg^{-1}$) | $K_L$ ($Lmg^{-1}$) | $r^2$ | $K_f$ | N | $r^2$ |
| MMT-$Fe^{2+}$ | 1μM–9μM | 6.6 ± 0.1 | 298K | 8 h | 67 | $0.47 \times 10^{-3}$ | $21.27 \times 10^{-5}$ | 1 | 0.0258 | $6.58 \times 10^{-6}$ | 0.97 |
| MMT-$Cu^{2+}$ | 1μM–9μM | 6.6 ± 0.1 | 298K | 8 h | 60 | $0.42 \times 10^{-3}$ | $47.61 \times 10^{-5}$ | 1 | 0.0217 | $6.47 \times 10^{-6}$ | 0.96 |
| MMT | 1μM–9μM | 6.6 ± 0.1 | 298K | 8 h | 31 | $0.22 \times 10^{-3}$ | $15.15 \times 10^{-5}$ | 1 | 0.0106 | $6.36 \times 10^{-6}$ | 0.93 |
| MMT-$Ca^{2+}$ | 1μM–9μM | 6.6 ± 0.1 | 298K | 8 h | 20 | $0.14 \times 10^{-3}$ | $17.85 \times 10^{-5}$ | 1 | 0.0037 | $5.71 \times 10^{-6}$ | 0.93 |
| MMT-$Mg^{2+}$ | 1μM–9μM | 6.6 ± 0.1 | 298K | 8 h | 04 | $0.29 \times 10^{-6}$ | $61.80 \times 10^{-6}$ | 1 | 0.0023 | $6.34 \times 10^{-6}$ | 0.93 |

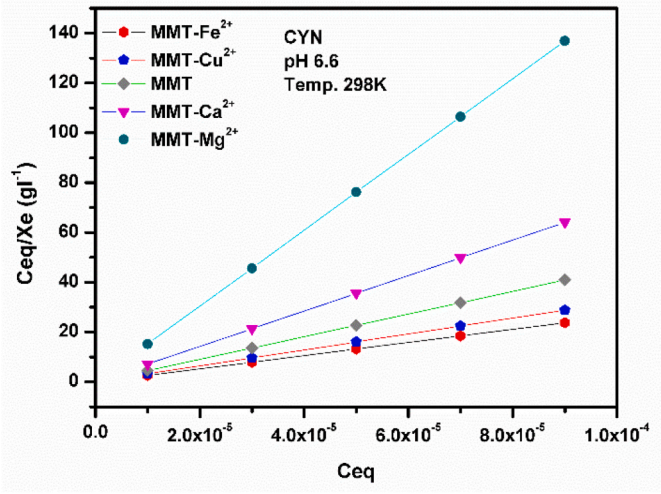



**Fig. 6.** Graph showing Langmuir adsorption isotherm of CYN for different adsorbents.

and the adsorption process fit a monolayer adsorption which is consistent with the Langmuir model. The adsorption capacity depends on the number of active sites on the solid surfaces.

### 3.3. Percent binding

Percent (%) binding referred to % adsorption and also called as sorption efficiency. This is the value of percent adsorbate adsorbed by the adsorbent. The percent binding i.e., the amount of adsorbate adsorbedafter adsorption was calculated using the following equation (Equation (3));

$$\%Binding = \left|\frac{C_{be} - C_{af}}{C_{be}}\right| \times 100. \quad (3)$$

Where, $C_{be}$ and $C_{af}$ are concentration before and after adsorption respectively. Table 1 shows the isotherm constants and % binding for adsorption of CYN on MMT and on different MMT-$M^{2+}$mineral matrices at pH 6.6 and temperature 298K.

In terms of % binding and $X_m$ data, the effectiveness of various adsorbents in adsorption of CYN on MMT/MMT-$M^{2+}$ was found in order;

MMT-$Fe^{2+}$> MMT-$Cu^{2+}$> MMT > MMT-$Ca^{2+}$> MMT-$Mg^{2+}$

The above results embarks better adsorption capability of transition metal ion incorporated MMT than the clay without metal ion and/or with alkaline earth dications. The maximum adsorption of CYN on MMT-$Fe^{2+}$following MMT-$Cu^{2+}$may be due to better coordination properties of transition metal ions having vacant d-orbital which tend to form π bonds and hence form stable complex with active sites of cytosine molecules as ligands. The $Ca^{2+}$ and $Mg^{2+}$ ions are known to present in MMT clay so it may not affect much the percentage binding capacity of clay when exchanged with similar metal cations. The exchanged cations might have been replaced the identical ions already present in the structure of MMT or they may have embedded or occupied other sites of clay. As density of $Ca^{2+}$ is low which prevent it to enter the interlayer regions but form a bridge at the edges of MMT through ionic interactions while $Mg^{2+}$ may replace the Al ion present in the clay leaving the extra cation and thus influence the affinity after intercalation reducing the adsorption capability of MMT clay.

Addition of transition metal dications, increases the adsorption of nucleic acid base throughout entire concentration range of the isotherm showing that they first get involved in complex formation with CYN and then the nucleic acid base gets adsorbed on clay surfaces or the repulsive force between the adsorbate and adsorbent or both of these mechanism may operate.

CYN can exist in two tautomeric forms (Fig. 7), in which the amino form being stable and in less energetic state predominates in solution phase [59]. Protonation in a molecule is known to increase the number of possible pairing modes leading to structural diversity [60]. In CYN protonation takes place at N3 position because of the maximum electron density. The amino group of CYN is hydrogen donor, and the ring nitrogen atom N3 is hydrogen acceptor. CYN also has a carbonyl group at C2 position, thus can form three hydrogen bonds with guanine. The hydrogen-bonding patterns of bases have important consequences for the three-dimensional structure of nucleic acids. Further, deamination of CYN can lead to the formation of uracil, which pairs with adenine(A) and implement RNA replication or 5-methyl cytosine can be converted directly to thymine(T) and thus facilitates normal biological traits in living systems.

Adsorption is also influenced by the nature of exchangeable metal cations intercalated in montmorillonite [8].In metal cations, higher oxidation state forms more stable complex with ligand than lower oxidation states and the stability of metal complex increases with decrease in size of the dication. The interactive forces between opposite charges also affect the surface interaction. Besides interactive forces, the size of metal ions may have remarkable effect on adsorption, which has also been observed from the above results.The sorption efficiency or the percent adsorption of CYN on different metal exchanged MMT surfaces has been represented inFig. 8 which shows that transition metal incorporated clay surfaces show better adsorption capacity than alkali Earth

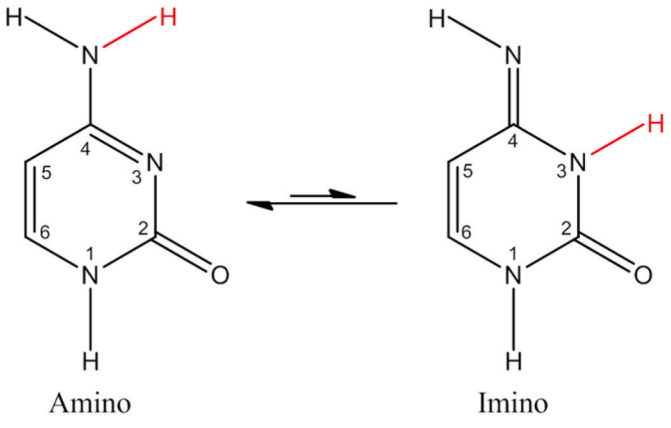



**Fig. 7.** Amino and iminoTautomers of CYN.

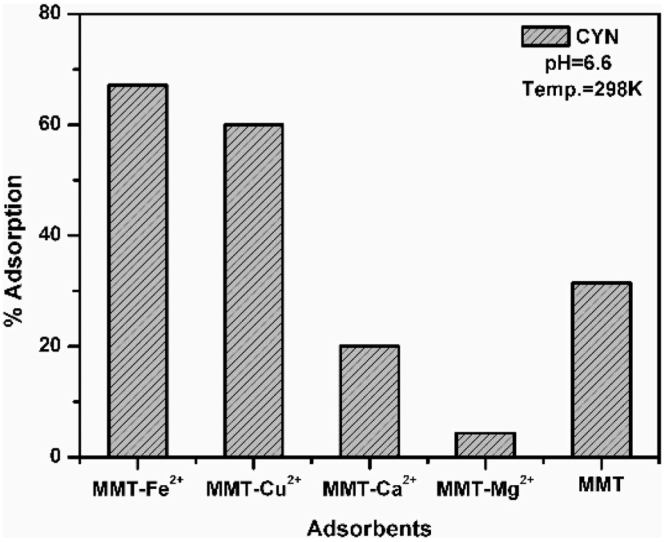



**Fig. 8.** Affinity of CYN towards different adsorbents.

meal incorporated clays.

In normal conditions, in the structure of MMT clay i.e., $[(Al–Mg)_2OH_2]^{1-}.[(Na–Ca)nH_2O]^{1-}$, the lattice silica and alumina layers appear to be sandwiched together with some vacant sites. This observation shows that incorporated metal ions increases the ionic strength of MMT surface and thus lowers the repulsive forces between organic molecule and clay surfaces.The above experimental results shows that adsorption of organic molecule from dilute aqueous solution on clay surfaces depends on the chemistry of adsorbing molecules (i.e.,interplay between electrostatic and non-electrostatic interactions based on characteristics of adsorbate and adsorbent) and the nature of adsorbent (having larger surface area) as well as thenature of adsorbate (higher the concentration, greater the extent of adsorption, provided that the temperature remains constant) which are affected by the physico-chemical parameters such as pH, temperature, concentration, ionic strength and presence of inorganic ions [33].

### 3.4. Physico-chemical characterization of adduct

Results of steady state adsorption isotherm studies and Langmuir adsorption parameters revealed MMT- $Fe^{2+}$ as the most effective adsorbent as compared to other metal induced [$M^{2+}$ = $Mg^{2+}$, $Ca^{2+}$, $Fe^{2+}$, $Cu^{2+}$] MMT clays for adsorption of CYN. Therefore, for further spectrophotometric characterization,the adduct with $Fe^{2+}$ incorporated MMT clay was chosen exclusively.

#### 3.4.1. SEM observations

Scanning electron microscopy (SEM) has been extensively used to study the surface morphology, surface topography and microstructural aspects near to surface only.The morphologies of CYN, MMT, $Fe^{2+}$modified MMT and CYN adsorbedMMT-$Fe^{2+}$ are shown in Fig. 9 in different sizes (5 μM, 10 μM, 20 μM). The morphology of CYN showed only rough and irregular forms (Fig. 9a). The characteristic morphology of MMT showed schistose structure and contain pores (Fig. 9b). However, intercalation of $Fe^{2+}$ ions in MMT led to irregular schistose structure with numerous cavities (Fig. 9c) and provide large surface area which was convenient for CYN adsorption into pore spaces resulted in an increase in the adsorption capacity of MMT-$Fe^{2+}$. The CYN adsorbed MMT-$Fe^{2+}$ (Fig. 9d) shows slight change in the typical surface morphology.

#### 3.4.2. XRD pattern

X-ray differection (XRD) is used for primary characterization of material properties viz., crystallographic structure, crystallite size and strain. The crystal morphology of clay was confirmed by X-ray diffraction analysis.The results of diffraction patterns of CYN, MMT, MMT-$Fe^{2+}$, CYN + MMT-$Fe^{2+}$ were given in Fig. 10, wherethe (001) crystal surface belongs to MMT and the higher affinity of MMT-$Fe^{2+}$ interlayer was observed for CYN.The XRD data of the adsorbate, adsorbent and adsorption adduct (Table 2) suggests slight increase in the basal spacing of the adsorbent after CYN adsorption.The crystallographic spacing (d)

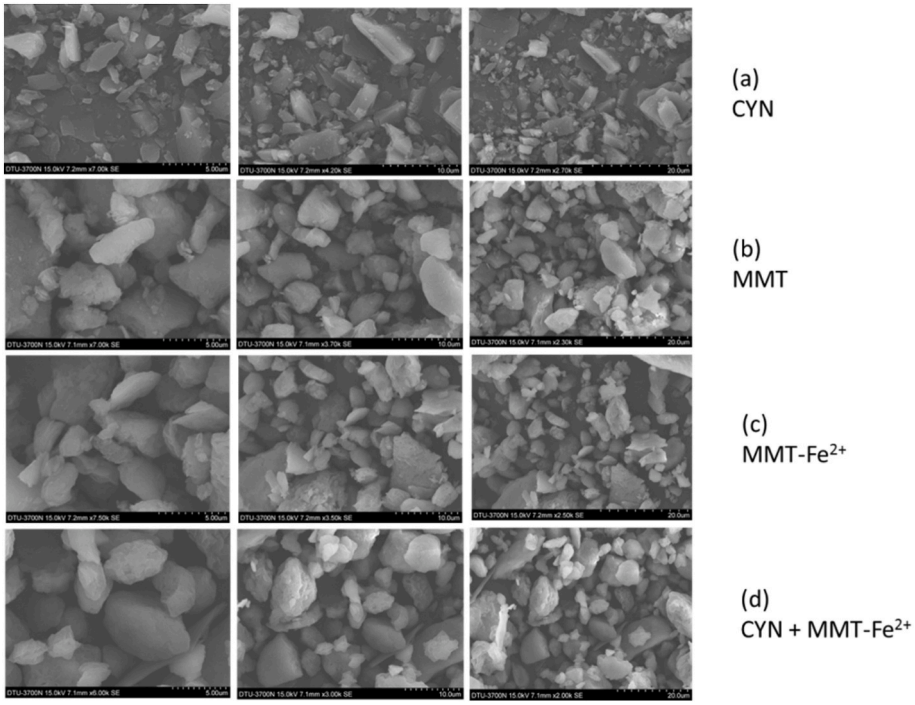



**Fig. 9.** Different SEM images of (a) CYN (5 μM, 10 μM, 20 μM), (b) MMT (5 μM, 10 μM, 20 μM), (c) MMT-$Fe^{2+}$(5 μM, 10 μM, 20 μM), and (d) adsorption adduct of MMT-$Fe^{2+}$ + CYN (5 μM, 10 μM, 20 μM) after adsorption.

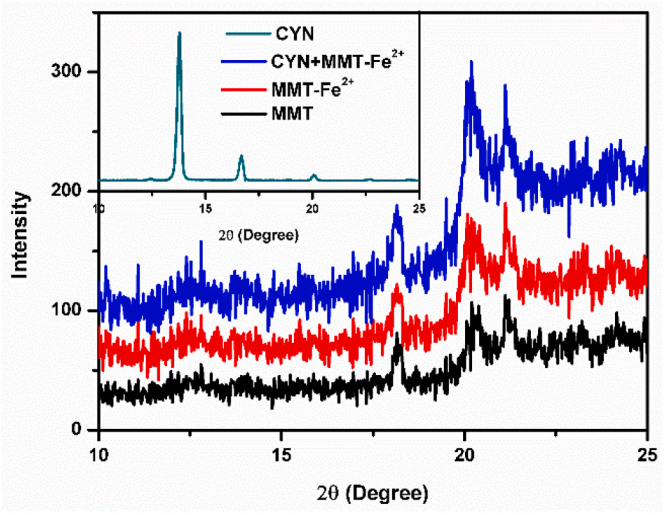



**Fig. 10.** XRD graph of CYN, MMT, MMT-$Fe^{2+}$ and CYN + MMT-$Fe^{2+}$ adduct after adsorption.

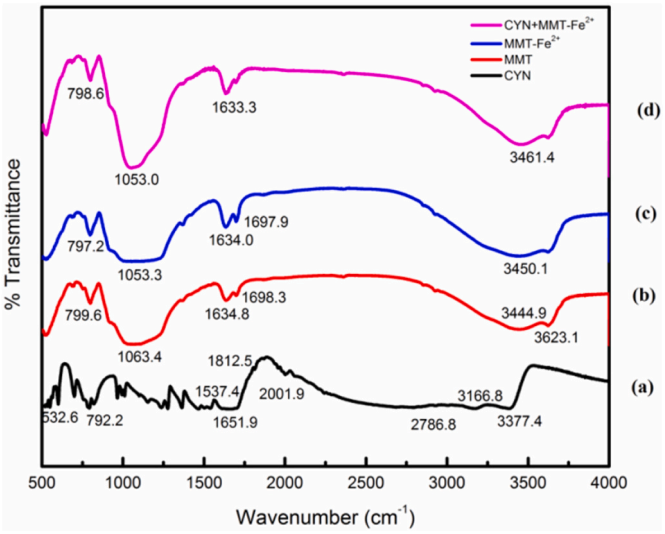



**Fig. 11.** FT-IR spectra of (a) CYN, (b) MMT, (c) MMT-$Fe^{2+}$ and (d) adsorption adduct, CYN + MMT-$Fe^{2+}$ after adsorption.

**Table 2**
XRD data showing 2θ and basal spacing (Å) for adsorbent and adsorption adduct.

| S/N | Probe | 2θ | d (Å) |
|---|---|---|---|
| 1. | CYN | 20.22 | 4.4577 |
| 2. | MMT | 20.2 | 4.4629 |
| 3. | MMT-$Fe^{2+}$ | 20.08 | 4.4876 |
| 4. | MMT-$Fe^{2+}$-CYN | 19.74 | 4.6052 |

was calculated by Bragg's equation (Equation (4)).

$$2d \sin\theta = n\lambda \quad (4)$$

Where, d is the spacing between differection lattice planes, θ is the measured differection angle, λ is the wavelength of X-ray radiation and n = 1, The interlayer space of MMT modified with $Fe^{2+}$ showed a slight increase from 4.4629 Å to 4.4876 Å and after CYN adsorption it further increases to 4.6052 Å. These observations indicates that after exchange with metal cation, the binding force of MMT layers weaken, meanwhile, make it form active metal site which is beneficial to adsorb CYN. All three peaks (Fig. 9) reveal no major differences, indicating that the $Fe^{2+}$ exchanged MMT and $Fe^{2+}$exchanged MMTadsorbed CYN will not destroy the original crystal structure of MMT.

Maximum affinity of CYN with MMT-$Fe^{2+}$may be due to complex formation with inorganic divalent metal ion (CYN–$Fe^{2+}$). However, $Fe^{2+}$ ions could have been oxidised into $Fe^{3+}$ ions after adsorbing inside the clay layers as suggested earlier by Zaia et al. [61]. CYN may interact with different sites of the clay surface, i.e., the interlayer region and the faces of the lamella.The expansion of the interlayer was greater may be due to the inclusion of CYN. After saturating edges of adsorbate the adsorbent might have entered the interlayer region and influenced the basal spacing.

### 3.4.3. FTIR spectra

The above results were complemented by vibrational data, where the possible interactive sites of CYN on MMT-$Fe^{2+}$ were investigated by FTIR spectra (Fig. 11). The absorption frequencies lying in between 1600 and 1800 $cm^{-1}$ are important for sensitivity of molecular interactions between clays and nucleobases, as pyrimidines absorb strongly in this region [62]. The spectra of CYN set forth a large number of vibrational frequencies (Fig. 11a), where the IR bands at 3377.4 $cm^{-1}$ and 1651.9 $cm^{-1}$ corresponds to N–H stretching and C=O stretching respectively [31,63], while band at 1537.4 $cm^{-1}$ points out N–H bending. Bands at 792.2 $cm^{-1}$ and 532.6 $cm^{-1}$ could be attributed to C–H and OCN deformation.

In the IR spectrum of MMT the band at 3623.1 $cm^{-1}$ coincides –OH stretching and bands at 3444.9 $cm^{-1}$ and 1634.8 $cm^{-1}$ could be due to stretching and bending vibrations of water adsorbed in MMT, whereas the band at 1063.4 $cm^{-1}$ adapts Si–O–Si stretching vibration [13] (Fig. 11b). After encapsulation of $Fe^{2+}$ ions in MMT a shift displayed in frequencies from 3444.9 $cm^{-1}$ to 3450.1 $cm^{-1}$ and 1063.4 $cm^{-1}$ to 1053.3 $cm^{-1}$ respectively, while the band at 3623.1 $cm^{-1}$ disappeared, revealing interaction of $Fe^{2+}$ with MMT (Fig. 11c). This interaction may be due to the attractive forces between opposite charges of $Fe^{2+}$ and the lone pair of electrons on oxygen or negative charges on silica layers binding alumina layer in between. For MMT, the negative charges are mainly coming from substitutions in the octahedral Al layer. The IR spectrum manifests a complete sweep out of number of vibrational frequencies in the finger print region of CYN and a remarkable shift in various absorption frequencies from 3377.4 $cm^{-1}$ to 3461.4 $cm^{-1}$, 1651.9 $cm^{-1}$ to 1633.6 $cm^{-1}$, 792.2 $cm^{-1}$ to 798.6 $cm^{-1}$, 532.6 $cm^{-1}$ to 526.9 $cm^{-1}$ and 440.9 $cm^{-1}$ to 466.6 $cm^{-1}$ after adsorption of CYN on MMT-$Fe^{2+}$ (Fig. 11d) explicitly indicate the interaction of CYN to MMT-$Fe^{2+}$ particularly via N–H and C=O bonds. A slight shift of band to higher frequencies indicates the alteration of structure and the chemical activation of CYN to MMT-$Fe^{2+}$. These consequences are in perfect accordance with the findings reported earlier [62,64].

The adsorption of CYN most likely be proceed via N–H and C=O group, either hydrogen bonded to the structural oxygen or as such tightly fixed with silica structure. The carboxyl groups are generally fixed with positively charged Al-oxy-hydroxy element by means of ionic bridges. Surface of MMT is entirely comprised of silica and in the interlayer region where oxygen confront each other, metal cations and polar molecules located cause lattice expansion. MMT is thus capable of forming intercalation with the probe which forms the basis of adsorption property of clays.

Since metal ions and nucleotidesare key components of various enzymes and most of the nucleotides are coenzymes, which can catalyse a number of biochemical reactions in the living systems [1,46], the present study give better understanding regarding the pathways of accumulation of biomonomers on solid surfaces through surface interaction which on aggregation themselves might have catalysed the formation of complex biomolecules via polymerization and condensation in enzyme free environment of early Earth and may influencesthe property of primigenial material being compatible and adaptable during the transition phase from non-living to living [39]. Due to the abundance of iron on

Earth and its role in biological systems, the transition and mineralogical stages of iron have played a key role in Earth surface systems and in marine geochemistry [65].

Moreover, the characterization of the spectroscopic features of biomolecule-mineral complexes also provides a support in remote sensing spectroscopy for detecting organic compounds on planetary surfaces or cometary grains and asteroid surfaces [66]. Life originated from lifeless environment through chemical evolution of complex life forming molecules and nucleobases might have played a critical role at the dawn of life due to their photoprotective properties. It is a long time span phenomena and without being protected and stabilized the evolutionary processes might not proceed further. Henceforth, upon adsorption on mineral matrices CYN could be stabilized and protected from hazardous electromagnetic radiations directly coming from Sun in ozone free primitive environment and thus become available for successive molecular evolution. In support, metal embedded clay minerals play a key role in this scenario by promoting the ability of adsorption.

Adsorption parameters and percent binding data endorse that CYN was adsorbed appreciably on MMT-$Fe^{2+}$ followed by MMT-$Cu^{2+}$ as compared to MMT alone and with $Ca^{2+}$ and $Mg^{2+}$. FTIR spectra emphasize that CYN interacted with $Fe^{2+}$ encapsulated MMT through electrostatic forces via N–H and C=O bonds. XRD data insinuated that CYN interact with different sites of the clay surface, i.e., the interlayer region and the faces of the lamella showing expansion of the interlayer was greater due to the inclusion of CYN on MMT-$Fe^{2+}$. Results demonstrate the higher affinity of MMT-$Fe^{2+}$ interlayer for CYN. SEM observations showed the surface morphology of biomonomers on $Fe^{2+}$ modifiedMMT surface after adsorption.

As metal ions in the form of metalloenzymes are very good reflectors of biogenic activities in living systems, the results substantiate the efficacy of biogenic divalent metal ions particularly transition metal ions in prebiological adsorption processes through which concentration and condensation of biomonomers might have facilitated the formation of complex biopolymers. Moreover, during the formation of planet Earth iron was the most abundant element which might have contributed a lot in chemical evolution and origin of life [65]. Our present results are also in good agreement with this scenario. Iron is also the most abundant multivalent element in planetary reservoirs like Mars, which may record signatures of processes, occurred during the formation and subsequent differentiation of the terrestrial planets [34,67], focuses on extinct or extant life on Mars [68]. Biological informational polymers such as nucleic acids have the potential to provide unambiguous evidence of life beyond Earth [69].

So, from above results it can be concluded that metal ions enhance the adsorption capability of clay minerals. In a wider perspective, studies related to surface interaction of minerals and biologically significant molecules provide appreciable support regarding prebiotic chemical evolution of complex biomolecules. The physico-chemical parameters of the study also decipher several aspects of MMT clay as drug carrier system with multifarious applications in biological assay. The work provides insight into the prebiotic biogeochemical as well as biological processes such as biomineralization, biomedical implants, primigenial chemical evolution implicated to origin of first life [33,62,70]. The interstellar bodies also evidently endue protobiomolecules viz. amino acids, nucleic acid bases, peptides, nucleotides, lipids and extracellular polymeric substances, firmly allied with inorganic templets [71,72]. The buffer using boric acid also has a potential catalytic role in the prebiotic peptide and nucleic acid synthesis and is an important component of sea water [73] and modified MMT act as an excellent adsorbent [74]. Therefore, surface interaction studies of such precious molecules on clay minerals are interesting and motivating to explore the extraterrestrial genesis and plausibility of life beyond Earth.

## 4. Conclusions

Present prebiotic simulation experiment signifies the importance of transition metal ions, having vacant d-orbitals, when embedded with MMT provide good adsorption sites which in turn could have played an important role for the protection of biomolecules and stabilization during prebiotic chemical evolution of life through adsorption. Adsorption parameters and percent binding data endorse that CYN was adsorbed appreciably on MMT-$Fe^{2+}$ followed by MMT-$Cu^{2+}$ as compared to MMT alone its $Ca^{2+}$ and $Mg^{2+}$exchanged MMT. FTIR spectra emphasize that CYN interacted with $Fe^{2+}$ encapsulated MMT through electrostatic forces via N–H and C=O bonds. XRD data insinuated that CYN interact with different sites of the clay surface, i.e., the interlayer region and the faces of the lamella showing expansion of the interlayer was greater due to the inclusion of CYN.

According to the previous adsorption studies, the ratio of purine-pyrimidine bases adsorbed on clays ranged from ~4.6 to 25.1, and the ratio of their occurrence in organisms ranges from 0.95 to 1.05 or 1.0 to 1.2. The weaker adsorption of pyrimidine bases on clays raises the question of possibility of genetic code with one nucleic acid base [31]. Further, this study indicates towards the possible role of biogenic metal incorporated clays to improve the adsorption capability of pyrimidine base for achieving nearby ratio of purine-pymidine base of nucleic acids in the event of chemical evolution of life in planet Earth.

## Funding

This study was funded by EMR-I, CSIR-HRDG group (Sanction No/F. No. 09/428(0075)2K17 and ACK No. 121412/2K15/1 dated 31-3-2017), CSIR Complex, Library avenue, Pusa, New Delhi 110,012, India through CSIR-SRF (Council of Scientific & Industrial Research-Senior Research Fellow) award.

## CRediT authorship contribution statement

**Preeti Bhatt:** Conceptualization, Methodology, and, Project administration, Writing – original draft, Investigation, Visualization, and, Funding acquisition, Software, Writing – review & editing, and, Formal analysis, Validation, Data curation, Resources, Supervision. **Chandra Kala Pant:** Conceptualization, Methodology, and, Project administration, Writing – review & editing, and, Formal analysis, Validation, Supervision. **Pramod Pandey:** Conceptualization, Methodology, and, Project administration, Writing – review & editing, and, Formal analysis, Validation, Data curation. **Yogesh Pandey:** Validation, Resources. **Satish Chandra Sati:** Validation, Data curation, Resources. **Mohan Singh Mehata:** Software, Writing – review & editing, and, Formal analysis, Validation, Data curation, Resources, Supervision, All authors contributed equally to the conception and design of the study and also commented on previous versions of the manuscript, All authors validated, read and approved the final version of the manuscript.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Preeti Bhatt reports financial support was provided by Council of Scientific and Industrial Research Human Resource Development Group. Preeti Bhatt reports a relationship with CSTR-HRDG that includes: funding grants.

## Data availability

Data will be made available on request.

## Acknowledgments

Authors are grateful to the Head of Chemistry Department, DSB Campus, Kumaun University, Nainital 263002, Uttarakhand, India, to endow requisite research facilities and also thankful to the Department

of Applied Physics, Delhi Technological University, New Delhi 110042, India, for required analysis of samples. Author P. B. is extremely thankful to EMR-I, CSIR-HRDG group (Sanction No/F.No. 09/428(0075) 2K17 and ACK No. 121412/2K15/1 dated 31-3-2017), CSIR Complex, Library avenue, Pusa, New Delhi 110012, India for providing financial assistance through CSIR-SRF (Council of Scientific & Industrial Research-Senior Research Fellow) award.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.matchemphys.2022.126720.

## References

[1] N. Kitadai, S. Maruyama, Origins of building blocks of life: a review, Geosci. Front. 9 (2018) 1117–1153.
[2] M.S. Dodd, D. Papineau, T. Grenne, J.F. Slack, M. Rittner, F. Pirajno, J. O'Neil, C.T. S. Little, Evidence for early life in Earth's oldest hydrothermal vent precipitates, Nature 543 (7643) (2017) 60–64.
[3] M. Franchi, E. Gallori, Origin, persistence and biological activity of genetic material in prebiotic habitats, Orig. Life Evol. Biosph. 34 (2004) 133–141.
[4] J.D. Bernal, The Physical Basis of Life, Routledge and Kegan Paul, London, UK, 1951, p. 364.
[5] J.P. Ferris, Mineral catalysis and prebiotic synthesis: montmorillonite-catalyzed formation of RNA, Elements 1 (2005) 145–149, 3.
[6] L. Sciascia, M.L.T. Liveri, M. Merli, Kinetic and equilibrium studies for the adsorption of nucleic acid bases onto K10 montmorillonite, Appl. Clay Sci. 53 (4) (2011) 657–668.
[7] W.H. Yu, N. Li, D.S. Tong, C.H. Zhou, C. Xiang (Cynthia) Lin, C.Y. Xu, Adsorption of proteins and nucleic acids on clay minerals and their interactions: a review, Appl. Clay Sci. 80–81 (2013) 443–452.
[8] H. Hashizume, Adsorption of nucleic acid bases, ribose, and phosphate by some clay minerals, Life 5 (2015) 637–650.
[9] B.K.G. Theng, The Chemistry of Clay-Organic Reactions, Adam Hilger, London, UK, 1974, p. 343.
[10] C. Aguzzi, P. Cerezo, C. Viseras, C. Caramella, Use of clays as drug delivery systems: possibilities and limitations, Appl. Clay Sci. 36 (2007) 22–36.
[11] B. Grégoire, V. Erastova, D.L. Geatches, S.J. Clark, H.C. Greenwell, D.G. Fraser, Insights into the behaviour of biomolecules on the early Earth: the concentration of aspartate by layered double hydroxide minerals, Geochimicaet Cosmochimica Acta 176 (2016) 239–258.
[12] D.G. Odom, M. Rao, J.G. Lawless, J. Oro, Association of nucleotides with homoionic clays, J. Mol. Evol. 12 (1979) 365–367.
[13] U. Pedreira-Segade, C. Feuillie, M. Pelletier, L.J. Michot, I. Daniel, Adsorption of nucleotides onto ferromagnesian phyllosilicates: significance for the origin of life, Geochimicaet Cosmochimica Acta 176 (2016) 81–95.
[14] P. Pandey, C.K. Pant, K. Gururani, P. Arora, N. Pandey, P. Bhatt, Y. Sharma, J. S. Negi, M.S. Mehata, Affinity of smectite and divalent metal ions ($Mg^{2+}$, $Ca^{2+}$, $Cu^{2+}$) with L-leucine: an experimental and theoretical approach relevant to astrobiology, Orig. Life Evol. Biosph. 45 (2015) 411–426.
[15] P.C. Joshi, M.F. Aldersley, J.W. Delano, J.P. Ferris, Mechanism of montmroillonite catalysis in the formation of RNA oligomers, J. Am. Chem. Soc. 131 (2009) 13369–13374.
[16] E.M. Pecini, M.J. Avena, Measuring the isoelectric point of the edges of clay mineral particles: the case of montmorillonite, Langmuir 29 (2013) 14926–14934.
[17] B.K.G. Theng, Clay-polymer interactions: summary and perspective, Clay Clay Miner. 30 (1) (1982) 1–10.
[18] S. Kalra, C.K. Pant, H.D. Pathak, M.S. Mehata, Studies on the adsorption of peptides of glycine/alanine on montmorillonite clay with or without co-ordinated divalent cations, Colloids Surf. A Physicochem. Eng. Asp. 212 (2003) 43–50.
[19] C.K. Pant, H. Lata, H.D. Pathak, M.S. Mehata, Heat-initiated prebiotic formation of peptides from glycine/aspartic acid and glycine/valine in aqueous environment and clay suspension, Int. J. Astrobiol. 8 (2) (2009) 107–115.
[20] K. Gururani, C.K. Pant, N. Pandey, P. Pandey, Heat induced formation of peptides from reaction mixture of glycine-glutamic acid and glycine-leucine in presence and absence of montmorillonite clay with or without metal ions under wetting drying cycles of primitive Earth, Int. J. Sci. Techno. Res. 1 (8) (2012) 159–163.
[21] K. Gururani, C.K. Pant, P. Pandey, N. Pandey, Heat induced synthesis of amino acids from reaction system comprised of acetylene, ammonia and water vapour in presence and absence of metal oxides (silica and alumina) under wetting drying conditions of primitive Earth, Asian J. Res. Chem. 5 (9) (2012) 1150–1155.
[22] P. Pandey, P. Bhatt, J. Yan, L. Peng, M.S. Mehata, C.K. Pant, Affinity of metal ions incorporated bentonite for enantiomeric interactions with alanine, Chem (2017), https://doi.org/10.26434/chemrxiv.5383627.v1.
[23] M. Paecht-Horowitz, The influence of various cations on the catalytic properties of clays, J. Mol. Evol. 11 (2) (1978) 101–107.
[24] M. Franchi, J.P. Ferris, E. Gallori, Cations as mediators of the adsorption of nucleic acids on clay surfaces in prebiotic environments, Orig, Life Evol. Biosph. 33 (2003) 1–16.
[25] L. Belmonte, S.S. Mansy, Metal catalysts and the origin of life, Elements 12 (2016) 413–418, 6.
[26] D.C. Mathew, Z. Luthey-Schulten, Influence of montmorillonite on nucleotide oligomerization reactions: a molecular dynamics study, Orig. Life Evol. Biosph. 40 (2010) 303–317.
[27] S. Mukhopadhyay, S. Gowtham, R.H. Scheicher, R. Pandey, S.P. Karna, Theoretical study of physisorption of nucleobases on boron nitride nanotubes: a new class of hybrid nano-biomaterials, Nanotechnology 21 (2010), 165703, 6.
[28] Ö. Güllü, A. Türüt, Electronic properties of Cu/n-InP metal semiconductor structures with cytosine biopolymer, ActaPhysicaPolonica A 128 (2015) 383–388.
[29] A. Rimola, M. Sodupe, P. Ugliengo, Role of mineral surfaces in prebiotic chemical evolution, 1, in: Silico Quantum Mechanical Studies. Life(Basel), vol. 9, 2019, https://doi.org/10.3390/life9010010. pii:E10.
[30] Lailach, G.E., Thompson, T.D., &Brindley, G.W. (1968a&1968b). Adsorption of pyrimidines, purines, and nucleosides by Li-, Na-, Mg-, and Ca-montmorillonite (Clay-Organic Studies XII) clays. Clay Miner.., 16, 285–293.
[31] L.O.B. Benetoli, H. de Santana, C.T.B.V. Zaia, D.A.M. Zaia, Adsorption of nucleic acid bases on clays: an investigation using Langmuir and Freundlich isotherms and FT-IR spectroscopy, Monatsh. Chem. 139 (2008) 753–761.
[32] H. Hashizume, S. van der Gaast, B.K.G. Theng, Adsorption of adenine, cytosine, uracil, ribose, and phosphate by Mg-exchanged montmorillonite, Clay Miner. 45 (2010) 469–475.
[33] H.J. Cleaves II, A.M. Scott, F.C. Hill, J. Leszczynski, N. Sahai, R. Hazen, Mineral–organic interfacial processes: potential roles in the origins of life, Chem. Soc. Rev. 41 (2012) 5502–5525.
[34] P. Pandey, C.K. Pant, K. Gururani, P. Arora, S. Kumar, Y. Sharma, H.D. Pathak, M. S. Mehata, Surface interaction of L-alanine on Hematite: an astrobiological implication, Orig. Life Evol. Biosph. 43 (2013) 331–339.
[35] P. Bhatt, C.K. Pant, Hemlata, S.C. Sati, Adsorption of bionanomolecule on clay minerals in relevance to chemistry-biology interfacial phenomenon, J. Chem. Eng. Chem. Res. 3 (11) (2016) 1075–1078.
[36] P. Thevenot, W.J. Hu, L.P. Tang, Surface chemistry influences implant biocompatibility, Curr. Top. Med. Chem. 8 (2008) 270–280.
[37] S. Jayrajsinh, G. Shankar, Y.K. Agrawal, L. Bakre, Montmorillonite nanoclay as a multifaceted drug delivery carrier: a review, J. Drug Deliv. Sci. Technol. 39 (2017) 200–209.
[38] M.A. Meyers, Po-Yu Chen, A. Yu-Min Lin, Y. Seki, Biological materials: structure and mechanical properties, Prog. Mater. Sci. 53 (2008) 1–206.
[39] B. Basile, A. Lazcano, J. Oró, Prebiotic syntheses of purines and pyrimidines, Adv. Space Res. 4 (12) (1984) 125–131.
[40] M.P. Robertson, S.L. Miller, An efficient prebiotic synthesis of cytosine and uracil, Nature 375 (6534) (1995) 772–774.
[41] J.P.T. Baú, P.R. Anizelli, H. de Santana, M.F. da Costa, D.A.M. Zaia, An experimental and theoretical vibrational study of interaction of cytosine and uracilwith artificial seawater: a prebiotic chemistry experiment, Vib. Spectrosc. 101 (2019) 92–99.
[42] A.W. Schwartz, H. Deuss, in: A.W. Schwartz (Ed.), Theory and Experiment in Exobiology, i, Wolters-Noordhoff, Groningen, 1971, p. 73.
[43] C. Varadachari, Origin of biological macromolecules on the young Earth, PINSA B64 (5&6) (1998) 249–266.
[44] H.J. Cleaves II, Prebiotic chemistry: geochemical context and reaction screening, Life 3 (2013) 331–345.
[45] Kamaluddin, M. Nath, A. Sharma, Adsorption of ribose and 2-deoxyribose 5'-nucleotides on metal ferrocyanides, Orig. Life Evol. Biosph. 71 (1990) 460–477.
[46] P. Chellan, P.J. Sadler, The elements of life and medicines, Philos. Trans. R. Soc. A373 (2015), 21040182.
[47] K.P. Carter, A.M. Young, A.E. Palmer, Fluorescent sensors for measuring metal ions in living systems, Chem. Rev. 114 (2014) 4564–4601.
[48] J. Anastassopoulou, T. Theophanides, The role of metal ions in biological systems and medicine, in: D.P. Kessissoglou (Ed.), Bioinorganic Chemistry, vol. 459, Kluwer Academic Publishers (ASIC, 1995, pp. 209–218.
[49] D.-W. Cho, C.-M. Chon, B.-H. Jeon, Y. Kim, M.A. Khan, H. Song, The role of clay minerals in the reduction of nitrate in groundwater by zero-valent iron, Chemosphere 81 (2010) 611–616.
[50] M.A. Khan, S.-w. Kim, R.A.K. Rao, R.A.I. Abou-Shanad, A. Bhatnagar, H. Song, B.-H. Jeon, Adsorption studies of dichloromethane on some commercially available GACs: effect of kinetics, thermodynamics and competitive ions, J. Hazard Mater. 178 (2010) 963–972.
[51] E.-R. Kenawy, A.A. Ghfar, S.M. Wabaidur, M.A. Khan, M.R. Siddiqui, Z. A. Alothman, A.A. Alqadami, M. Hamid, Cetyltrimethylammonium bromide intercalated and branched polyhydroxystyrene functionalized montmorillonite clay to sequester cationic dyes, J. Environ. Manag. 219 (2018) 285–293.
[52] M.B. Kurade, J.-Q. Xiong, S.P. Govindwar, H.-S. Roh, G.D. Saratale, B.-H. Jeon, H. Lim, Uptake and biodegradation of emerging contaminant sulfamethoxazole from aqueous phase using, Ipomoea aquatica.Chemosphere 225 (2019) 696–704.
[53] A. Parbhakar, J. Cuadros, M.A. Sephton, W. Dubbin, B.J. Coles, D. Weiss, Adsorption of L-lysine on montmorillonite, Colloids Surf. A Physicochem. Eng. Asp. 307 (2007) 142–149.
[54] H. Swenson, N.P. Stadie, Langmuir's Theory of adsorption: a centennial review, Langmuir 35 (16) (2019) 5409–5426.
[55] I. Langmuir, The adsorption of gases on plane surfaces of glass, mica and platinum, J. Am. Chem. Soc. 40 (1918) 1361–1403.
[56] A.H. Herz, R.P. Danner, G.A. Janussonis, Adsorption from Aqueous Solution [Adv. Chem. Ser., No. 79], Am. Chem. Soc, Washington DC, 1968, p. 173.
[57] H.M.F. Freundlich, Over the adsorption in solution, J. Phys. Chem. 57 (1906) 385–470.

[58] S.P. Patil, V.S. Shrivastavaa, G.H. Sonawane, S.H. Sonawane, Synthesis of novel Bi2O3–montmorillonite nanocomposite with enhanced photocatalytic performance in dye degradation, J. Environ. Chem. Eng. 3 (2015) 2597–2603.
[59] V. Singh, B.I. Fedeles, J.M. Essigmann, Role of tautomerism in RNA biochemistry: a review, RNA 21 (2014) 1–13.
[60] R. Krishnamurthy, Role of pKa of nucleobases in the origins of chemical evolution, Accounts Chem. Res. 45 (12) (2012) 2035–2044.
[61] C.E.A. Carneiro, G. Berndt, I.G. de Souza Jr., C.M.D. de Souza, A. Paesano Jr., A.C. S. da Costa, E. di Mauro, H. de Santana, C.T.B.V. Zaia, D.A.M. Zaia, Adsorption of adenine, cytosine, thymine, and uracil on sulfide-modified montmorillonite: FT-IR, Mössbauer and EPR spectroscopy and X-Ray diffractometry studies, Orig. Life Evol. Biosph. 41 (2011) 453–468.
[62] A. Pucci, S. Branciamore, M. Casarosa, L.P. D'Acqui, E. Gallori, Implications for an RNA-clay world: interaction of cytosine with clay minerals, Journal of Cosmology 10 (2010) 3398–3407.
[63] T. Yamada, K. Shirasaka, A. Takano, M. Kawai, Adsorption of cytosine, thymine, guanine and adenine on Cu (10) studied by infrared reflection absorption spectroscopy, Surf. Sci. 561 (2004) 233.
[64] J.P. Ferris, G. Ertem, V. Agarwal, Mineral catalysis of the formation of dimers of 5′-AMP in aqueous solution: the possible role of montmorillonite clays in the prebiotic synthesis of RNA, Orig. Life Evol. Biosph. 19 (2) (1989) 165–178.
[65] K.G. Taylor, K.O. Konhauser, Iron in Earth surface systems: a major player in chemical and biological processes, Elements 7 (2) (2011) 83–88.
[66] T. Fornaro, J.R. Brucato, S. Branciamore, A. Pucci, Int. J. Astrobiol. 12 (2013) 78.
[67] M. Anand, S.S. Russell, R.L. Blackhurst, M.M. Grady, Searching for signatures of life on Mars: an Fe-isotope perspective, Phil. Trans. R. Soc. B 361 (2006) 1715–1720.
[68] F. Westall, F. Foucher, N. Bost, M. Bertrand, D. Loizeau, J.L. Vago, G. Kminek, F. Gaboyer, K.A. Campbell, J.-B. Bréhéret, P. Gautret, C.S. Cockell, Biosignatures on Mars: what, where and how? Implications for the search for martian life, Astrobiology 15 (2015) 998–1029.
[69] A. Mojarro, G. Ruvkun, M.T. Zuber, C.E. Carr, Nucleic acid extraction from synthetic Mars analog soils for in situ life detection, Astrobiology 17 (8) (2017) 747–760.
[70] E. Greiner, K. Kumar, M. Sumit, A. Giuffre, W. Zhao, J. Pedersen, N. Sahai, Adsorption of L-glutamic acid and L-aspartic acid to γ-Al2O3, Geochimicaet Cosmochimica Acta 133 (2014) 142–155.
[71] V.K. Pearson, M.A. Sephton, A.T. Kearsley, P.A. Bland, I.A. Franchi, I. Gilmour, Clay mineral–organic matter relationships in the early solar system, Meteorit. Planet. Sci. 37 (2002) 1829–1833.
[72] Z. Martins, O. Botta, M.L. Fogel, et al., 1–2, in: Extraterrestrial Nucleobases in the Murchison Meteorite. Earth and Planetary Science Letters, vol. 270, Elsevier, Amsterdam, the Netherlands, 2008, pp. 130–136.
[73] B. Prieur, Is boric acid the missing link in prebiotic chemistry?, in: First Steps in the Origin of Life in the Universe Springer, Dordrecht, 2001, pp. 103–106.
[74] X. Liu, F. Zhou, R. Chi, J. Feng, Y. Ding, Q. Liu, Preparation of modified montmorillonite and its application tp rare earth adsorption, 12, Minerals 9 (2019) 747.

**Supporting Information**

# An electrochemical immunosensor for detection of *H. pylori* secretory protein VacA on g-$C_3N_4$/ZnO nanocomposite modified Au electrode.

**Kirti Saxena[a], Arun Kumar[b], Nidhi Chauhan[a], Manika Khanuja[b], Bansi D. Malhotra[c], Utkarsh Jain[a*]**

[a]*Amity Institute of Nanotechnology, Amity University Uttar Pradesh (AUUP), Sector-125, Noida – 201313, , India*

[b]*Centre for Nanoscience and Nanotechnology, Jamia Millia Islamia, New Delhi-110025*

[c]*Nanobioelectronics Laboratory, Department of Biotechnology, Delhi Technological University, Delhi 110042, India*

*Corresponding author: karshjain@gmail.com, ujain@amity.edu

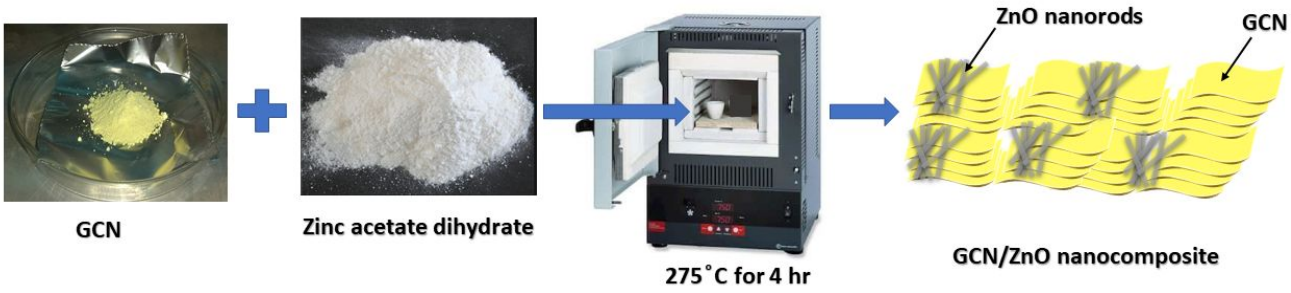



**Figure S1.** A schematic representation of synthesis of g-$C_3N_4$/ZnO nanocomposite.

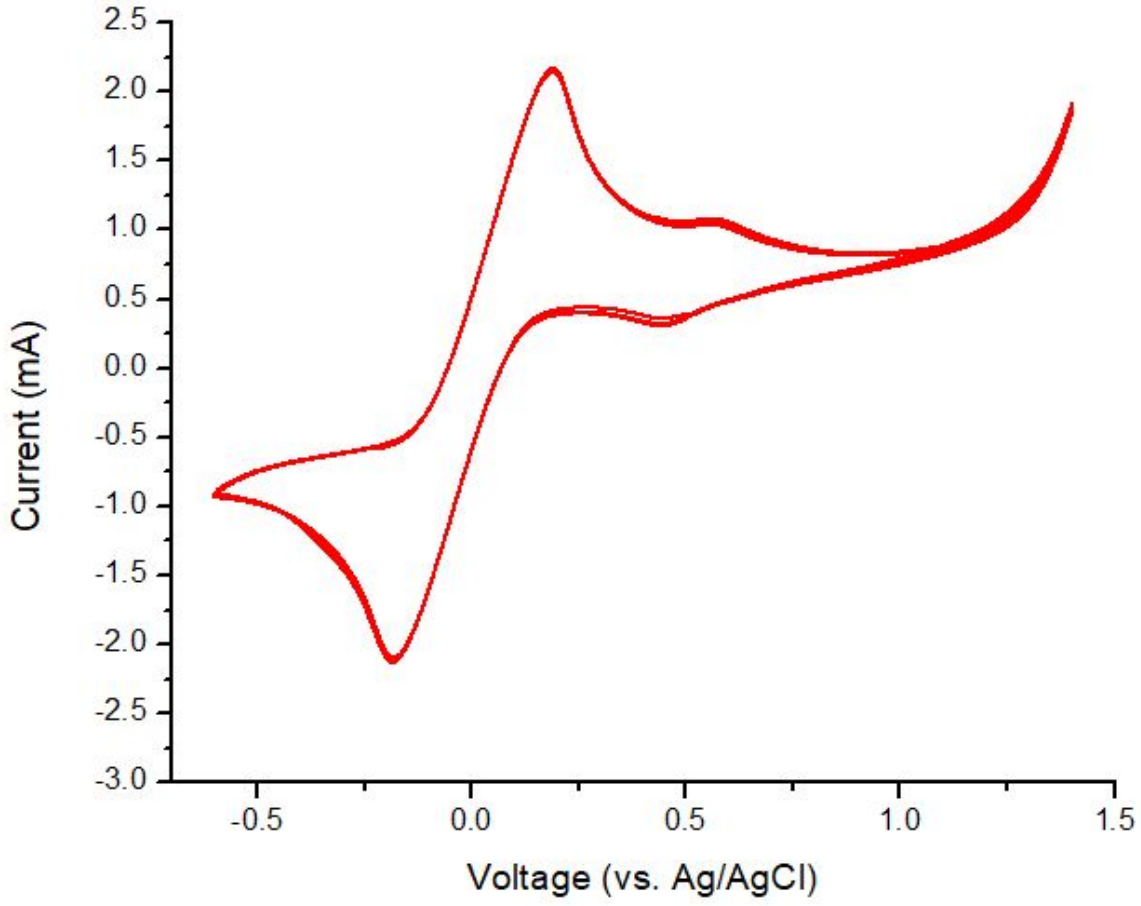



**Figure S2.** Electrodeposition of g-$C_3N_4$/ZnO nanocomposite at 50 mV/s in the potential range of -0.6-1.4 V for 10 cycles.

Archives of Computational Methods in Engineering
https://doi.org/10.1007/s11831-022-09812-w

REVIEW ARTICLE

# An Extensive Data-Based Assessment of Optimization Techniques for Distributed Generation Allocation: Conventional to Modern

**Vivek Saxena**[1] · **Narendra Kumar**[1] · **Uma Nangia**[1]

Received: 2 September 2021 / Accepted: 26 August 2022
© The Author(s) under exclusive licence to International Center for Numerical Methods in Engineering (CIMNE) 2022

## Abstract

Distributed generation (DG) is a comprehensive, compact, and salient segment of the centralized power system due to its extensive prospective benefits in constrained/unconstrained power quality parameters. Traditional power plants have their own technical, environmental, and economic diminutions for the expansion of power generation in the distribution network. Furthermore, unreserved fossil fuel escorted the electricity companies towards the renewable resources of energy generation. The enhancement of power quality parameters, cost-saving, and fulfilling the energy demand can be achieved by the integration of optimized DG in the distribution network. This paper presents the conventional to modern mathematical approaches enacted for the objective variables' empowerment via optimization of size and location of DG. The optimization intricacies, parameters, constraints, and benefits are also highlighted in this paper. The exhaustive data-based assessment carried out in this paper is a new work in the literature.

## Abbreviations

| | |
|---|---|
| ALO | Ant-Lion Optimization |
| ALRPF | Allow reverse power flow |
| AO | Anyone |
| AVRPF | Avoid reverse power flow |
| BCBV | Branch current to bus voltage |
| BCS | Bilateral Contract Scenario |
| BDG | Bio gass Distributed Generation |
| BGA | Binary Genetic Algorithm |
| BIBC | Bus injection to branch current |
| BLA | Bi-level Approach |
| CAC | Considering all constraints |
| CCT | Critical clearing time |
| CGA | Continuous Genetic Algorithm |
| CGSA | Classical Grid Search Algorithm |
| CL | Constant Load |
| CML | Commercial Load |
| COP | Cost Optimization |
| CPLSM | Combined power loss sensitivity method |
| CPSO | Clonal Particle Swarm Optimization |
| DABC | Discrete Artificial Bee Colony |
| DHPF | Decoupled harmonic power flow |
| DS | Dispatchable system |
| ECIM | Equivalent Current Injection Method |
| EL | Exhaustive Load Flow |
| ELR | Energy loss reduction |
| FLC | Fuzzy logic controller |
| GLBIW | Global Local Best Inertia Weight |
| GOA | Grasshopper Optimization Algorithm |
| GSA | Gravitation Search Algorithm |
| HDG | Hydro DG |
| HeS | Heuristic search |
| HL | High load |
| HPSO | Hybrid Particle Swarm Optimization |
| IB | Immune based |
| IGA | Improved Genetic Algorithm |
| IHRA | Improved Hereford Ranch Algorithm |
| IL | Industrial load |
| Itr. | Iteration |
| IVM | Index Vector Method |
| LaPF | Lagging power factor |
| LC | Load Concentration |
| LCCA | Life cycle cost analysis |
| LDC | Local distribution company |
| LePF | Leading power factor |
| LFC | Limited feeder capacity |
| LFG | Land fill gas |
| LI | Loss Incentive |
| LL | Low load |
| LLC | Line loading capacity |
| LLR | Line loss reduction |

✉ Vivek Saxena
vvksaxena1234@gmail.com

[1] Department of Electrical Engineering, Delhi Technological University, New Delhi, India

Published online: 09 September 2022

| | |
|---|---|
| LSM | Loss Sensitivity Method |
| MALO | Multi-objective Ant Lion Optimization |
| MCPSO | Modified Clonal Particle Swarm Optimization |
| MFO | Moth-Flame Optimization |
| MGT | Mini Gas turbine |
| ML | Medium Load |
| MMNRES | Multi membered non recombinative evolution strategy |
| MNM | Modified novel method |
| MPI | Multi objective performance index |
| MVO | Multi Verse Optimization |
| MXL | Mixed load |
| ND | Non differential |
| NPr | No preference |
| NPV | Net profit value |
| OIW | Constant Inertia Weight |
| PFR | Power flow reduction |
| PI | Power system performance index |
| PP | Probabilistic planning |
| PR | Pollutant reduction |
| PSBIT | Power Stabililty Based Index Technique |
| PSF | Price scaling factor |
| QR | Quick restoration |
| RI | Reliability improvement |
| RIW | Random Inertia Weight |
| RL | Residential load |
| RP | Repeated power flow |
| RPF | Reverse power flow |
| ΣDG | Number of Distributed Generator |
| SI | Stability improvement |
| $S_L^0$ | Peak load |
| SQP | Sequential Quadratic Programming |
| TI | Total incentive |
| Tiv | Time invariant |
| TLBO | Teaching learning-based Optimization |
| TP | Traditional planning |
| TSI | Transient stability improvement |
| Tv | Time variant |
| TVIW | Time Varying Inertia Weight |
| UCO | Uncontrolled output |
| UFC | Unlimited feeder capacity |
| Vdm | Maximum voltage drops |
| VL | Voltage limit |
| Vpm | V % mean |
| VSIM | Voltage Sensitivity Index Method |
| Wc | Clipping wind turbine generation output |
| $W_L$ | Weightage factor for power loss |
| WoDG | Without Distributed Generation |
| WSP | Without solar penetration |
| Wt | Turning off wind turbine generation output |
| WTBDG | Wind turbine DG + Bio mass DG |
| $W_{V1}$ | Weightage factor for voltage deviation |
| $W_{V2}$ | Weightage factor for voltage variation |

## 1 Introduction

In the modern era, distributed generation (DG) is an indispensable part of the electric power system. It is a futuristic approach for the modern power system due to its enormous advantages (technological benefits, financial benefits, and environmental benefits). The presence of DG in the distribution network enhances the stability and utilization of renewable energy compared to the previous conventional centralized generation (CG). DG bridges the gap between the generation of electric power and day-by-day increasing load demands. These advantages can be achieved by DG selection, site selection, analysis of load demand, renewable energy sources (RES) selection, and optimization for the placement of DG. The integration of DG in the distribution network also facilitates the reduced active power losses and reactive VA losses, enhanced voltage profile, and improved reliability and stability [1]. To create a renewably energized world such a revolutionary approach must be deliberately promoted at political level, social level, and individual levels. The separation from a conventional hierarchy of CG is escorted by the perception, recognition, and dissemination of RES in the distribution network. Moreover, DG unification with the centralized grid generation requires an advanced protection system that can devise the secured implementation to attain promising results despite the presence of various power quality constraints [2, 3].

The relative comparison between CG and DG could be done by the following factors [4]:

(a) *Output Capacity*

The generation capacity of CG is 100 MW-1000 GW while in DG it is up to 300 MW.

(b) *Technology Used*

In CG Hydroelectric plant, thermoelectric plant and nuclear power plant have been used while in DG the electric power has been generated by Diesel engine, Gas engine and RES.

(c) *Location*

CG is situated distant from the consumer and normally established in sector of non-renewable or renewable resources while DG is located near to the consumer facilities.

(d) *Generation to Distribution*

Step-up transformers have been used for high voltage transmission to substation in CG further step-down transformers are used for distribution from substation to distribution having reduced line

power losses. DG may or may not be connected to grid for low voltage distribution system.

### 1.1 Contribution of Paper

In this paper, various types of conventional to modern mathematical optimization approaches to DG allocation are considered for enhancing the power quality parameters, environmental parameters, and financial parameters of the power network in the presence of different constraints.

### 1.2 Paper Organization

In Section 2, employed aspirations of DG planning have been elaborated. Section 3 presents the optimization approaches (conventional as well as modern mathematical) with methodology, constraints, and optimized results. In Section 4, conclusion of the paper is presented.

## 2 Employed Aspirations for DG Planning

Multitudinous driving factors have been attained by the optimization of DG allocation and size selection in power system network. The aspirations of DG planning can be recapitulated as technological, economic, and environmental perspectives. The optimization DG allocation benefits are extensively illustrated in Fig. 1.

### 2.1 Technological Aspect

DG aims to very significant technical advantages which embrace the voltage profile enhancement, active and reactive power losses minimization, power factor optimization (PFO), line loss reduction and network MVA capacity maximization. The various researches for technological advantages are summarized as:

#### 2.1.1 Voltage Profile Enhancement

Voltage profile enhancement (VPE) is an imperative parameter in the power quality of the distribution system. The voltage level is boosted by increasing the penetration level of DG during optimized allocation in the distribution network. In the presence of DG, the VPE is selected as an objective function and the results are significant by providing the bidirectional power flow to/from the power grid during peak hours or valley hours of load [5]. The voltage stability is increased by using the incremental voltage (dv/dp) sensitivities method during the integration of DG [6]. Voltage profile along with voltage stability is also improved by utilizing the positive sequence of voltage ratio [7], power voltage curve [8], and voltage sensitivity index [9]. The voltage sag problem is reduced in a low voltage distribution network during various faults [10]. VPE along with power loss reduction (PLR) is facilitated by considering voltage amplitude as a function of injected power [11], while the max operator is used for such optimization in the Tehran electricity distribution grid [12, 13] and in radial distribution network [14]. The authors proposed a voltage stability index to enhance the voltage level by DG allocation [15]. P–V buses [16] and multiple micro turbines [17] have been incorporated for the improvement of voltage profile and such type of distribution network consist a renewable energy based optimized DG allocation. Voltage rise issue is also encountered with the integration of DG to meet the power requirement during time variant and invariant load [18].

#### 2.1.2 Power Loss Reduction

The integration and optimized allocation of DG in the distribution network enables the minimization of various power losses. Reducing the network losses, reactive power losses, and line loading is done by the optimum sizing and siting of DG in a meshed network [19] and to the load of concentrated buses in ref. [20]. Total energy losses are reduced by active and reactive VA injection at selected bus [9, 10] while real-time solar radiation and atmospheric temperature are considered to reduce the real power losses in a voltage distribution system of Thailand with the contemplation of power quality constraints [15]. A solar photo voltaic (SPV) system is used to reduce power losses with voltage limit consideration [21] while for power injection, DG size of 10–80% of system load demand is selected to diminish the actual power losses [22]. Moreover, actual power losses are decreased by the integration of multiple distributed generation (MDG) systems in continuous and discrete optimization [23]. Transmission line losses are reduced by ECIM which is a function of power quality parameters [24]. However, real power injection and reactive power compensation are done by wind power generation [6]. Total power losses are reduced by computing approximate losses for each bus [25] and compensating real component of branch currents [26]. Real power losses are minimized by using a metaheuristic optimization approach (MHOA) [27]. It is also minimized for hourly power flow with different penetration levels in different operating modes of wind turbines [28]. Annual energy losses are reduced with time-varying characteristics of load curve [29], random behavior of wind speed [30] and intermittent nature of RES [31], avoiding RPF [32], and allowing RPF [33].

#### 2.1.3 Power Factor Optimization

Power factor is also an important power quality parameter which is necessary to optimized in the presence of DG. To

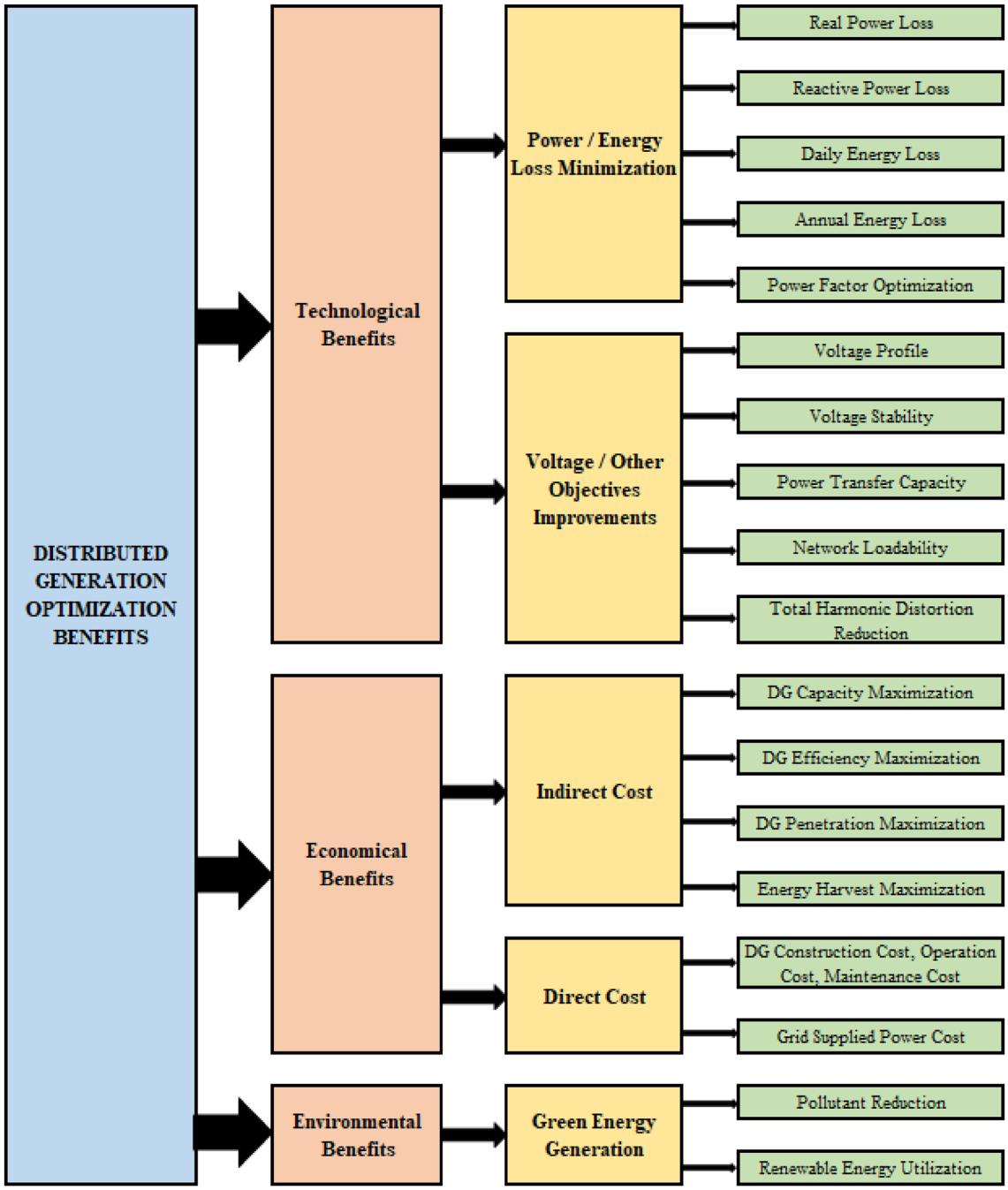



**Fig. 1** Benefits of DG optimization

reduce the real power losses, authors optimized the value of power factor by considering the maximum and minimum value of operating power factor while taking inequality boundary conditions, related practical concerns and rounded off issues into account [27]. DG size and power factor is simultaneously optimized by assuming the prespecified constant value of power factor at various load level [29]. Energy losses are reduced by optimizing the power factor in a battery energy storage (BES) integrated SPV system at various load level and the comparison is demonstrated at unity power (UPF), lagging power and leading power factor [34]. Impact of power factor variation is shown on power losses and voltage profile by taking UPF or non UPF of the system [35]. Optimal value of power factor is identified by considering the all-possible values of power factor using curve fitted technique and exact solution method [22]. In [9], a comparative analysis

has been shown for UPF and 0.9 lagging power factor with different DG optimization approaches.

#### 2.1.4 Total Harmonic Distortion Reduction

Total harmonic distortion reduction (THDR) is another determined objective of DG planning. The forward/backward sweep approach has been used to optimize the total harmonic distortion (THD) in a distributed system that comprises the amalgamation of passive elements and a harmonic current generator. Moreover, a harmonic spectrum is developed for a nonlinear (NL) load of deviated frequency driver and a convertible speed driver having branch current is a function of harmonic current [31]. THD is optimized in an SPV-based DG system where the various solar radiation levels are considered. THD in voltage profile and current profile have been measured by considering the background harmonics in an 11-kW grid-connected inverter [15].

### 2.2 Economical Aspect

Cost optimization (COP) is the pillar of the foundation for planning, execution, and maintenance of DG in a distribution system. Total cost (comprises installation cost, maintenance cost, and sag reduction cost) have been reduced by optimum DG allocation and the results are significant [10]. The distribution companies (DisCo)'s cost minimization and profit maximization are exhibited using BLA for DG planning in ref. [36]. DG capital cost with the state-dependent cost is considered in ref. [37] while fuel cost reduction (considering utilization factor) is achieved by the integration of DG in Japan's east power station [38]. The reimbursement time and anticipated profit rate are calculated by forming a multi-objective (MO) DG optimization for the sake of DisCo and owner benefits [39].

### 2.3 Environmental Aspect

Environment protection has been one of the most crucial aspects of human beings' existence on earth because of the aggressive use of fossil fuels in the traditional power system for meeting the peak demand of end-users. DG could be a prospective measure for the perforation of a natural source of energy into the electrical network of the power system for reducing greenhouse gaseous pollutants emissions and weather change [40]. DG has a great potential for the utilization of RES in power generation and is capable to build a low carbon emission grid. Carbon emission has been calculated for the complete life cycle of DG with the help of the carbon emission intensity factor having a procedure in which the life cycle of the product and pollutant part of the product have been pre-measured. With the penetration of RES in the power system, CO2 emission is reduced up to 1.8 million tons [41].

## 3 Optimization Approaches

Optimization approaches can be categorized into the following classification; conventional, modern mathematical, and hybrid.

### 3.1 Conventional Approaches

Conventional approaches of optimization are traditional basic search methods for the optimization of DG parameters. The proposed techniques have been executed by various researchers under this category. The classification of convention algorithm is given in Fig. 2. Moreover, a significant review and data-based assessment of such approaches for DG optimization are conferred in Table 1.

#### 3.1.1 Analytical Method (AM)

In this paper, several AM is reviewed that are used for optimum DG allocation in a distribution network. Total power losses are reduced by an equivalent current injection method in which the BIBC matrix and BCBV have been applied for finding the value of injected current and tested for three distribution systems without considering the admittance matrix and Jacobin matrix [24, 42]. The exact loss formula (ELF) is used for the DG optimization and this method is independent of the type of DG and able to generate the active power and reactive VA [25]. To minimize the energy losses in a three-phase unbalanced system feasible optimization interval (FOI) approach is used and bridging the gap between feeder demand characteristic and SPV based DG (SPVDG) characteristic [32]. DG optimization is done by power injection method (PIM) in a dispatchable system or non-dispatchable system with the consideration of the time-varying nature of the load and supply [29]. FOI technique has been implemented in which the RPF is considered and the power is injected at coupling node followed by the calculation of line losses by Carson equations [35]. An algebraic approach (AA) has been applied for VPE while in [7] an iterative method (IM) has been used for improving the power quality parameters and line loadabilty. In continuation of DG optimization, analytical techniques like a primal–dual interior-point algorithm (MPDIPA) [11], AA [43], multi-objective index (IMO) with self-correction algorithm (SCA) [34], and heuristic curve fitted method (HCFM) [22] are analyzed. In continuation, an IA is used to get the optimum size and location of DG that enhance the reliability and voltage profile of a distribution network. The approach is evaluated

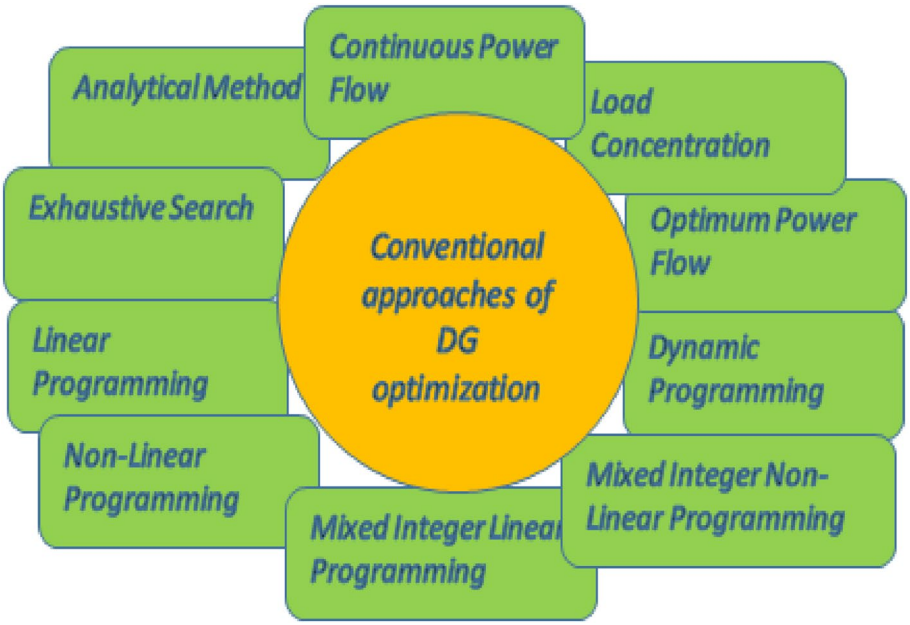



**Fig. 2** Conventional approaches of DG optimization

on IEEE 34 bus system and the results are significant on the of different indices [44].

#### 3.1.2 Exhaustive Search (ES)

In [45], the brute force method (BFM) is implemented for the SPVDG system incorporation of MATLAB programming with the consideration of daily load and supply curve without violating the European standard on power quality (EN 50,160 standards). It is like the backward/ forward sweep algorithm includes the large the R/X ratio of long feeder by assuming the π model of a distributed system. A two-stage technique is formed for the DG optimization which is dependent on the supply and load variation, the internal stage consists of ES while the external stage has clustering approach (CA) [46]. Power quality parameters are improved by using the probabilistic approach (PA) [15] and weightage ES [47] by optimal DG allocation. In continuation, Monte Carlo simulation and C language are also used. A Newton–Raphson (N-R) method along with IM has been enacted for the MO optimization problem to improvise the cost-effective power quality parameters [34].

#### 3.1.3 Linear Programming (LP)

The utilization of the distribution network is changed after the dissimilation of DG. The authors presented a methodology for the optimization of DG and getting the maximum harvesting of energy by considering the various power quality and financial constraints. The energy reaping is dependent on DG size, load level, DG location, occurred losses, and financial parameters. Moreover, the limitations which are occurred due to financial concerns are considered to implementing an IM for solving the linear optimization problem [48]. This methodology has the significance to validate the results of nature-inspired optimization techniques. The authors evaluated a genetic algorithm technique for the DG allocation in the actual distribution network of Egypt and the outcomes are justified with the help of LP. Moreover, MO optimization is carried out for optimizing the various parameters of power quality like LLR, SRI, VPE, and PFR [49].

#### 3.1.4 Non-Linear Programming (NLP)

Adaptive reactive power compensation must also be a benefit that would be facilitated in the DG incorporation in the distribution network. Continuing this, the authors demonstrated a DG optimization technique that is based on NLP to solve the MO function. Different objective functions are converted into a single objective function because of power loss reduction and enhanced voltage regulation. It has been highlighted that the application of this optimization technique is valid for different types of DGs by employing the fuzzification technique and adequate weighting scale to facilitate the one-level approach for various objective functions. However, the process of optimization is intended with the help of constraint optimization (CONOPT) which is a tool of general algebraic modeling system (GAMS) software used for nonstop power flow runs [50].

#### 3.1.5 Mixed-Integer Linear Programming (MILP)

In [36], a two-tier approach having the mutual approach among DisCo with the owner has been accomplished. This

**Table 1** Data-based assessment of DG optimization by adopted conventional approaches

| Ref | Conventional optimization method | Test systems | Different cases | Optimum DG size | Optimum DG location | Strengthened parameters | Compared approaches |
|---|---|---|---|---|---|---|---|
| [5] | AA | 48 km feeder | Vref = 0.94 (p.u.) | 144.7 (kVA) | 46.8 km | VPE | WoDG |
| | | | Vref = 0.95 (p.u.) | 292.2 (kVA) | 44.8 km | | |
| | | | Vref = 0.96 (p.u.) | 458.3 (kVA) | 42.6 km | | |
| | | | Vref = 0.97 (p.u.) | 650 (kVA) | 40.1 km | | |
| [7] | IA | IEEE 34 | Itr. −1 | 600 (kVA) | 890 | PLR, LLC, VPE, | VSIM |
| | | | Itr. −2 | 600 (kVA) | 852 | | |
| | | | Itr. −3 | 1200 (kVA) | 814 | | |
| [8] | MINLP | IEEE 41 | Dispatchable system (0.95 LePF) | 4.5 (MVA) | 40 | VPE | WoDG |
| | | | WTDG (0.95 LePF, UPF) | 8.8, 1.1 (MVA) | 19, 40 | | |
| | | | SPV (0.95 LePF, UPF) | 49.7, 1.06, 2.38 (MVA) | 19, 28, 40 | | |
| [9] | MNM | IEEE 33 | UPF | 2494.8 (kVA) | 6 | COP, PLR, VPE | WoDG |
| | | IEEE 69 | | 1832.53 (kVA) | 61 | | |
| | CPLSM | IEEE 33 | UPF | 1800 (kVA) | 8 | | |
| | | IEEE 69 | | 1850 (kVA) | 61 | | |
| | IVM | IEEE 33 | UPF | 1550 (kVA) | 30 | | |
| | | IEEE 69 | | 1850 (kVA) | 61 | | |
| | VSIM | IEEE 33 | UPF | 1000 (kVA) | 16 | | |
| | | IEEE 69 | | 1450 (kVA) | 65 | | |
| [11] | MPDIPA | IEEE 123 | ΣDG = 4 | 65.37,34.75, 12.17, 31.43 (kW) | 60, 36, 57,42 | PLR, VPE | WoDG |
| [15] | PA | IEEE 51 | SPV ΣDG = 1 | 0.8 (MW) | 38 | PLR, VPE, THDR | AM |
| | | | SPV ΣDG = 2 | 0.7, 0.9 (MW) | 38,19 | | |
| | | | SPV ΣDG = 2 with THD | 0.7, 0.5 (MW) | 38, 19 | | |
| [22] | HCFM | IEEE 69 | UPF | 1900 (kVA) | 61 | PFO, PLR, VPE | AM |
| | | | 0.85 LaPF | 2300 (kVA) | 61 | | |
| | | IEEE 32 | 0.85 LaPF | 2000 (kVA) | 29 | | |
| [24] | ECIM | IEEE 12 | ΣDG = 1 | 0.2272 (MW) | 9 | PLR | Classical grid search Method |
| | | IEEE 34 | ΣDG = 1 | 2.8848 (MW) | 21 | | |
| | | IEEE 69 | ΣDG = 1 | 1.8078 (MW) | 61 | | |
| [25] | ELF | IEEE 30 | ΣDG = 1 | 3.3 (MW) | 12 | PLR | Loss sensitivity, Repeated load flow |
| | | IEEE 33 | ΣDG = 1 | 2.49 (MW) | 6 | | |
| | | IEEE 69 | ΣDG = 1 | 1.81 (MW) | 61 | | |
| [29] | PIM | IEEE 69 | BΣDG = 2 | 0.89, 1.05 (MVA) | 62,35 | ELR, PFO | WoDG |
| | | | WTΣDG = 2 | 0.86, 0.99 (MVA) | 62,35 | | |
| | | | WTBΣDG = 4 | 0.49, 0.56, 0.71, 0.82 (MVA) | 62, 35, 62, 35 | | |
| [32] | FOI | IEEE 29 | SPV ΣDG = 1 | 0.2905 (MW) | 17 | PLR, VPE | WoDG |
| [33] | FOI | IEEE 29 | ΣDG = 1 (ALRPF) | 340.4 (kW) | 26 | PLR, VPE | AVRPF |
| | | | ΣDG = 1 (AVRPF) | 290.5 (kW) | 17 | | |
| | | IEEE 14 | ΣDG = 1 (ALRPF) | 803.1 (kW) | 9 | | |
| | | | ΣDG = 1 (AVRPF) | 601.7 (kW) | 7 | | |

**Table 1** (continued)

| Ref | Conventional optimization method | Test systems | Different cases | Optimum DG size | Optimum DG location | Strengthened parameters | Compared approaches |
|---|---|---|---|---|---|---|---|
| [34] | IMO & SCA | IEEE 33 | SPV + BES at UPF | 4.336 (MW) + 1.803 (MW) | 12, 20, 24 | PFO, PLR, VPE | Standard IEEE 1547 |
| | | | SPV + BES at LaPF | 4.336 (MW) + 1.804 (MW) | 12, 20, 24 | | |
| [36] | BLA | IEEE 34 | HL | 1.5 (MW) | 21 | COP | WoDG |
| | | | ML | 1.5 (MW) | 24 | | |
| | | | LL | 1.5 (MW) | 21 | | |
| [42] | BIBC, BCBV | IEEE 33 | Injecting P only | 0.981, 0.981, 0.981, 0.325 (MW) | 12, 30, 24, 5 | PFO, PLR, VPE | Repeated power flow |
| | | | Injecting P & Q | 1.16, 1.14, 1.13, 0.29 (MW) | 30, 11, 24, 31 | | |
| | | IEEE 69 | Injecting P only | 1.01, 0.797, 0.511, 0.318 (MW) | 61, 62, 17, 50 | | |
| | | | Injecting P & Q | 1.23, 0.99, 0.61, 0.88 (MW) | 61, 62, 17, 15 | | |
| [43] | AM | IEEE 30 | ΣDG = 1 | 15 (MW) | 5 | PLR | WoDG |
| [44] | IA | IEEE 33 | ΣDG = 2 | 2.7, 0.39 (MW) | 6, 30 | VPE, PLR | FFM, CS |
| [45] | BFM | IEEE 30 | ΣDG = 1 | 1 (MW) | 9 | PLR | Heuristic search |
| [46] | CA & ES | 24 Node | ΣDG = 5 | 40, 100, 15, 100, 400 (kW) | 4, 7, 9, 11, 13 | COP, PLR, VPE | GA, MINLP |
| [47] | ES | IEEE 6 | ΣDG = 2 | 3.4, 0.85 (MW) | 3, 5 | PLR, VPE | WoDG |
| | | IEEE 14 | ΣDG = 2 | 25.9, 25.9 (MW) | 10, 14 | | |
| | | IEEE 30 | ΣDG = 7 | 16.194, 8.097, 10.796, 8.097, 5.398, 2.699, 5.398 (MW) | 17, 18, 20, 24, 26, 27, 30 | | |
| [48] | LP | IEEE 7 | BΣDG = 1 | 8 (MW) | 7 | COP, PLR, VPE | WoDG |
| | | | ΣDG = 1 (LFG) | 650 (kW) | 3 | | |
| | | | HΣDG = 1, 2, 3 | 2, 1.5, 0.6 (MW) | 2, 3, 2 | | |
| | | | WTGn = 1, 2, 3 | 4.5, 8.5, 9.4 (MW) | 7, 2, 6 | | |
| [53] | OPF | IEEE 12 | FLCOPF (NPr) | 1.8, 30.7, 14.3 (MVA) | 1, 10, 11 | COP, PLR | IM |
| | | | Direct FLCOPF (NPr) | 0, 30.9, 17.6 (MVA) | 1, 10, 11 | | |
| | | | FLCOPF (CEL-1) | 2.6, 30.7, 11.7 (MVA) | 1, 10, 11 | | |
| | | | Direct FLCOPF (CEL-1) | 20.7, 2.3, 17.7 (MVA) | 1, 10, 11 | | |
| [54] | FLCOPF | IEEE 12 | OPF (NPr) | 12.3, 52.3, 38.2 (MVA) | 1, 10, 11 | COP, PLR | OPF |
| | | | FLCOPF (NPr) | 13.6, 49.1, 38.4 (MVA) | 1, 10, 11 | | |
| | | | OPF (@ bus 1) | 34.9, 52, 11.6 (MVA) | 1, 10, 11 | | |
| | | | FLCOPF (@ bus 1) | 35.9, 48.8, 12.2 (MVA) | 1, 10, 11 | | |
| [55] | OPF | IEEE 14 | ΣDG = 1 | 202.62 (MW) | 4 | COP, PLR | WoDG |
| | | | ΣDG = 2 | 195.05 (MW) | 4 | | |
| | | | ΣDG = 3 | 141.28 (MW) | 9 | | |

**Table 1** (continued)

| Ref | Conventional optimization method | Test systems | Different cases | Optimum DG size | Optimum DG location | Strengthened parameters | Compared approaches |
|---|---|---|---|---|---|---|---|
| | | | ΣDG = 4 | 41.94 (MW) | 14 | | |
| | | | ΣDG = 5 | 50.38 (MW) | 14 | | |
| | | | ΣDG = 6 | 42.84 (MW) | 14 | | |
| | | | ΣDG = 7 | 25.33 (MW) | 14 | | |
| [57] | OPF | IEEE 69 | ΣDG = 3 (P = 99%, α = 0.1) | 2.6614 (MW) | 26, 35, 62 | COP, PLR | WODG |
| | | | ΣDG = 5 (P = 99%, α = 0.1) | 3.9761 (MW) | 4, 26, 40, 49, 62 | | |
| | | | ΣDG = 7 (P = 99%, α = 0.1) | 5.5305 (MW) | 4, 26, 30, 35, 40, 49, 65 | | |
| | | | ΣDG = 9 (P = 99%, α = 0.1) | 6.0027 (MW) | 4, 13, 17, 26, 30, 40, 49, 58, 52 | | |
| | | | ΣDG = 3 (P = 99.999%, α = 0.01) | 2.6614 (MW) | 26, 35, 62 | | |
| | | | ΣDG = 5 (P = 99.999%, α = 0.01) | 4.0069 (MW) | 4, 26, 40, 49, 62 | | |
| | | | ΣDG = 7 (P = 99.999%, α = 0.01) | 5.5305 (MW) | 4, 26, 30, 35, 40, 49, 65 | | |
| | | | ΣDG = 9 (P = 99.999%, α = 0.01) | 6.0027 (MW) | 4, 13, 17, 26, 30, 40, 49, 58, 52 | | |
| [58] | LC | IEEE 13 | LCn @ 5 | 2, 1 (MW) | 5, 11 | PLR, VPE | Heuristic search |
| | | IEEE 25 | LCn @ 12 | 2, 2, 2, 0.5(MW) | 12, 15, 14, 7 | | |
| | | | LCn @ 17 | 2, 1.5 (MW) | 17, 22 | | |
| | | IEEE 30 | LCn @ 2 | 50, 50, 50, 10(MW) | 2, 9, 6, 28, 13 | | |
| [59] | CPF | IEEE 34 | ΣDG = 1 | 25 (MW) and 20 (MVA)r | 26 | PLR, LLC, VPE | WoDG |
| | | | ΣDG = 2 | 25 (MW) and 20 (MVA)r | 26, 33 | | |
| | | | ΣDG = 3 | 25 (MW) and 20 (MVA)r | 26, 33, 17 | | |
| [90] | N-R & IS | IEEE 6 | ΣDG = 1 | 6 (MW) | 3 | COP, PLR, VPE | AA |
| | | IEEE 14 | ΣDG = 1 | 16 (MW) | 8 | | |
| | | IEEE 30 | ΣDG = 1 | 35 (MW) | 11 | | |

bi-level MILP approach constitutes the upper level having DG location and central point of power injection while the lower level consists of payment minimization of DisCo to the energy market. A decision-making approach between the owner for profit maximization and DisCo for payment minimization is considered.

#### 3.1.6 Mixed-Integer Non-Linear Programming (MINLP)

The MINLP optimization technique is adopted in various researches [8, 51]. It will affect the DisCo investment planning for transformer purchasing and feeder up-gradation. A pool market model is developed having Lagrangian multiplier (LM) and line loss sensitivity with the consideration of equality and inequality constraints which is optimized by GAMS using a Sparse nonlinear optimizer solver for generation companies and DisCo profit [51]. Beta and Weibull probability distribution functions have been employed to calculate the probabilistic characteristics of natural sources for DG optimization [8].

#### 3.1.7 Dynamic Programming (DP)

Increasing the penetration of renewable energy-based DG in distributed generation is one of the futuristic aspects of energy generation. A reconfiguration of the chronological network has been proposed to reduce the limitation of natural sources. To enhance the power quality parameters, a Markov decision process is adopted to optimize the DG operation. Moreover, the present and future costs are considered to decide every step of the Markov model. Consequently, dynamic programming is used to optimize the proposed recursive model and evaluated on IEEE 33 and IEEE 123 test systems [52].

#### 3.1.8 Optimum Power Flow (OPF)

Several researchers [53, 54, 55, 56, 57] presented the OPF technique for the allocation of DG. CONOPT with a generalized gradient method is used for finding the network generation capacities. For maximizing the generation capacities, the NL optimization executed by finding the multiple local optima is used in a highly meshed reliable system. Initially no contingencies in the problem formulation and solved using scrutiny constrained OPF models then selection of contingency without constraint and include the most violated constraints. In [53, 54], Fault level constraint optimum power flow (FLCOPF) is considered for sinks source concept having bi-level procedures OPF and generation reduction optimization algorithm to solve the quadratic benefit function by considering capacity expansion locations (CELs). For social welfare and profit maximization, local marginal price and short-run marginal price with LM are considered [55]. In continuation of the above social welfare, maximization involves the quadratic curves of benefit, bid, and cost given by DISCOs, seller, and DG owner respectively while profit maximization includes the two-block (inner and outer) approach. OPF technique with genetic algorithm for a variable number of DG [56] and as ordinal optimization (OO) for three-level approach [57] has been executed. Moreover, out of search space, one of all possible solutions is in the top percent α with a probability level of P.

#### 3.1.9 Load Concentration (LC)

In [20], the Kalman filter algorithm is executed for the optimization of DG allocation. This algorithm has the properties of smoothing and noise rejection which are used to solve the linear time-varying equations. Initially, the losses are calculated by the N-R method then the state vector estimation is completed in the measurement update and time update stages which are evaluated by the root mean square error. In [58], the concept of the equivalent load is implemented in two-step approaches of load centroid (LCn) and performance index. Firstly, the equivalent load is calculated then the PI is approachable for finding the active power loss and average node voltage variation.

#### 3.1.10 Continuous Power Flow (CPF)

Referring to the continuation parameters and two levels iterative approach (IA) most sensitive buses have been found for optimum location of DG approaching towards bifurcation point from a stable point. CPF optimized system having predictor and corrector step could be incorporated as a compensator or as a large source for DG units [59].

### 3.2 Modern Mathematical Approaches

Modern mathematical approaches are basically artificial intelligent techniques founded on the performance of society and nature. Such type of approaches can be predominantly designated as given in Fig. 3. An assessment of modern mathematical approaches for DG optimization is manifested in Table 2.

#### 3.2.1 Evolutionary Programming (EP)

EP is one of the matured artificial intelligence optimization algorithms based on the artificers of real anthology having metamorphosis, contention, and evolution levels which have been used for non-continual, uneven, and undistinguishable optimization problems unlike the conventional methods of optimization. Reactive power planning has been executed using the probability transition rule in [60] comprises the following steps: initialization, statistics, mutation, completion, and determination. The authors demonstrated the EP on 12.66 kV, IEEE 69 test system with two modes of operation involving turning off DG based on wind turbine (WTDG) and clipping of the output using an index-based scheme and sensitivity analysis by keeping the ratio of dispatched wind energy to load (WPDLR) within designated limit [28].

#### 3.2.2 Genetic Algorithm (GA)

GA has the natural assortment that relates to the greater category of EP and has been used to evaluate for the optimum DG allocation in the power system [61]. In continuation, GA has been demonstrated to produce higher grade optimization by leaning on bio-simulated steps like mutation, crossover, and selection. It has been implemented for NL optimization in a Tehran regional electricity company considering the financial objectives that include the pricing of power connectors and the energy which is not provided by the system [14]. Moreover, the linear problem has been optimized for the enhancement of the power quality parameter in the West Delta sub-transmission network [49]. Minimization of

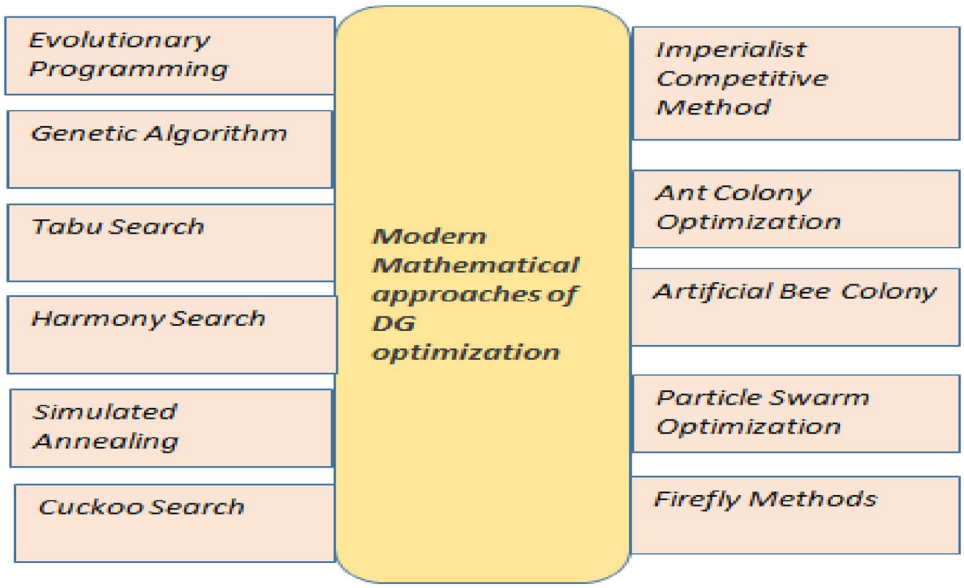



**Fig. 3** Modern mathematical approaches of DG optimization

voltage sag effect that is calculated by voltage divider rule with the line loss minimization and cost-saving subjected to the limitations of power flow, voltage level. The optimum size of DG is achieved by a stochastic optimization technique with a point of common coupling consideration during various fault conditions [10]. The various types of load models have been considered for the optimization results with the consideration of source as negative sink and index-based performance evaluation [62]. GA has been integrated with OPF [56] and AM [16] for DG optimization. In [63], step by step approach has been used for the optimization of the size and location of DG in main and sub-network distribution systems using single-step restoration and supplied by utility-owned DG to encounter the power outage problem due to cold load pick up conditions. A non-dominated sorting GA (NSGA) and forward/ backward sweep method have been executed for the renewable DG / nonrenewable DG (NRDG) allocation and size selection in [31]. Zonal division of peak load [64], moment method and central limit theorem [65] have been incorporated with DG planning for the cost, power quality parameters, and reliability optimization. Moreover, SPV and WTDG have been characterized as DG for Weibull reliability, and maintenance costs optimization results in annual operating cost (AOC) minimization.

#### 3.2.3 Tabu Search (TS)

TS was proposed in 1986 that is based on the planning of forbidden moves violating the cycling to form of strong inhibition search in which enough memory is saved for the tabu from down to up in the tabu list. TS technique is employed for the optimization of DG placement considering the residential, commercial, and industrial load having three types of algorithms proceed simultaneously which involves rounding calculation, local minimum finding, and optimum installation respectively [66]. The procedure involves initialization, neighborhood solution, loss calculations, repetition, and updating of the current solution till a maximum number of iterations to get the optimization. If the neighborhood solution is healthier than the present set of values then the neighborhood result is adopted that will satisfy the aspiration in the tabu list. The authors presented a methodology, for the optimization of DG resources and reactive power resources optimization with the involvement of short, intermediate (record best rail solution) and long-term memory (more than 3 successive iterations) having controlled output (CO) and Z scenarios [67]. In this work, the forbidden move i.e., the addition of a new move and deletion of the old one is involved having the updated solution in the next iteration of the tabu list whose length is directly proportional to the better solution.

#### 3.2.4 Harmony Search (HS)

Geem et al. suggested this heuristic technique, is built on imitating the extemporization of the music player for finding the best condition of music by an artistic estimation. Musicologist Tirro and French composer Jean Philippe Rameau documented the history of American Jazz and invented the classical method of harmony examined on Traveling salesman problem (TSP) [68]. Improved multi-objective harmony search (IMOHS) has been suggested for the optimization of DG placement having qualitative comparison with NSGA II. IMOHS includes the search process of novel global HS which includes the mutation probability and excludes the harmony memory considering rate parameters. Mainly two

**Table 2** Data-based assessment of DG optimization by adopted modern mathematical approaches

| Ref | Modern mathematical optimization method | Test systems | Different cases | Optimum DG sizes | Optimum DG locations | Strengthened parameters | Optimized results | Compared approaches |
|---|---|---|---|---|---|---|---|---|
| [6] | DE | IEEE 6 | ΣDG = 2 | 0.48995, 0.63858 (p.u.) | 3, 6 | PLR, VPE | 0.020545 p.u | Base Care WoDG, BBPSO, MMN-RES |
| | | IEEE 30 | ΣDG = 6 | 0.16362, 0.11973, 0.21795, 0.20327, 0.013192, 0.29578 (p.u.) | 16, 18, 19, 23, 25, 27 | | 0.088605 p.u | |
| [10] | GA | IEEE 34 | ΣDG = 4 | 500 (kW) each | 20, 25, 8, 17 | COP, PLR, VPE | 102.76 (kW) | |
| | | IEEE 30 | ΣDG = 5 | 510 (kW) each | 18, 11, 25, 21 | | 104.6 (kW) | |
| [12] | GA | IEEE 13 | V % Mean = 98.823 | 3200 (kW) | 13 | PLR, VPE | 100.25 (kW) | |
| | | | V % Mean = 98.81 | 1600, 1600 (kW) | 13, 9 | | 92.9 (kW) | |
| [13] | GA | IEEE 13 | V % Mean = 98.823 | 3200 (kW) | 13 | PLR, VPE | 100.25 (kW) | |
| | | | V % Mean = 98.81 | 1600, 1600 (kW) | 13, 9 | | 92.9 (kW) | |
| [18] | GA | IEEE 12 | ΣDG = 1 (Tiv) | 0.24 (MW) | 9 | COP, PLR, VPE | 0.0046 (MW) | WoDG |
| | | | MDG (Tiv) | 0.0755, 0.0484, 0.0676, 0.0594 (MW) | 6, 8, 9, 10 | | | |
| | | | ΣDG = 1 (Tv) | 0.13 (MW) | 9 | | 0.0035 (MW) | |
| | | | MDG (Tv) | 0.0401, 0.0601, 0.0401, 0.0267 (MW) | 6, 8, 10, 12 | | | |
| [19] | IHRA | IEEE 6 | ΣDG = 4 | 15 (MW) | 2, 3, 4, 6 | PLR | 0.0498 (MW) | GA |
| | | IEEE 14 | ΣDG = 4 | 26 (MW) | | | 11.12% | |
| | | IEEE 30 | ΣDG = 4 | 28.3 (MW) | | | 9.08% | |
| [23] | PSO | IEEE 69 | ΣDG = 1 | 1904.2 (kW) | 61 | PLR | 23.9 (kW) | SQP |
| | | | ΣDG = 2 | 1582, 322 (kW) | 61, 21 | | 13.6 (kW) | |
| | | | ΣDG = 3 | 1278, 301, 324 (kW) | 61, 64, 21 | | 12.8 (kW) | |
| [27] | ABC | IEEE 69 | ΣDG = 1, load = 3802 (kW) + 2694 (kVA)r | 2200 (kVA) | 61 | COP, PFO, PLR | 23.9199 (kW) | GA, ELF |
| | | | ΣDG = 1, load = 150% | 3400 (kVA) | 61 | | 54.7271 (kW) | |
| | | | MDG, load = 3802 (kW) + 2694 (kVA)r | 2100, 600 (kVA) | 61, 17 | | 7.99901 (kW) | |
| | | | MDG, load = 150% | 3200, 900 (kVA) | 61, 17 | | 17.9966 (kW) | |
| | | | ΣDG = 1 + capacitor (load = 100%) | 2200 (kVA), 300 (kVA)r | 61, 18 | | 18.5512 (kW) | |
| | | | ΣDG = 1 + capacitor load = 150%) | 3300 (kVA), 600 (kVA)r | 61, 16 | | 42.1286 (kW) | |
| | | | Active Power DG (load = 100%) | 1800 (kW) + 1350 (kVA)r | 61 | | 23.282 (kW) | |
| | | | Active Power DG (load = 150%) | 2700 (kW) + 1950 (kVA)r | 61 | | 53.2194 (kW) | |
| | | | DG size is fixed | 1000 (kVA) | 61 | | 80.29 (kW) | |
| [28] | EP | IEEE 69 | Wc, WPDLR = 0 | 0.25 (MW) each (1 SPV, 1 WTDG) | 53 | PLR, VPE | 65.69 (MWh) | GA, ES |
| | | | Wc, WPDLR = 0.1 | 0.25 (MW) each (1 SPV, 1 WTDG) | 51, 53 | | 51.16 (MWh) | |

**Table 2** (continued)

| Ref | Modern mathematical optimization method | Test systems | Different cases | Optimum DG sizes | Optimum DG locations | Strengthened parameters | Optimized results | Compared approaches |
|---|---|---|---|---|---|---|---|---|
| | | | Wc, WPDLR = 0.2 | 0.25 (MW) each (1 SPV, 1 WTDG) | 54, 50 | | 44.81 (MWh) | |
| | | | Wc, WPDLR = 0.3 | 0.25 (MW) each (1 SPV, 1 WTDG) | 53, 50 | | 42.58 (MWh) | |
| | | | Wc, WPDLR = 0.4 | 0.25 (MW) each (1 SPV, 1 WTDG) | 53, 50 | | 42.17 (MWh) | |
| | | | Wt, WPDLR = 0 | 0.25 (MW) each (1 SPV, 1 WTDG) | 53 | | 65.69 (MWh) | |
| | | | Wt, WPDLR = 0.1 | 0.25 (MW) each (1 SPV, 1 WTDG) | 53, 53 | | 61.4 (MWh) | |
| | | | Wt, WPDLR = 0.2 | 0.25 (MW) each (1 SPV, 1 WTDG) | 53, 53 | | 56.04 (MWh) | |
| | | | Wt, WPDLR = 0.3 | 0.25 (MW) each (1 SPV, 1 WTDG) | 54, 50 | | 48.17 (MWh) | |
| | | | Wt, WPDLR = 0.4 | 0.25 (MW) each (1 SPV, 1 WTDG) | 53, 50 | | 43.79 (MWh) | |
| | | | Wc, WPDLR = 0 | 0.125 (MW) each (2 SPV, 2 WTDG) | 53, 54 | | 65.69 (MWh) | |
| | | | Wc, WPDLR = 0.1 | 0.125 (MW) each (2 SPV, 2 WTDG) | 53, 54 and 51, 53 | | 51.14 (MWh) | |
| | | | Wc, WPDLR = 0.2 | 0.125 (MW) each (2 SPV, 2 WTDG) | 52, 54 and 50, 53 | | 44.7 (MWh) | |
| | | | Wc, WPDLR = 0.3 | 0.125 (MW) each (2 SPV, 2 WTDG) | 50, 53 and 50, 53 | | 42.53 (MWh) | |
| | | | Wc, WPDLR = 0.4 | 0.125 (MW) each (2 SPV, 2 WTDG) | 53, 54 and 50, 52 | | 42.15 (MWh) | |
| | | | Wt, WPDLR = 0 | 0.125 (MW) each (2 SPV, 2 WTDG) | 53, 54 | | 65.69 (MWh) | |
| | | | Wt, WPDLR = 0.1 | 0.125 (MW) each (2 SPV, 2 WTDG) | 54, 54 and 53, 53 | | 58.43 (MWh) | |
| | | | Wt, WPDLR = 0.2 | 0.125 (MW) each (2 SPV, 2 WTDG) | 52, 54 and 52, 53 | | 48.64 (MWh) | |
| | | | Wt, WPDLR = 0.3 | 0.125 (MW) each (2 SPV, 2 WTDG) | 53, 53 and 50, 51 | | 44.27 (MWh) | |
| | | | Wt, WPDLR = 0.4 | 0.125 (MW) each (2 SPV, 2 WTDG) | 51, 54 and 50, 53 | | 42.59 (MWh) | |
| [31] | GA | IEEE 31 | NRDG | 1.5, 0.5, 1, 1.5, 1.5 (MW) | 8,12,13,28,30 | PLR, THDR | 11,960 (MWh) | WoDG |

**Table 2** (continued)

| Ref | Modern mathematical optimization method | Test systems | Different cases | Optimum DG sizes | Optimum DG locations | Strengthened parameters | Optimized results | Compared approaches |
|---|---|---|---|---|---|---|---|---|
| | | | NRDG + WTDG | 1.5, 0.7, 1.4, 0.6, 0.6 (MW) | 8,15,17,20,30 | | 12,784 (MWh) | |
| | | | NRDG + SPV | 1.15, 0.3, 0.4, 0.85, 0.3 (MW) | 12,13,15,17,31, | | 15,410 (MWh) | |
| | | | NRDG + WTDG + SPV | 0.3, 1.5, 1.4, 0.8, 1.5 (MW) | 6,12,28,30,31 | | 15,121 (MWh) | |
| [35] | PSO | IEEE 16 | ΣDG = 1 | 12.97 (MW) | 9 | COP, PFO, PLR, VPE | 168.1 (kW) | IAM |
| | | | ΣDG = 2 | 12.97, 5.86 (MW) | 9, 6 | | 111.7 (kW) | |
| | | | ΣDG = 3 | 13.1, 5.897, 4.29 (MW) | 9, 6, 16 | | 76.4 (kW) | |
| | | | ΣDG = 1 | 2.591 (MW) | 6 | | 111.1 (kW) | |
| | | | ΣDG = 2 | 1.002, 1.0195 (MW) | 12, 30 | | 87.5 (kW) | |
| | | | ΣDG = 3 | 0.88, 1.0928, 1.0098 (MW) | 13, 24, 30 | | 73.2 (kW) | |
| | | | ΣDG = 1 | 1.8062 (MW) | 61 | | 78.6 (kW) | |
| | | | ΣDG = 2 | 1.8062, 0.511 (MW) | 61, 17 | | 67 (kW) | |
| | | | ΣDG = 3 | 1.8062, 0.511, 0.719 (MW) | 61, 17, 50 | | 65.5 (kW) | |
| [38] | ABC | IEEE 30 | ΣDG = 1 | 200 (MW) | 17 | COP, PR | | Without solar penetration |
| | | | ΣDG = 2 | 136, 97 (MW) | 5, 23 | | | |
| | | | ΣDG = 3 | 198, 127, 176 (MW) | 14, 22, 23 | | | |
| | | | ΣDG = 4 | 68, 109, 1193, 102 (MW) | 11, 13, 26, 29 | | | |
| [39] | PSO | IEEE 33 | ERR = 15% | 1, 1, 0.9 (MW) | 7, 33, 15 | PLR, RI, SI VPE | TLPI = 0.1642 p.u | WoDG |
| | | | ERR = 20% | 0.9, 0.9, 1 (MW) | 6, 32, 14 | | TLPI = 0.1568 p.u | |
| | | | ERR = 25% | 0.9, 0.9, 1 (MW) | 6, 32, 14 | | TLPI = 0.1568 p.u | |
| | | | ERR = 30% | 0.9, 0.9, 1 (MW) | 6, 32, 14 | | TLPI = 0.1568 p.u | |
| | | | ERR = 35% | 1, 1, 0.9 (MW) | 6, 32, 13 | | TLPI = 0.1479 p.u | |
| | | | ERR = 40% | 1, 1, 0.9 (MW) | 7, 31, 13 | | TLPI = 0.14665 p.u | |
| | | | ERR = 45% | 1, 1, 1 (MW) | 6, 29, 12 | | TLPI = 0.162 p.u | |
| | | | ERR = 50% | 1, 1, 1 (MW) | 6, 29, 12 | | TLPI = 0.1635 p.u | |
| [49] | GA | IEEE 52 | GA (VPE = 24.4%, SRI = 63.1%) | 0.3105 p.u | 50 | LLR, SRI, VPE, PFR | PFR = 42.016%, LLR = 81.5% | LP |
| | | | LP (VPE = 24.26%, SRI = 63.075%) | 0.3 p.u | 50 | | PFR = 42.709%, LLR = 80.7% | |
| [62] | GA | IEEE 13 | CL | 0.63 p.u | 7 | PLR | 0.0161 p.u | WoDG |
| | | | IL | 0.61 p.u | 8 | | 0.0167 p.u | |
| | | | RL | 0.59 p.u | 8 | | 0.0167 p.u | |

**Table 2** (continued)

| Ref | Modern mathematical optimization method | Test systems | Different cases | Optimum DG sizes | Optimum DG locations | Strengthened parameters | Optimized results | Compared approaches |
|---|---|---|---|---|---|---|---|---|
| | | | CML | 0.58 p.u | 8 | | 0.0171 p.u | |
| | | | MXL | 0.62 p.u | 8 | | 0.0168 p.u | |
| | | IEEE 37 | CL | 0.62 p.u | 14 | | 0.001889 p.u | |
| | | | IL | 0.63 p.u | 25 | | 0.00166 p.u | |
| | | | RL | 0.63 p.u | 25 | | 0.001664 p.u | |
| | | | CML | 0.63 p.u | 25 | | 0.001646 p.u | |
| | | | MXL | 0.63 p.u | 25 | | 0.001663 p.u | |
| [63] | GA | IEEE 33 | ΣDG = 2 | 700, 500 (kVA) | 10, 31 | QR, PLR | - | WoDG |
| [64] | GA | IEEE 39 | MGT | 500 (kW) | 36 | COP, RI | AOC = 141,920 $ / year | WoDG |
| | | | MGT | 1000 (kW) | 19 | | AOC = 236,520 $ / year | |
| [65] | GA | IEEE 97 | WTΣDG-1 | 1 (MVA) | Any bus | COP, PR, RI | AOC = 15,000 $ / (MVA) | WoDG |
| | | | WTΣDG-2 | 1.5 (MVA) | Any bus | | AOC = 12,750 $ / (MVA) | |
| | | | WTΣDG-3 | 2 (MVA) | Any bus | | AOC = 10,500 $ / (MVA) | |
| | | | SPV ΣDG-1 | 16 (MVA) | Any bus | | AOC = 20,500 $ / (MVA) | |
| | | | SPV ΣDG-2 | 18 (MVA) | Any bus | | AOC = 18,750 $ / (MVA) | |
| | | | SPV ΣDG-3 | 19 (MVA) | Any bus | | AOC = 19,625 $ / (MVA) | |
| [66] | TS | Model 1 (1 S/S, 4 Feeder and 28 Sections) | AL-1 (ΣDG = 10) | 4000 (kW) | - | PLR | 1725 (kW) | SA |
| | | | AL-2 (ΣDG = 10) | 4000 (kW) | | | 1823 (kW) | |
| | | | AL-3 (ΣDG = 10) | 4000 (kW) | | | 1823 (kW) | |
| | | Model 1 (1 S/S, 4 Feeder and 28 Sections) | AL-1 (ΣDG = 20) | 4000 (kW) | | | 1211 (kW) | |
| | | | AL-2 (ΣDG = 20) | 4000 (kW) | | | 1292 (kW) | |
| | | | AL-3 (ΣDG = 20) | 4000 (kW) | | | 1299 (kW) | |

**Table 2** (continued)

| Ref | Modern mathematical optimization method | Test systems | Different cases | Optimum DG sizes | Optimum DG locations | Strengthened parameters | Optimized results | Compared approaches |
|---|---|---|---|---|---|---|---|---|
| | | Model 2 (4 S/S, 6 Feeder and 78 Sections) | AL-2 (ΣDG = 15) | 13.7 (kW) | | | 26,554 (kW) | |
| | | | AL-3 (ΣDG = 15) | 13.7 (kW) | | | 35,250 (kW) | |
| | | Model 2 (4 S/S, 6 Feeder and 78 Sections) | AL-2 (ΣDG = 35) | 14.9 (kW) | | | 13,044 (kW) | |
| | | | AL-3 (ΣDG = 35) | 14.9 (kW) | | | 18,647 (kW) | |
| [67] | TS | IEEE 33 | UCO, SL0 = 3715 + j 2300 (kVA) | 275, 325, 150, 100, 450 (kW) | 8, 16, 24, 27, 32 | COP, PLR, VPE | 54.43 (kW) | |
| | | | $S_L^1$ = 0.8 (3715 + j 2300) (kVA) | 275, 325, 150, 100, 450 (kW) | 8, 16, 24, 27, 32 | | 22.87 (kW) | |
| | | | $S_L^2$ = 0.6 (3715 + j 2300) (kVA) | 275, 325, 150, 100, 450 (kW) | 8, 16, 24, 27, 32 | | 6.93 (kW) | |
| | | | $S_L^3$ = 0.4 (3715 + j 2300) (kVA) | 275, 325, 150, 100, 450 (kW) | 8, 16, 24, 27, 32 | | 5.47 (kW) | |
| | | | $S_L^4$ = 0.2 (3715 + j 2300) (kVA) | 275, 325, 150, 100, 450 (kW) | 8, 16, 24, 27, 32 | | 17.46 (kW) | |
| | | | CO, SL 0 = 3715 + j 2300 (kVA) | 275, 400, 75, 50, 500 (kW) | 8, 16, 24, 27, 32 | | 50.32 (kW) | |
| | | | $S_L^1$ = 0.8 (3715 + j 2300) (kVA) | 275, 400, 75, 50, 500 (kW) | 8, 16, 24, 27, 32 | | 21.59 (kW) | |
| | | | $S_L^2$ = 0.6 (3715 + j 2300) (kVA) | 275, 400, 75, 50, 500 (kW) | 8, 16, 24, 27, 32 | | 7.88 (kW) | |
| | | | $S_L^3$ = 0.4 (3715 + j 2300) (kVA) | 275, 225, 75, 50, 325 (kW) | 8, 16, 24, 27, 32 | | 3.24 (kW) | |
| | | | $S_L^4$ = 0.2 (3715 + j 2300) (kVA) | 175, 75, 75, 50, 125 (kW) | 8, 16, 24, 27, 32 | | 0.67 (kW) | |
| | | | CO with 300 (kW), 300 (kVA)r Source, SL 0 = 3715 + j 2300 (kVA) | 300, 300, 125, 275, 300 (kW) | 8, 16, 24, 27, 32 | | 53.1 (kW) | |
| | | | $S_L^1$ = 0.8 (3715 + j 2300) (kVA) | 300, 300, 125, 250, 300 (kW) | 8, 16, 24, 27, 32 | | 22.7 (kW) | |
| | | | $S_L^2$ = 0.6 (3715 + j 2300) (kVA) | 300, 300, 125, 250, 300 (kW) | 8, 16, 24, 27, 32 | | 6.83 (kW) | |
| | | | $S_L^3$ = 0.4 (3715 + j 2300) (kVA) | 200, 125, 125, 215, 75 (kW) | 8, 16, 24, 27, 32 | | 2.68 (kW) | |
| | | | $S_L^4$ = 0.2 (3715 + j 2300) (kVA) | 250, 225, 125, 250, 225 (kW) | 8, 16, 24, 27, 32 | | 0.8 (kW) | |
| [69] | HS | IEEE 33 | Load = 3.72 (MW), 2.3 (MVA)r | 0.9369, 0.6672, 1.0117 (MW) | 6, 14, 24, 31 | PLR, VPE | 0.06783 p.u | GA/ PSO, GA, PSO |
| | | IEEE 69 | Load = 3.8 (MW), 2.69 (MVA)r | 1.4552, 0.4769, 0.3124 (MW) | 61, 64, 21 | | 0.0105 p.u | |
| [72] | ICM | IEEE 33 | ΣDG = 1 | 200 (kW) | - | PLR | 201.38 (kW) | GA |
| | | IEEE 69 | ΣDG = 1 | 200 (kW) | | | 211.45 (kW) | |
| [73] | ICM | IEEE 69 | ΣDG = 3 | 0.385, 1.186, 1 (MW) | 26, 35, 62 | PLR | LI = 7.535 | GA + OPF, OO, IB, PSO, GA, GAMS, IGA |

**Table 2** (continued)

| Ref | Modern mathematical optimization method | Test systems | Different cases | Optimum DG sizes | Optimum DG locations | Strengthened parameters | Optimized results | Compared approaches |
|---|---|---|---|---|---|---|---|---|
| | | | ΣDG = 5 | 1.059, 0.85, 0.873, 0.717, 0.9 (MW) | 4, 26, 40, 48, 62 | | LI = 9.438 | |
| | | | ΣDG = 7 | 0.998, 0.686, 0.676, 0.871, 0.719, 0.811, 0.75 (MW) | 4, 17, 27, 40, 48, 58, 65 | | LI = 9.837 | |
| | | | ΣDG = 9 | 0.996, 0.752, 0.645, 1.258, 0.496, 0.498, 0.406, 0.855, 0.73 (MW) | 4, 17, 27, 30, 35, 41, 50, 58, 65 | | LI = 9.769 | |
| [74] | ICM | IEEE 33 | ZONE-1 | 941.5 (kW) | 30 | PLR, VPE | 34.52 (kW) | WoDG |
| | | | ZONE-2 | 603.091 (kW) | 14 | | 4.82 (kW) | |
| | | | ZONE-3 | 691.607 (kW) | 25 | | 2.41 (kW) | |
| [76] | ABC | IEEE 39 | ΣDG = 1 | 2.5775 (MW) | 6 | PLR, VPE | 105.02(kW) | AM |
| | | | ΣDG = 2 | 1.9707, 0.5757 (MW) | 6, 15 | | 89.96(kW) | |
| | | | ΣDG = 3 | 1.7569, 0.5757, 0.7826 (MW) | 6, 15, 25 | | 79.25(kW) | |
| | | | ΣDG = 4 | 1.0765, 0.5757, 0.7824, 0.6538 (MW) | 6, 15, 25, 32 | | 66.58(kW) | |
| [77] | ABC | IEEE 33 | ΣDG = 1 | 2.3970 p.u | 5 | LLC, PLR, VPE | 0.06 p.u | WoDG |
| | | | ΣDG = 3 | 1.0645, 0.7322, 0.9823 (p.u.) | 29, 13, 23 | | 0.013 p.u | |
| | | | ΣDG = 5 | 1.39, 0.66, 0.52, 0.73,0.74 (p.u.) | 1, 24, 14, 7, 30 | | 0.007 p.u | |
| [78] | PSO | IEEE 12 | WTΣDG = 2 | 336 (kW) | 8 | PFO, PLR, VPE | 14.7 (kVA) | GA, AM, PSBIT |
| | | IEEE 15 | WTΣDG = (4 + 3) | 1176 (kW) | 4, 7 | | 54.8 (kVA) | |
| | | IEEE 33 | WTΣDG = (5 + 3 + 5 + 5) | 2873 (kW) | 7, 16, 24, 30 | | 127.1 (kVA) | |
| | | IEEE 69 | WTΣDG = (4 + 2 + 6 + 6 + 4) | 3696 (kW) | 7, 9, 48, 62, 64 | | 131.3 (kVA) | |
| | | IEEE 12 | SPV = (4 + 3) | 370 (kW) | 5, 10 | | 12..5 (kVA) | |
| | | IEEE 15 | SPV = (4 + 4 + 5 + 3 + 4 + 4) | 1267 (kW) | 5, 6, 8, 9, 12, 15 | | 50.9 (kVA) | |
| | | IEEE 33 | SPV = (6 + 6 + 3 + 6 + 6 + 6 + 6 + 6 + 6 + 6) | 2693 (kW) | 8, 13, 16, 17, 20, 24, 25, 27, 31, 33 | | 127 (kVA) | |
| | | IEEE 69 | SPV = (3 + 5 + 5 + 6 + 5 + 6 + 5 + 6 + 5 + 6) | 3062 (kW) | 6, 8, 12, 26, 52, 53, 59, 61, 62, 65, 68 | | 148.6 (kVA) | |
| [81] | CS | IEEE 38 | Wv1 = 1 & WL = Wv2 = 0 | 10, 20, 30, 40 (MW) | 18, 8, 27, 36 | PLR, VPE | | Binary GA, continuous GA, PSO, WoDG |
| | | | WL = 1 & Wv1 = Wv2 = 0 | 10, 20, 30, 40 (MW) | 24, 29, 7, 18 | | | |

**Table 2** (continued)

| Ref | Modern mathematical optimization method | Test systems | Different cases | Optimum DG sizes | Optimum DG locations | Strengthened parameters | Optimized results | Compared approaches |
|---|---|---|---|---|---|---|---|---|
| | | | Wv1 = 1 & WL = Wv1 = 0 | 10, 20, 30, 40 (MW) | 29, 8, 12, 15 | | | |
| | | | Wv1 = 0.5 & WL = Wv2 = 0.25 | 10, 20, 30, 40 (MW) | 34, 30, 16, 9 | | | |
| | | IEEE 69 | Wv1 = 1 & WL = Wv2 = 0 | 10, 20, 30, 40 (MW) | 30, 44, 59, 29 | | | |
| | | | WL = 1 & Wv1 = Wv2 = 0 | 10, 20, 30, 40 (MW) | 58, 51, 38, 19 | | | |
| | | | Wv1 = 1 & WL = Wv1 = 0 | 10, 20, 30, 40 (MW) | 32, 68, 63, 31 | | | |
| | | | Wv1 = 0.5 & WL = Wv2 = 0.25 | 10, 20, 30, 40 (MW) | 9, 67, 55, 40 | | | |
| [83] | FF | IEEE 69 | ΣDG = 1 | 1.8753 (MW) | 61 | PLR, VPE | 0.0832 (MW) | GA |
| | | | ΣDG = 2 | 1.7496, 0.9269 (MW) | 61, 67 | | 0.0747 (MW) | |
| [91] | HA | IEEE 9 | HL | 4, 2, 4, 4 (MVA) | 2, 4, 5, 8 | COP, PLR | 2.42 (MVA) | Bilateral contract scenario |
| | | | ML | 2, 4, 4 (MVA) | 4, 5, 8 | | | |
| | | | LL | 4, 1, 4, 4 (MVA) | 1, 3, 6, 7 | | | |
| | | | PSF = 0.7—1.66 | 4, 4, 4, 3 (MVA) | 2, 5, 8, 4 | | | |
| | | | PSF = 0.7—1.75 | 4, 4, 4, 3, 2, 1 (MVA) | 2, 5, 8, 4, 7, 1 | | | |
| [92] | GA | IEEE 16 | HL | 0.62 p.u | 7 | PLR | 5.76 p.u | WoDG |
| | | | ML | 0.53 p.u | 7 | | 2.63 p.u | |
| | | | LL | 0.4 p.u | 7 | | 1.7 p.u | |
| | | IEEE 37 | HL | 0.5 p.u | 15 | | 14.12 p.u | |
| | | | ML | 0.47 p.u | 15 | | 6.2 p.u | |
| | | | LL | 0.63 p.u | 15 | | 3.45 p.u | |
| | | IEEE 75 | HL | 0.45 p.u | 32 | | 29.93 p.u | |
| | | | ML | 0.63 p.u | 30 | | 13.06 p.u | |
| | | | LL | 0.59 p.u | 30 | | 8.09 p.u | |
| [93] | MVO | IEEE 33 | ΣDG = 1 | 2500 (kW) | 6 | VPE, PLR | 0.941 (p.u.), 111.01 (kW) | MFO, PSO, HPSO |
| | | | ΣDG = 2 | 852.06, 1156.95 (kW) | 13, 30 | | 0.9684 (p.u.), 87.16 (kW) | |
| [94] | MALO | IEEE 33 | ΣDG = 5 | 220, 50, 30, 20, 1020 (kW) | 5, 9, 12, 23, 30 | VPE, PLR | 0.083, 74.96 kW | WoDG |
| [95] | PSO | IEEE 33 | ΣDG = 5 (OIW) | 2.8171 (MW) each | 6, 10, 18, 22, 31 | VPE, PLR | 112.691 (kW) | WoDG |
| | | | ΣDG = 5 (RIW) | 2.9109 (MW) each | | | 112.558 (kW) | |
| | | | ΣDG = 5 (GLBIW) | 2.7947 (MW) each | | | 112.719 (kW) | |
| | | | ΣDG = 5 (TVIW) | 2.8201 (MW) each | | | 112.534 (kW) | |
| [96] | MCPSO | IEEE 30 | ΣDG = 1 | 46.95 (MW) | 23 | VPE, PLR | 12.93 (MW) | AM, PSO |
| | | IEEE 14 | ΣDG = 1 | 34 (MW) | 14 | | 10.093 (MW) | |

**Table 2** (continued)

| Ref | Modern mathematical optimization method | Test systems | Different cases | Optimum DG sizes | Optimum DG locations | Strengthened parameters | Optimized results | Compared approaches |
|---|---|---|---|---|---|---|---|---|
| [97] | PSO | IEEE 34 | ΣDG = 1 | 3 (MW) | 8 | VPE, PLR | 167 (kW) | WoDG |
| [98] | CPSO | IEEE 14 | ΣDG = 1 | 33.95 (MW) | 6 | VPE, PLR, COP | 10.81 (MW) | IA, ELF, PSO |

steps namely domination rank and crowding distance are involved in this technique. Out of two harmonies, better harmony is dominated by another harmony, non-dominated harmony is stored in harmony memory while the dominated harmony is not abandoned and got a second chance for extemporization [69].

#### 3.2.5 Simulated Annealing (SA)

In 1983, Kirkpatrick et al. have been researched the SA optimization technique having four steps namely concise configuration, random selection, determining function, and schedule of the annealing process. The authors demonstrated five DG applications to extract the various benefits of DG allocation in the distribution network by using simulated annealing optimization technique. The load flow analysis is used to calculate the voltage drops and power losses having MO function. This technique is evaluated on IEEE 33 test system and concluded the increased penetration of renewable energy in a radial power system. The results exhibited the optimal size and location of DG by considering the number of DG [70].

#### 3.2.6 Imperialist Competitive Method (ICM)

In [71, 72, 73, 74], an optimization technique has been implemented which is based on the nature of imperialistic nature of imperialist for the addiction of imperialism to make an empire. In the whole approach, the colonies approach toward their imperialist then the cost of each empire is calculated to interchange the positions of weaker one to the strongest one till only one empire is exist like the optimized value in DG. In continuation, the profit of distribution network operators is maximized and implemented in the UK under Ofgen which includes power flow constraints, operating limit, voltage profile, feeder capacity, and the number of DG steps with crossover probability and mutation probability function programmed in MATLAB [73]. In [74], losses are calculated by load flow technique in which KVL and KCL are used to measure the upstream voltage resulting in complex power measurement then ICM is implemented for DG optimization in zonal distribution network using islanding operation of the sensitive load.

#### 3.2.7 Ant Colony Optimization (ACO)

ACO technique has been characterized as; initialization, cost function calculation, constraint assignment, feasible state selection, and neighborhood structure with the properties like search, memory storage, feasible neighborhood, termination condition, probabilistic decision, ant routing table, and pheromone update. ACO technique is implemented for restructuring the distribution network framework to

minimize power losses. ACO exhibited the characteristic to evade premature stability and trapping in local optima. This technique is tested and implemented on IEEE 33 bus system and the results are significant and power losses are reduced in comparison to without restructuring the distribution network. The convergence time also indicates the faster solution by this technique. Moreover, network losses are reduced from 140 to 110 kW by reconfiguration of DG in the distribution system [75].

#### 3.2.8 Artificial Bee Colony (ABC)

Karboga presented an optimization technique that is based on the communal etiquette of honey bees and encompasses three key components namely food source, searchers, and non-searchers which are having self-organizational behavior depending on alternate feedback, random search, and sharing of information. A bi-level approach having ABC optimization technique is demonstrated on IEEE 33 system for DG allocation. In the first step, DG location is found and this process is followed by the DG size optimization considering 4 different cases to get the possible outcomes in less computation time [76]. Moreover, indexes are introduced in ABC optimization to improve the power quality parameters. In continuation, the authors attracted attention to the reformation of the industrial and commercial processes. The distribution loss index is included with the objective function because of its high presence in Iran [77].

#### 3.2.9 Particle Swarm Optimization (PSO)

In 1995, Kennedy et al. have been investigated PSO technique which is another social behavior simulation to optimization technique comprises the binds between artificial life to bird gathering, fish training and swarming behavior which require only ancient mathematical operators with less memory space and speed. The predecessor involves velocity matching, craziness, cornfield vector, auxiliary variable removal, multidimensional search, distance acceleration and present better version. The foundation of PSO having the pillars of social concept, swarm intelligence principles (proximate, quality, diverse response, stability, and adaptability) and computational behavior. In continuation author also discussed the various modified approach of PSO like multi-objective PSO (MPSO), constraint handling PSO, stretching PSO, cooperative PSO, comprehensive learning PSO and hybrid PSO. Tribal PSO (TPSO) with OO [37], constriction factor PSO [35] and MPSO [39, 78] are executed for renewable source-based DG optimization. Moreover, MPSO has been implemented for optimizing the power quality parameter with a constraint of expected rate of return (ERR) results in terms of total loss power index (TLPI). Synchronous compensator, synchronous generator and synchronous based DG units are used in the optimization process [79]. The continuous and discrete optimization by using N-R method is incorporated for DG optimization and penalty factor calculation [23].

#### 3.2.10 Cuckoo Search (CS)

The researchers Yang et al. have been projected an optimization technique build on the communal bearing of a cuckoo bird in addition to Levy flight conduct for better results due to a good combination of randomization and escalation having fewer constraints CS involves three steps namely cuckoo breeding behavior, Levy flight and search in which a cuckoo bird randomly search the nest, place its egg and if the entertainer bird did not find that egg fit for the nest then the entertainer bird will repudiate the cuckoo's egg. The renewable energy-based DG optimization has been done by the CS algorithm to get technical, financial, and societal advantages. These tests are performed on IEEE 22 and IEEE 69 bus systems having the mono, dual and multi-DG injection. The proposed technology results are efficaciously compared with the established PSO technique [80]. Authors recommended the CS for optimization of the objective function which includes the upgraded voltage profile and abridged power loss with different weightage factors namely; (WL), (WV1), and (WV2) [81]. Moreover, the importance of indices base-voltage evaluation is also highlight in the power quality. The outcomes are significant and compared with well-established GA and PSO. Consequently, the CS technique has the following benefits:

It has the property to cover more distance in step randomization.

- The requirement of optimization parameters is lesser as compared to other approaches so more adaptability is present among the researchers.
- It is inspired by the meta-population approach in which every nest provides a group of solutions.
- The tuned parameters are not dependent on the meeting frequency, so accuracy consideration is not a necessary step.
- It may lead to getting the high acceptance due to its simplicity and robust nature.

#### 3.2.11 Firefly Methods (FFM)

FFM is based on the gleaming etiquettes of a firefly in which three simple assumptions have been followed; all firefly is androgynous and each one produces a flare to attract others, the attraction force is proportional to the brightness and each firefly will attract toward the glower one and in the absence of brighter one then the firefly will take the random step in search space. A modified FFM is used for DG allocation for the active and reactive power compensation along with shunt capacitors. FFM optimization technique considered the four different compensation cases of the distribution network (absence of compensator, presence of shunt capacitor, presence of DG, and presence of integrated DG and shunt capacitor) and evaluated on IEEE 85 test system. The outcomes are significant as compared to PSO and MINLP [82]. Moreover, voltage profile and line losses are considered for the optimization of DG size and allocation. The evaluation has been done on IEEE 69, radial distribution system, and concluded with the results having the same efficacy as in GA [83].

### 3.3 Hybrid Approaches

In the optimal DG planning, heterogeneous hybrid AI techniques have been proposed by various researchers that are based on the amalgamation of two optimization approaches which work in progression mode or equidistant mode to solve more complex problems. In hybrid category GA is integrated with TS (GATS) [84], PSO (GAPSO) [85], OPF (GAOPF) [86] and fuzzy approach (FZ) [87]. A collateral approach GA and TS is evaluated for the optimization of objective function in the form of GA chromosomes and TS neighbors [84]. For the optimized power quality parameters GAPSO is adopted in which GA is used for the DG allocation and PSO for size optimization with 30 and 20 population size respectively [85]. In GAOPF the decision variables are found using GA and OPF is implemented for the optimum solution. In [88], jumping frog PSO (JFPSO) forwarded the DG location for the optimization of DG by OPF. The deterioration in voltage level has been considered in PSO process followed by the GSA in PSOGSA hybrid optimization technique [89] while the enhanced transient stability is achieved with the cascade process of PSO and shuffled frog leaping (SFL) using CCT index on Dlg SILENT software [17]. Consolidations of FZ with TS (FZTS) has been wielded for the MO, DG optimization [26]. Hybrid optimization techniques are reviewed and summarized in Table 3 based on adopted techniques, test systems, optimized results, and comparison with other optimization techniques.

## 4 Conclusions

It has been concluded that DG is advantageous for the improvisation of the power system characteristics. It provides numerous optimization techniques implemented in DG allocation with the purpose to meet the limitations. It also acts as a tool to augment the stability, reliability, and consistency of the distribution network. Moreover, it is also to analyze the classification of the intricacies for functioning optimization algorithms. The pictorial taxonomy of data-based assessment of DG optimization approaches are given in Fig. 4.

In this paper, various optimization techniques have been successfully reviewed along with comparative analysis of conventional, modern mathematical and hybrid methods utilized for significant parameters optimization of DG that aims at technological, financial, and environmental benefits with the optimized result parameters. Moreover, conventional approaches are simple, easy to execute and highly précised but agonized due to single objective slow convergence optimization while modern mathematical approaches can solve the multi-objective complex problems, with a smaller number of iterations having some limitations like harder to code, various settling parameters and impetuous convergence. Hybrid optimizations approaches can handle more complex problems with faster convergence but it may undergo complexity including a smaller number of literatures.

Reviewed studies indicate that the integration of renewable energy will maximize the benefits of DG planning in the distribution network but it is also implying the necessity of a reliable assessment tool for renewable energy. The intermittent nature of renewable energy is required to encounter and effective energy storage may lead to the elimination of intermittency.

Additionally, the demand response is also not significantly considered by the researchers. Due to the high cost of the energy storage system, demand response is also a major contributor to the smart distribution system. There is a possibility to develop a system that included the planning and optimal dispatch of renewable DG with energy storage and demand response.

**Table 3** Data-based assessment of DG optimization by hybrid approaches

| Ref | Hybrid optimization method | Test systems | Different cases | Optimum DG sizes | Optimum DG locations | Strengthened parameters | Optimized results | Compared approaches |
|---|---|---|---|---|---|---|---|---|
| [14] | FZ + AM | IEEE 12 | – | 0.22 (MW) | 9 | PLR, VPE | 0.01077 (MW) | WoDG |
| | | IEEE 33 | | 2.59 (MW) | 6 | | 0.111 (MW) | |
| | | IEEE 69 | | 1.87 (MW) | 61 | | 0.0832 (MW) | |
| [16] | GA + AM | IEEE 4 | Vdm = 0.007422 p.u | 4000, 3000 (kW) | 3,4 | COP, RI, VPE | 74.26 (kW) | WoDG |
| | | | Vdm = 0.007422 p.u | 3000, 2000 (kW) | 12, 7 | | 129.37 (kW) | |
| [20] | PSO + SFL | IEEE 33 | CL | 312.8, 334.4, 323, 279, 200 (kW) | 14, 16, 33, 8, 31 | PLR, VPE, TSI | 0.053273 (MW) | PSO, SFL |
| | | | MXL | 275.1, 252.3, 306.2, 237.6, 290 (kW) | 33, 12, 14, 26, 13 | | 0.055567 (MW) | |
| [21] | FZ + AM | IEEE 33 | VL > ± 5% | 2.4818 (MW) | 6 | PLR, VPE | 110.6318 (kW) | WoDG |
| | | | VL ± 5% | 3.15 (MW) | 6 | | 115.2 (kW) | |
| [26] | PSO + FZ | IEEE 33 | ΣDG = 1 | 1.2931 (MW) | 32 | PLR, VPE | 127.0919 (kW) | PSO |
| | | | ΣDG = 2 | 0.3836, 1.1506 (MW) | 32, 30 | | 117.3946 (kW) | |
| | | | ΣDG = 3 | 0.2701, 1.1138, 0.1503 (MW) | 32, 30, 31 | | 117.3558 (kW) | |
| | | | ΣDG = 4 | 0.2706, 0.8432, 0.1503, 0.5982 (MW) | 32, 30, 31, 18 | | 90.4794 (kW) | |
| [30] | MINLP + OPF | IEEE 41 | PP | 4.4, 1.1, 2.2 (MW) | 19, 23, 40 | PLR, VPE | 1079.7 (MW)h | WoDG |
| | | | TP | 6.6, 5.5, 9.9 (MW) | 19, 23, 40 | | 1527.2 (MW)h | |
| [37] | TPSO + OO | IEEE 86 | TRIBE PSO (SPV + BDG) | 200, 200 (kVA) | 61, 85 | COP | Total cost = 9,355,000 $ | TRIBE PSO, OO |
| | | | OO (SPV + BDG) | 400, 100 (kVA) | 72, 85 | | 9,504,000 $ | |
| | | | TRIBE PSO + OO (SPV + BDG) | 200, 200 (kVA) | 61, 82 | | 9,355,000$ | |
| [50] | NLP + OPF | IEEE 34 | ΣDG = 2 | 3112, 6.613 (MW) | 17, 18 | PLR, VPE | 0.279 (MW) | WoDG |
| [56] | GA + OPF | IEEE 69 | ΣDG = 3 | 2.661 (MW) | 26, 35, 62 | COP, PLR | TI (£/h) = 8.72 | |
| | | | ΣDG = 5 | 4.067 (MW) | 4, 26, 35, 40, 62 | | TI (£/h) = 10.73 | |
| | | | ΣDG = 7 | 4.566 (MW) | 5, 13, 27, 35, 40, 57, 65 | | TI (£/h) = 11.27 | |
| | | | ΣDG = 9 | 4.833 (MW) | 4, 6, 13, 21, 27, 35, 40, 57, 62 | | TI (£/h) = 11.51 | |
| [79] | PSO + CPF | IEEE 33 | ΣDG = 1 | 3.0317 (MW) | 12 | PLR | 70.949 (kW) | WoDG |
| | | | ΣDG = 2 | 0.9143, 1.5345 (MW) | 27, 22 | | 29.82 (kW) | |
| | | IEEE 69 | ΣDG = 1 | 2.2215 (MW) | 56 | | 23.594 (kW) | |
| | | | ΣDG = 2 | 0.6247, 2.1213 (MW) | 53, 56 | | 7.342 (kW) | |
| [84] | GA + TS | IEEE 13 | BDG | 300 (kW) | | PLR | 84.6 (kW) | GA |
| | | | WTDG | 200 (kW) | | | 83.5 (kW) | |
| | | | SPV | 80 (kW) | | | 95.2 (kW) | |
| | | IEEE 34 | BDG | 200 (kW) | | | 220.9 (kW) | |
| | | | WTDG | 200 (kW) | | | 195.7 (kW) | |
| | | | SPV | 200 (kW) | | | 141.4 (kW) | |

**Table 3** (continued)

| Ref | Hybrid optimization method | Test systems | Different cases | Optimum DG sizes | Optimum DG locations | Strengthened parameters | Optimized results | Compared approaches |
|---|---|---|---|---|---|---|---|---|
| [85] | GA + PSO | IEEE 33 | ΣDG = 4 | 0.6639, 0.6628, 1.0232, 0.8671 (MW) | 32, 14, 24, 26 | PLR | 0.0682 p.u | GA, PSO |
| [86] | GA + OPF | IEEE 9 | UFC | 2 (4 (MVA) & 1 (MVA)) | 4, 8 | COP, PLR | 52.73 GWh | MINLP |
| | | | LFC | 1 (4 & 1), 2 (2 & 1), 1 (1 & 1) (MVA) | 2, 4, 7, 1 | | 45.35 GWh | |
| | | | CAC | 3 (4 & 1), 1 (3 & 1) (MVA) | 2, 4, 8, 1 | | 31.55 GWh | |
| [87] | GA + FZ | IEEE 6 | ΣDG = 1 | 8.5 (MW) | 3 | COP, PLR | 0.112 (MW) | WoDG |
| | | IEEE 30 | ΣDG = 2 | 40, 35 (MW) | 10, 6 | | | |
| [88] | PSO + OPF | IEEE 30 | ΣDG = 3 | 9.75, 9.26, 8.81 (MW) | 19, 24, 30 | COP, PLR | 11.05 (MW) | |
| | | | ΣDG = 5 | 9.95, 7.85, 6.23, 2.01, 7.75 (MW) | 7, 19, 24, 26, 30 | | 10.92 (MW) | |
| | | | ΣDG = 7 | 8.78, 7.15, 2, 2, 4.52, 2.01, 7.73 (MW) | 7, 19, 21, 23, 24, 26, 30 | | 10.91 (MW) | |
| | | | ΣDG = 9 | 7.78, 5.34, 2, 2, 2, 2, 3.67, 2, 7.21 (MW) | 7, 18, 19, 21, 22, 23, 24, 26, 30 | | 10.9 (MW) | |
| [89] | PSO + GSA | IEEE 69 | Test Case 1 (Tiv) | 0.2, 1, 1.8 (MW) | 21, 49, 61 | LLC, PLR, PR, VPE | MPI = 0.5463 | PSO, GSA |
| | | | Test Case 1 (Tv) | 0.9, 0.2, 1.3 (MW) | 4, 21, 61 | | MPI = 0.6014 | |
| | | | Test Case 2 (Tiv) | 0.3, 1.2, 1.8 (MW) | 21, 49, 61 | | MPI = 0.4495 | |
| | | | Test Case 2 (Tv) | 1.3, 1.8, 0.2 (MW) | 21, 49, 61 | | MPI = 0.4909 | |
| | | | Test Case 3 (Tiv) | 1.7, 0.8, 0.8 (MW) | 3, 60, 61 | | MPI = 0.6888 | |
| | | | Test Case 3 (Tv) | 1.6, 0.8, 0.3 (MW) | 3, 61, 64 | | MPI = 0.7131 | |
| | | | Test Case 3 (Tiv) | 0.3, 0.2, 1 (MW) | 21, 61, 48 | | MPI = 0.4771 | |
| | | | Test Case 3 (Tv) | 0.6, 1.5, 0.6 (MW) | 50, 61, 47 | | MPI = 0.5282 | |
| [99] | ABC + TLBO | IEEE 33 | WTΣDG-1 | 2558.5 (kW) | 6 | VPE, PLR, COP, | 67.83 (kW), 35,971$ | EA, GA, PSO |
| | | | WTΣDG-2 | 858.3, 1089.1 (kW) | 13, 30 | | 28.63 (kW), 15,049$ | |
| | | | WTΣDG-3 | 1069.9, 1029.9, 793.8 (kW) | 13, 30, 24 | | 11.74 (kW), 6171$ | |
| | | | SPV ΣDG-1 | 2590.2 (kW) | 6 | | 111.027 (kW), 58,536$ | |
| | | | SPV ΣDG-2 | 851.5, 1157.6 (kW) | 13, 30 | | 87.16 (kW), 45,814$ | |
| | | | SPV ΣDG-3 | 801.7, 1053.6, 1091.3 (kW) | 13, 30, 24 | | 72.78 (kW), 38,256$ | |
| [100] | ALO + PSO + FLC | IEEE 33 | SPV ΣDG = 2, UPF | 385, 2154 (kW) | 32, 7 | PLR, COP | 90.98 (kW), 12,062 $ | PSO, ALO + PSO |
| | | | WT ΣDG = 2, UPF | 951, 696 (kW) | 31, 17 | | 89.3 (kW), 8496 $ | |

**Table 3** (continued)

| Ref | Hybrid optimization method | Test systems | Different cases | Optimum DG sizes | Optimum DG locations | Strengthened parameters | Optimized results | Compared approaches |
|---|---|---|---|---|---|---|---|---|
| | | | SPV ΣDG = 2, LaPF = 0.85 | 924 + j1223, 665 + j710 (kVA) | 32, 14 | | 47.6 (kW), 8037 $ | |
| | | | WT ΣDG = 2, LaPF = 0.85 | 993 + j1667, 606 + j913.5 (kVA) | 8, 30 | | 35.5 (kW), 8038 $ | |
| | | | SPV ΣDG = 2, LePF = 0.85 | 1669-j413, 1179-j29 (kVA) | 7, 30 | | 116 (kW), 13,710 $ | |
| | | | WT ΣDG = 2, LePF = 0.85 | 2300 + j273, 934-j38 (kVA) | 4, 30 | | 109 (kW), 15,611 $ | |
| [101] | GOA + CS | IEEE 33 | ΣDG = 1, Half load = 50% | 1716 (kW) | 2 | VPE, PLR, COP | 123.62 (kW), 34.33$ | GA, PSO, GOA, CS |
| | | IEEE 33 | ΣDG = 1, Full load | 926.99 (kW) | 24 | | 139.59 (kW), 18.5$ | |
| | | IEEE 33 | ΣDG = 1, 150% load | 926.99 (kW) | 24 | | 139.59 (kW), 18.5$ | |
| | | IEEE 69 | ΣDG = 1, Half load = 50% | 1930.7 (kW) | 17 | | 141.43 (kW), 38.6 $ | |
| | | IEEE 69 | ΣDG = 1, Full load | 1990.7 (kW) | 6 | | 147.66 (kW), 39.8$ | |
| | | IEEE 69 | ΣDG = 1, 150% load | 1890.6 (kW) | 12 | | 151.66 (kW), 37.8$ | |

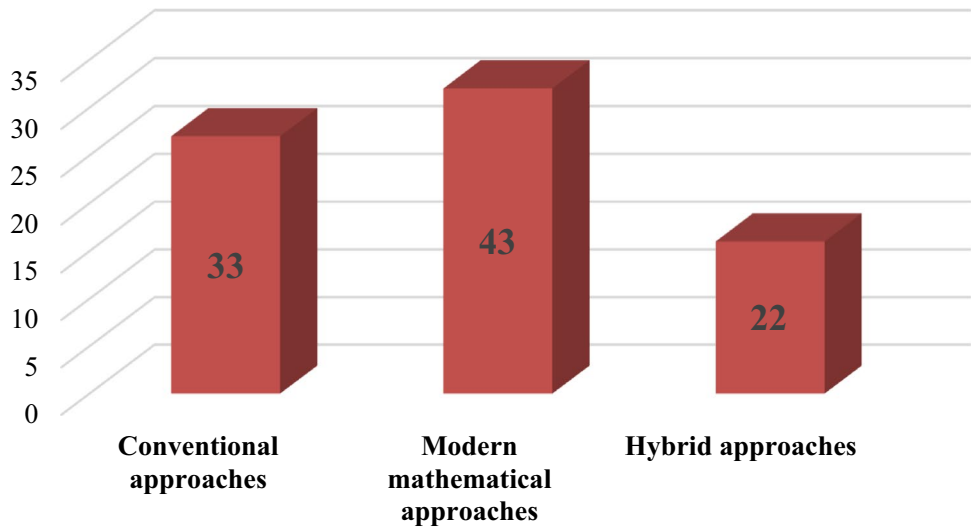



**Fig. 4** Taxonomy for data-based assessment of DG optimization approaches

## Declarations

**Conflict of interest** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

1. El-Khattam W, Salama MMA (2004) Distributed generation technologies, definitions and benefits. Electr Power Syst Res 71:119–128
2. Saxena V, Kumar N, Nangia U, Singh J (2021) Analysis of Smart Electricity Grid Framework Unified with Renewably Distributed Generation. In: Agrawal R, Singh CK, Goyal A (eds) Advances

in Smart Communication and Imaging Systems Lecture Notes in Electrical Engineering. Springer, Singapore
3. Saxena V, Kumar N, Nagia U (2021) Smart Grid: A Sustainable Smart Approach. Journal of Physics *Conf Ser.* **2007(1):** 012042
4. Gonen T. (2014) Electrical power transmission system engineering: analysis and design. BocaRaton, Florida: T & F Group, LLC.
5. Muttaqi KM, Le AD, Negnevitsky M, Ledwich G (2014) An algebraic approach for determination of dg parameters to support voltage profiles in radial distribution networks. IEEE Trans Smart Grid 5:1351–1360
6. Arya LD, Koshti A, Choube SC (2012) Distributed generation planning using differential evolution accounting voltage stability consideration. Int J Electr Power Energy Syst 42:196–207
7. Juanuwattanakul P (2012) Masoum M (2012) increasing distributed generation penetration in multiphase distribution networks considering grid losses, maximum loading factor and bus voltage limits. Gener Transm Distrib IET 6:1262–1271
8. Al Abri R, El-Saadany EF, Atwa YM (2013) Optimal placement and sizing method to improve the voltage stability margin in a distribution system using distributed generation. Power Syst IEEE Trans 28:326–334
9. Murthy VVSN, Kumar A (2013) Comparison of optimal dg allocation methods in radial distribution systems based on sensitivity approaches. Int J Electr Power Energy Sys 53:450–467
10. Biswas S, Goswami SK, Chatterjee A (2012) Optimum distributed generation placement with voltage sag effect minimization. Energy Convers Manag 53:163–174
11. Liu Z, Wen F, Ledwich G, Ji X (2011) Optimal sitting and sizing of distributed generators based on a modified primal-dual interior point algorithm. In: Electric Utility Deregulation and Restructuring and Power Technologies (DRPT). Proc of the 4th Int Conf on IEEE 1360–1365.
12. Alinejad-Beromi Y, Sedighizadeh M, Bayat M, Khodayar M (2007) Using genetic alghoritm for distributed generation allocation to reduce losses and improve voltage profile. in: 42nd International Universities Power Engineering Conference, 2007, pp. 954–959.
13. Sedighizadeh M, Rezazadeh A (2008) Using genetic algorithm for distributed generation allocation to reduce losses and improve voltage profile. World Acad Sci Eng Technol 37:251–256
14. Injeti SK, Kumar NP (2011) Optimal planning of distributed generation for improved voltage stability and loss reduction. Int J Comput Appl 15:40–46
15. Hengsritawat V, Tayjasanant T, Nimpitiwan N (2012) Optimal sizing of photovoltaic distributed generators in a distribution system with consideration of solar radiation and harmonic distortion. Int J Electr Power Energy Syst 39:36–47
16. Borges CLT, Falcão DM (2006) Optimal distributed generation allocation for reliability, losses, and voltage improvement. Int J Electr Power Energy Syst 28:413–420
17. Nayeripour M, Mahboubi-Moghaddam E, Aghaei J, Azizi-Vahed A (2013) Multi-objective placement and sizing of DGs in distribution networks ensuring transient stability using hybrid evolutionary algorithm. Renew Sustain Energy Rev 25:759–767
18. Singh RK, Goswami SK (2010) Optimum allocation of distributed generations based on nodal pricing for profit, loss reduction, and voltage improvement including voltage rise issue. Int J Electr Power Energy Syst 32:637–644
19. Kim JO, Park SK, Park KW, Singh C (1998) Dispersed generation planning using Improved Hereford Ranch Algorithm. in: IEEE World Congress on Computational Intelligence (Cat. No.98TH8360), 1998, pp. 678–683.
20. Lee S-H, Park J-W (2009) Selection of optimal location and size of multiple distributed generations by using kalman filter algorithm. Power Syst IEEE Trans 24:1393–1400
21. Jamil M, Kirmani S (2012) Optimal allocation of SPV based DG system for loss reduction and voltage improvement in radial distribution systems using approximate reasoning. In: IEEE 5th India International Conference on Power Electronics (IICPE), 2012, pp. 1–5.
22. Abu-Mouti F, El-Hawary M (2011) Heuristic curve-fitted technique for distributed generation optimisation in radial distribution feeder systems. IET Gener Transm Distrib 5:172–180
23. AlRashidi M, AlHajri M (2011) Optimal planning of multiple distributed generation sources in distribution networks: a new approach. Energy Convers Manag 52:3301–3308
24. Gözel T, Hocaoglu MH (2009) An analytical method for the sizing and siting of distributed generators in radial systems. Electr Power Syst Res 79:912–918
25. Acharya N, Mahat P, Mithulananthan N (2006) An analytical approach for dg allocation in primary distribution network. Int J Electr Power Energy Syst 28:669–678
26. Lalitha MP, Reddy V, Usha V, Reddy NS (2010) Application of fuzzy and pso for dg placement for minimum loss in radial distribution system. ARPN J Eng Appl Sci 5:32–37
27. Abu-Mouti FS, El-Hawary ME (2011) Optimal distributed generation allocation and sizing in distribution systems via artificial bee colony algorithm. IEEE Trans Power Deliv 26:2090–2101
28. Khatod DK, Pant V, Sharma J (2013) Evolutionary programming based optimal placement of renewable distributed generators. IEEE Trans Power Syst 28:683–695
29. Hung DQ, Mithulananthan N, Lee KY (2014) Optimal placement of dispatchable and nondispatchable renewable dg units in distribution networks for minimizing energy loss. Int J Electr Power Energy Syst 55:179–186
30. Atwa Y, El-Saadany E (2011) Probabilistic approach for optimal allocation of wind-based distributed generation in distribution systems. IET Renew Power Gener 5:79–88
31. Abdelsalam AA, El-Saadany EF (2013) Probabilistic approach for optimal planning of distributed generators with controlling harmonic distortions. Gener Transm Distrib IET 7:1105–1115
32. Al-Sabounchi A, Gow J, Al-Akaidi M, Al-Thani H (2011) Optimal sizing and location of a PV system on three-phase unbalanced radial distribution feeder avoiding reverse power flow. in: IEEE Electrical Power and Energy Conference, 2011, pp. 74-79.
33. Al-Sabounchi A, Gow J, Al-Akaidi M (2014) Simple procedure for optimal sizing and location of a single photovoltaic generator on radial distribution feeder. Renew Power Gener IET 8:160–170
34. Hung DQ, Mithulananthan N, Bansal R (2014) Integration of pv and bes units in commercial distribution systems considering energy loss and voltage stability. Appl Energy 113:1162–1170
35. Jain N, Singh S, Srivastava S (2013) A generalized approach for dg planning and viability analysis under market scenario. Ind Electron IEEE Trans 60:5075–5085
36. Rider MJ, López-Lezama JM, Contreras J, Padilha-Feltrin A (2013) Bilevel approach for optimal location and contract pricing of distributed generation in radial distribution systems using mixed-integer linear programming. Gener Transm Distrib IET 7:724–734
37. Zou K, Agalgaonkar AP, Muttaqi KM, Perera S (2012) Distribution system planning with incorporating dg reactive capability and system uncertainties. Sustain Energy IEEE Trans 3:112–123
38. Phonrattanasak P, Miyatake M, Sakamoto O (2013) Optimal location and sizing of solar farm on japan east power system using multiobjective bees algorithm. in: IEEE Energytech, 2013, pp. 1-6.
39. Ameli A, Bahrami S, Khazaeli F, Haghifam M-R (2014) A multiobjective particle swarm optimization for sizing and placement of dgs from dg owner's and distribution company's viewpoints. IEEE Trans Power Deliv 29:1831–1840

40. Liew PY, Theo WL, Alwi SRW, Lim JS, Manan ZA, Klemeš JJ (2016) Total site heat integration planning and design for industrial, urban and renewable systems. Renew Sustain Energy Rev. https://doi.org/10.1016/j.rser.2016.05.086
41. Cao Y, Wang X, Li Y, Tan Y, Xing J, Fan R (2016) A comprehensive study on low- carbon impact of distributed generations on regional power grids: a case of Jiangxi provincial power grid in China. Renew Sustain Energy Rev 53:766–778
42. Shaaban H, Elsherif A (2015) Location and Sizing of Distributed Generation Units in Primary Distribution Networks for Loss Reduction and Voltage Enhancement. in: 17th International Middle-East Power SystemsConference (MEPCON'15) Mansoura University, Egypt. 2015, pp. 35–44.
43. Caisheng W, Nehrir MH (2004) Analytical approaches for optimal placement of distributed generation sources in power systems. IEEE Trans Power Syst 19:2068–2076
44. Maurya H, Pal NS, Bhadoria VS (2020) Performance Analysis of Optimal Placement of Multiple DGs Using PSO. Decision Analytics Applications in Industry. Springer, Singapore
45. Prenc R, Skrlec D, Komen V (2013) Optimal PV system placement in a distribution network on the basis of daily power consumption and production fluctuation. in: Eurocon 2013, pp. 777-783.
46. Rotaru F, Chicco G, Grigoras G, Cartina G (2012) Two-stage distributed generation optimal sizing with clustering-based node selection. Int J Electr Power Energy Syst 40:120–129
47. Pesaran HAM, Mohd Zin AA, Khairuddin A, Shariati O (2014) Optimal sizing and siting of distributed generators by a weighted exhaustive search. Electr Power Compon Syst 42:1131–1142
48. Keane A, O'Malley M (2007) Optimal utilization of distribution networks for energy harvesting. IEEE Trans Power Syst 22:467–475
49. Abou El-Ela AA, Allam SM, Shatla MM (2010) Maximal optimal benefits of distributed generation using genetic algorithms. Electr Power Syst Res 80:869–877
50. Esmaili M (2013) Placement of minimum distributed generation units observing power losses and voltage stability with network constraints. IET Gener Transm Distrib 7:813–821
51. Porkar S, Poure P, Abbaspour-Tehrani-fard A, Saadate S (2010) A novel optimal distribution system planning framework implementing distributed generation in a deregulated electricity market. Electr Power Syst Res 80:828–837
52. Wang C, Lei S (2020) MDP-based distribution network reconfiguration with renewable distributed generation: approximate dynamic programming approach. IEEE Trans on Smart Grid 11(4):3620–3631
53. Vovos PN, Bialek JW (2005) Direct incorporation of fault level constraints in optimal power flow as a tool for network capacity analysis. IEEE Trans Power Syst 20:2125–2134
54. Vovos PN, Harrison GP, Wallace AR, Bialek JW (2005) Optimal power flow as a tool for fault level-constrained network capacity analysis. IEEE Trans Power Syst 20:734–741
55. Gautam D (2007) Mithulananthan N (2007) Optimal DG placement in deregulated electricity market. Electr Power Syst Res 77:1627–1636
56. Harrison GP, Piccolo A, Siano P, Wallace AR (2008) Hybrid GA and OPF evaluation of network capacity for distributed generation connections. Electr Power Syst Res 78:392–398
57. Jabr RA, Pal BC (2009) Ordinal optimisation approach for locating and sizing of distributed generation. Gener Transm Distrib IET 3:713–723
58. Elmitwally A (2013) A new algorithm for allocating multiple distributed generation units based on load centroid concept. Alex Eng J 52:655–663
59. Hedayati H, Nabaviniaki SA, Akbarimajd A (2008) A method for placement of dg units in distribution networks. IEEE Trans Power Deliv 23:1620–1628
60. Lai LL, Ma JT (1997) Application of evolutionary programming to reactive power planning-comparison with nonlinear programming approach. IEEE Trans Power Syst 12:198–206
61. Silvestri A, Berizzi A, Buonanno S (1999) Distributed generation planning using genetic algorithms. in: PowerTech Budapest 99. Abstract Records. (Cat. No.99EX376), 1999, pp. 257.
62. Singh D, Verma KS (2009) Multiobjective optimization for dg planning with load models. IEEE Trans Power Syst 24:427–436
63. Kumar V, Kumar HCR, Gupta I, Gupta HO (2010) Dg integrated approach for service restoration under cold load pickup. IEEE Trans Power Deliv 25:398–406
64. Jen-Hao T, Tain-Syh L, Yi-Hwa L (2002) Strategic distributed generator placements for service reliability improvements. Power Engg Society Summer Meet, IEEE 2:719–724
65. Jin T, Tian Y, Zhang CW, Coit DW (2013) Multicriteria planning for distributed wind generation under strategic maintenance. Power Deliv IEEE Trans 28:357–367
66. Nara K, Hayashi Y, Ikeda K, Ashizawa T (2001) Application of tabu search to optimal placement of distributed generators. IEEE Power Engg Society Winter Meet 2:918–923
67. Golshan MEH, Arefifar SA (2006) Distributed generation, reactive sources and network configuration planning for power and energy-loss reduction. IEE Proc Gener Transm Distrib 153:127–136
68. Geem ZW, Kim JH, Loganathan G (2001) A new heuristic optimization algorithm: harmony search. SIMULATION 76:60–68
69. Nekooei K, Farsangi MM, Nezamabadi-Pour H, Lee KY (2013) An improved multiobjective harmony search for optimal placement of dgs in distribution systems. Smart Grid IEEE Trans 4:557–567
70. Dharageshwari K and Nayanatara C (2015) Multiobjective optimal placement of multiple distributed generations. in: International Conference on Circuits, Power, and Computing Technologies [ICCPCT-2015], pp. 1-7.
71. Atashpaz-Gargari E, Lucas C (2007) Imperialist competitive algorithm: An algorithm for optimization inspired by imperialistic competition. in: IEEE Congress on Evolutionary Computation, 2007, pp. 4661-4667.
72. Jahani R, Hosseinzadeh A, Gharegozi A, Abadi MM (2011) Ica-based allocation of dgs in a distribution system. Am J Sci Res 33:64–75
73. Soroudi A, Ehsan M (2012) Imperialist competition algorithm for distributed generation connections. Gener Transm Distrib IET 2012(6):21–29
74. Rahmatian M, Ebrahimi E, Ghanizadeh AJ, Gharehpetian GB (2012) Optimal sitting and sizing of DG units considering islanding operation mode of sensitive loads. in: Iranian Conference on Smart Grids, 2012, pp. 1-5.
75. Yang J (2020) Research on Optimized Reconfiguration of Distributed Distribution Network Based on Ant Colony Optimization Algorithm. in: International Conference on Computer Engineering and Application (ICCEA), 2020, pp. 20-23.
76. Lalitha MP, Reddy NS, Reddy VV (2010) Optimal dg placement for maximum loss reduction in radial distribution system using ABC algorithm. International Journal of Reviews in Computing, 3:44-52.
77. Sohi MF, Shirdel M, Javidaneh A (2011) Applying BCO algorithm to solve the optimal DG placement and sizing problem. in: 5th International Power Engineering and Optimization Conference, 2011, pp. 71-76
78. Kayal P, Chanda C (2013) Placement of wind and solar based dgs in distribution system for power loss minimization and

voltage stability improvement. Int J Electr Power Energy Syst 53:795–809
79. Hien NC, Mithulananthan N, Bansal RC (2013) Location and sizing of distributed generation units for loadabilty enhancement in primary feeder. IEEE Syst Jour 7:797–806
80. Chowdhury A, Roy R and Mandal K K (2020) Optimal allocation of wind based DG for enhancement of technical, economic & social benefits u sing Cuckoo Search Algorithm for radial distribution networks. in: International Conference on Computer, Electrical & Communication Engineering (ICCECE), 2020, pp. 1-6.
81. Moravej Z, Akhlaghi A (2013) A novel approach based on cuckoo search for dg allocation in distribution network. Int J Electr Power Energy Syst 44:672–679
82. Sukraj K, Yuvaraj T, Hariharan R and Thirumalai M (2019) Simultaneous Allocation of Shunt Capacitor and Distributed Generator in Radial Distribution Network Using Modified Firefly Algorithm. in: International Conference on Smart Structures and Systems (ICSSS), 2019, pp. 1-5.
83. Sulaiman MH, Mustafa MW, Azmi A, Aliman O, Abdul Rahim SR (2012) Optimal allocation and sizing of distributed generation in distribution system via firefly algorithm. in: IEEE International Power Engineering and Optimization Conference Melaka, Malaysia, 2012, pp. 84-89.
84. Gandomkar M, Vakilian M, Ehsan M (2005) A genetic-based tabu search algorithm for optimal dg allocation in distribution networks. Electr Power Compon Syst 33:1351–1362
85. Moradi MH, Abedini M (2012) A combination of genetic algorithm and particle swarm optimization for optimal dg location and sizing in distribution systems. Int J Electr Power Energy Syst 34:66–74
86. Naderi E, Seifi H, Sepasian MS (2012) A dynamic approach for distribution system planning considering distributed generation. IEEE Trans Power Deliv 27:1313–1322
87. Kyu-Ho K, Yu-Jeong L, Sang-Bong R, Sang-Kuen L, Seok-Ku Y (2002) Dispersed generator placement using fuzzy-ga in distribution systems. IEEE Power Eng Soc Summer Meeting 3:1148–1153
88. Gomez-Gonzalez M, López A, Jurado F (2012) Optimization of distributed generation systems using a new discrete PSO and OPF. Electr Power Syst Res 84:174–180
89. Tan WS, Hassan MY, Rahman HA, Abdullah MP, Hussin F (2013) Multi-distributed generation planning using hybrid particle swarm optimisation-gravitational search algorithm including voltage rise issue. IET Gen Transm Distrib 7:929–942
90. Ghosh S, Ghoshal SP, Ghosh S (2010) Optimal sizing and placement of distributed generation in a network system. Int J Electr Power Energy Syst 32:849–856
91. El-khattam W, Bhattacharya K, Hegazy Y, Salama MMA (2004) Optimal investment planning for distributed generation in a competitive electricity market. IEEE Trans Power Syst 19:1674–1684
92. Singh D, Singh D, Verma K (2007) Ga based optimal sizing and placement of distributed generation for loss minimization. Int J Electr Comput Eng 2:556–562
93. Roy K, Srivastava L, Dixit S (2020). Optimal Placement and Sizing of Distributed Generation Using Multi-Verse Optimization. in: 12th International Conference on Computational Intelligence and Communication Networks (CICN), 2020, pp. 268-272.
94. Liu H, Xu L, Zhang C (2019). Optimal allocation of distributed generation based on multi-objective ant lion algorithm. in: IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia), 2019, pp. 1455-1460.
95. Maurya H et al (2019) Effect of Inertia Weight of PSO on Optimal Placement of DG. IOP Conf Ser Mater Sci Eng 594:012011
96. Bhadoria VS, Jaiswal SP, Pal NS, Shrivastava V (2022) Application of modified clonal PSO in distributed generator placement for enhancement of efficiency and voltage stability in distribution system. Lecture Notes in Electrical Engineering, vol 766:703-718. Springer, Singapore.
97. Bhadoria VS, Jaiswal SP, Pal NS, Shrivastava V (2017) Reliability improvement of distribution system by optimal sitting and sizing of disperse generation. Int J Reliability, Quality Safety Engg 24(06):1740006
98. Bhadoria VS, Jaiswal SP, Pal NS, Shrivastava V (2019) Artificial immune system-based approach for size and location optimization of distributed generation in distribution system. Int J Syst Assur Eng Manag 10:339–349
99. Khasanov M, Kamel S, Xie K (2019). Allocation of Distributed Generation in Radial Distribution Networks Using an Efficient Hybrid Optimization Algorithm**.** in: IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia), 2019, pp. 1300-1305.
100. Samala RK, Kotapuri MR (2020) Optimal allocation of distributed generations using hybrid technique with fuzzy logic controller radial distribution system. SN Appl Sci 2:191
101. Suresh MCV, Edward JB (2020) A hybrid algorithm based optimal placement of DG units for loss reduction in the distribution system. Appl Soft Comput 91:106191

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

Springer Nature or its licensor 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.

*2022 2nd International Conference on Intelligent Technologies (CONIT)*
*Karnataka, India. June 24-26, 2022*

# An Iterative Method for Image Denoising Based on Ramanujan Sums

Priyansh Jaiswal
*Department of Electrical Engineering*
*Delhi Technological University*
New Delhi, India
priyanshjaiswal2310@gmail.com

Noman Ahmed Ansari
*Department of Electrical Engineering*
*Delhi Technological University*
New Delhi, India
noman.ahmad0808@gmail.com

Anup Kumar Mandpura
*Department of Electrical Engineering*
*Delhi Technological University*
New Delhi, India
amandpura@gmail.com

Neeta Pandey
*Department of Electronics and Communication*
*Delhi Technological University*
New Delhi, India
neetapandey@dce.ac.in

2022 International Conference on Intelligent Technologies (CONIT) | 978-1-6654-8407-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CONIT55038.2022.9847857

***Abstract*—The suppression of noise in the received noisy image is termed as image denoising. Recently Ramanujan Sums has garnered a lot of attention in signal processing applications due to its attributes such as: integer valued, symmetry and periodicity. This paper presents an iterative image denoising technique based on Ramanujan Sums. Further edge detection using the Ramanujan Sums is performed in conjunction to image denoising to improve the over-all image quality. The proposed iterative image denoising algorithm has better structural similarity index (SSI) in comparison to a 3X3 averaging filter. Simulation results also show that the residual error in each iteration decreases with increase in iterations.**

***Index Terms*—Ramanujan sums, denoising, Mean Squared Error, 3X3 Average Filter, Edge Detection.**

## I. Introduction

Denoising is the method of noise elimination from the source signal. Denoising an image is a tedious task when the noise also affects the high-frequency component of the image. Achieving a balance between suppressing as much noise as possible without losing too much information is the main purpose of denoising. Image processing finds applications in various fields such as: medical images, computer vision, pattern recognition. Various image denoising techniques including spatial domain, frequency domain and machine learning based approach have been studied in [1]–[5]. Recently the Ramanujan sum (RS) has garnered a lot of attention in various domains of signal processing due to its attributes such as: integer valued, symmetricity and periodicity making its way in numerous applications ranging from signal processing to biomedical field [6]. Further the authors in [7] extended the application of the properties of Ramanujan subspace for representing signal along with identifying periodic components. The authors in [8] presented application of Ramanujan sums from simple period estimation to complex biomedical applications. In [9], the authors performed hardware implementation of three different Sparse Ramanujan sequences (SRS) based transforms. The resource utilization of the SRS based transforms is compared with discrete cosine transform (DCT) and discrete wavelet transform (DWT). The authors in [9] concluded that SRS transforms are hardware efficient than DCT and DWT. In [10], the authors used the Ramanujan periodic transform (RPT) for denoising a mixed sine wave signal. The denoising performance using RPT was found to be better in comparison to denoising using discrete Fourier transform (DFT). The authors in [11] defined a class of Ramanujan operators that has attributes of first order derivative and second order derivative and demonstrates the use of the aforementioned properties in detection of edges and estimation of noise. In this paper we propose an iterative method based on Ramanujan sums for image denoising. The proposed method has a smaller residual error and better structural similarity index in comparison to a simple averaging filter.

Further in Section 2 we present the mathematical preliminaries and properties of Ramanujan Sums. Then in section 3 we present the proposed iterative denoising algorithm using Ramanujan Sums. Results and discussions are presented in Section 4. Lastly, Conclusion is presented in Section 5.

## II. Mathematical Preliminaries

### A. Ramanujan Sums

The Ramanujan sums, for q $\geq$ 1 can be expressed as: [6]

$$c_q(n) = \sum_{k=1,(k,q)=1}^{q} e^{j2\pi knq} = \sum_{k=1,(k,q)=1}^{q} W_q^{-kn}, \qquad (1)$$

where, k and q are co-prime.

From (1) it is inferred that the sum $c_q(n)$ is periodic in nature with period $q$.

As $W_q^{-(q-k)} = W_q^k = (W_q^{-k})^*$ it can be thereby concluded that $c_q(n)$ has real-valued nature. Now, $c_q(n)$ can also be described as follows:



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:59:55 UTC from IEEE Xplore. Restrictions apply.

TABLE I: RAMANUJAN COEFFICIENTS FOR FIRST TEN VALUES OF q

| | | | |
|---|---|---|---|
| $c_1$ | 1 | $c_6$ | 2, 1,-1 ,-2, -1, 1 |
| $c_2$ | 1,-1 | $c_7$ | 6, -1,-1, -1,-1, -1,-1 |
| $c_3$ | 2,-1,-1 | $c_8$ | 4, 0 ,0,0, -4,0, 0,0 |
| $c_4$ | 2,0,-2,0 | $c_9$ | 6, 0, 0, -3,0, 0,-3, 0,0 |
| $c_5$ | 4, -1, -1,-1 ,-1 | $c_{10}$ | 4, 1,-1,1, -1,-4,-1, 1,-1, 1 |

$$c_q(n) = \sum_{k=1,(k,q=1)}^{q} cos(2\pi kn)/q \tag{2}$$

So summing up $c_q(n)$ can be written in these forms:

$$c_q(n) = \sum_{k=1,(k,q)=1}^{q} W_q^{kn} = \sum_{k=1,(k,q)=1}^{q} W_q^{-kn} = \sum_{k=1,(k,q=1)}^{q} cos(2\pi kn)/q \tag{3}$$

So from (2) it is inferred that the sums are real and symmetric, that is $c_q(n)$ = $c_q(-n)$, or which can be seen in a view of periodicity as,

$$c_q(n) = c_q(q-n) \tag{4}$$

Table I shows the Ramanujan coefficients $c_q(n)$ for first ten values of $q$.

### *B. Derivative Properties*

Ramanujan operator $\hat{R}_q$ is defined for $q$ ($q \neq 1$) as a linear and shift invariant operator whose kernel $\hat{c}_q(n)$ is given as one period of $c_q(n)$ [11]. Mathematically, $\hat{c}_q(n)$ can be expressed as:

$$\hat{c}_q(n) = \begin{Bmatrix} c_q(n),\ 0 \leq n \leq q-1 \\ 0,\ otherwise \end{Bmatrix} \tag{5}$$

For example the kernel for $q = 2$ and $q = 3$ are given by:

$$\hat{c}_2 = (1, -1),\ \hat{c}_3 = (2, -1, -1)$$

In Equation (6), the application of $\hat{R}_q$ on a signal $x(n)$ is mathematically expressed as:

$$(\hat{R}_q x)(n) \triangleq \sum_k \hat{c}_q(k) x(n-k) \tag{6}$$

where $\hat{R}_q$ has the properties of the first derivative.

Ramanujan operator $\tilde{R}_q$ defined for odd $q$ ($q \neq 1$), have been shown to have properties of second order derivative [11] with a particular shift. Let $\tilde{c}_q(n)$ represents the basis of shifted Ramanujan operators mathematically expressed as: [11]

$$\tilde{c}_q(n) = \begin{Bmatrix} c_q(n - \frac{q-1}{2}),\ 0 \leq n \leq q-1 \\ 0,\ otherwise \end{Bmatrix} \tag{7}$$

For $q = 3$ and $q = 5$, $\tilde{c}_q(n)$ are given by:

$$\tilde{c}_3 = (-1, 2, -1),\ \tilde{c}_5 = (-1, -1, 4, -1, -1)$$

The first derivative property of RS is applied for noise estimation and the second derivative property finds its application in edge detection.

## III. Image Denoising

The main goals of image denoising are: effective noise suppression, preservation of edges and maintaining the visual quality of image [12]. When the image signal contains additive noise only, spatial filtering is preferred. Linear filters are preferred for suppression of additive-noise while non-Linear filters are preferred for multiplicative noise and function-based noise present in the image signal [13]. In general spatial filters can effectively suppress noise but at the expense of blurring pixels. Median filter is an example of non-Linear filter which is used for dealing with extreme outliers.

Another approach for image denoising is based on transform based filtering. Wavelet denoising is the most widely used technique [14]. It is basically reconstructing the signal through a non-linear operation in wavelet domain. The steps involve computing the wavelet decomposition coefficients of the given signal followed by thresholding the coefficients. Lastly reconstruction of signal is done by taking the Inverse wavelet transform of these thresholded coefficients.

In this paper we extend the denoising algorithm presented in [11]. We propose to iteratively perform denoising and also recover the edge information using the derivative properties of Ramanujan Sums. Let $x(n)$ be the signal which is inflicted by zero mean additive white Gaussian noise $y(n)$ and variance $\hat{\sigma}$. The noisy signal is represented by $s(n) = x(n) + y(n)$ of length $N$. The k divisors of $N$ are denoted by $d_1, d_2, ...d_k$. For each of the divisors, the normalized kernel of the corresponding Ramanujan operators are denoted by $(\tilde{c}_{d_1}, \tilde{c}_{d_2}, ...\tilde{c}_{d_k})$.

Fig. 1 shows the flowchart of the proposed iterative algorithm. In the second step of the algorithm, using equation (6) $\hat{R}_q$ is applied on the noisy signal $s(n)$. Each $(\hat{R}_{d_i} s)(n)$ is then downsampled by the corresponding factor $d_i$.

$$y_{d_i}(n) = (\hat{R}_{d_i} s)(nd_i)$$

Then for every divisor, Median Absolute Deviation (MAD) is computed and the standard deviation of noise ($\hat{\sigma}$) is computed as [11]:

$$\hat{\sigma} = \sum_{i=1}^{k} \frac{MAD(y_{di})}{0.6745k}$$

Thresholding is carried out on each $y_{d_i}(n)$ where entries less than $\hat{\sigma}$ are made zero and entries greater than $\hat{\sigma}$ remain unchanged. All these steps of the algorithm are iteratively carried out till the desired residual error is achieved. Finally, image sharpening with edges obtained using Ramanujan sums



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:59:55 UTC from IEEE Xplore. Restrictions apply.

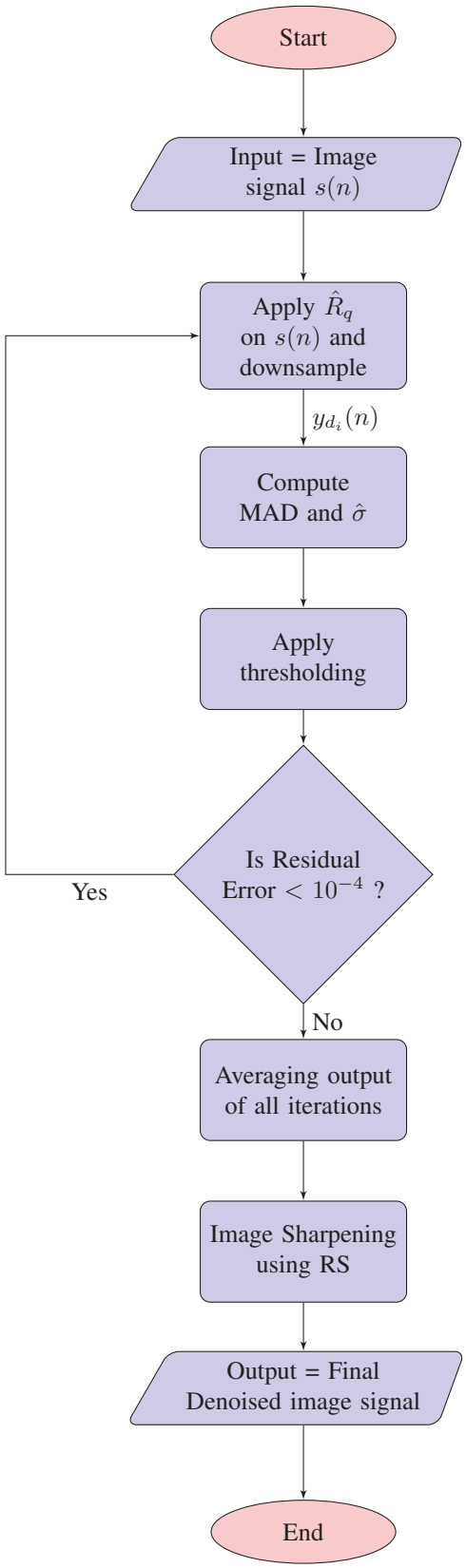



Fig. 1: Flowchart of Iterative denoising algorithm

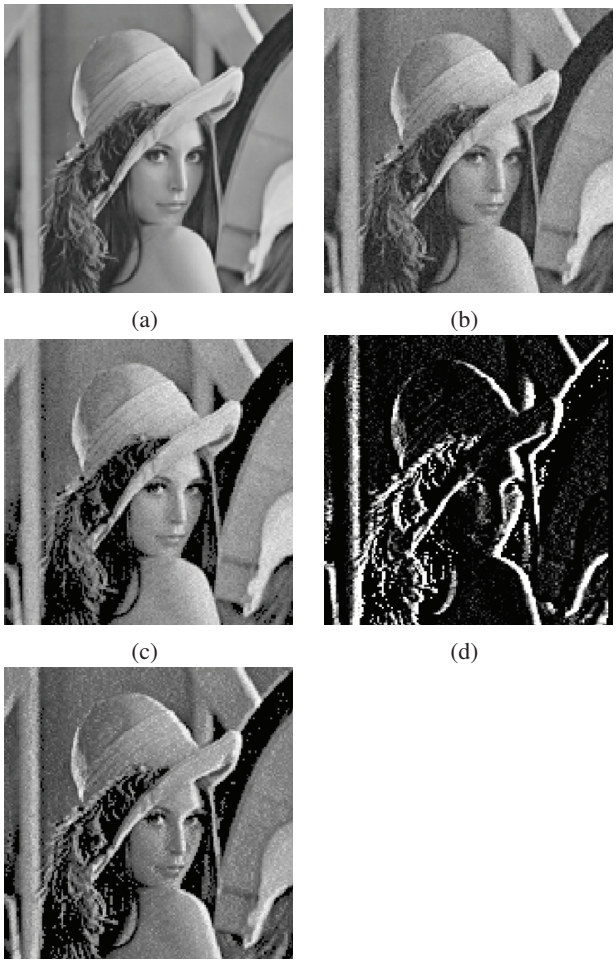



Fig. 2: (a) Input Grayscale image (b) Image with AWGN (mean = 0, variance = 0.001) (c) Denoised Image (d) Vertical Edges using second derivative (e) Final denoised image with edges retained

is performed to obtain the final output image.

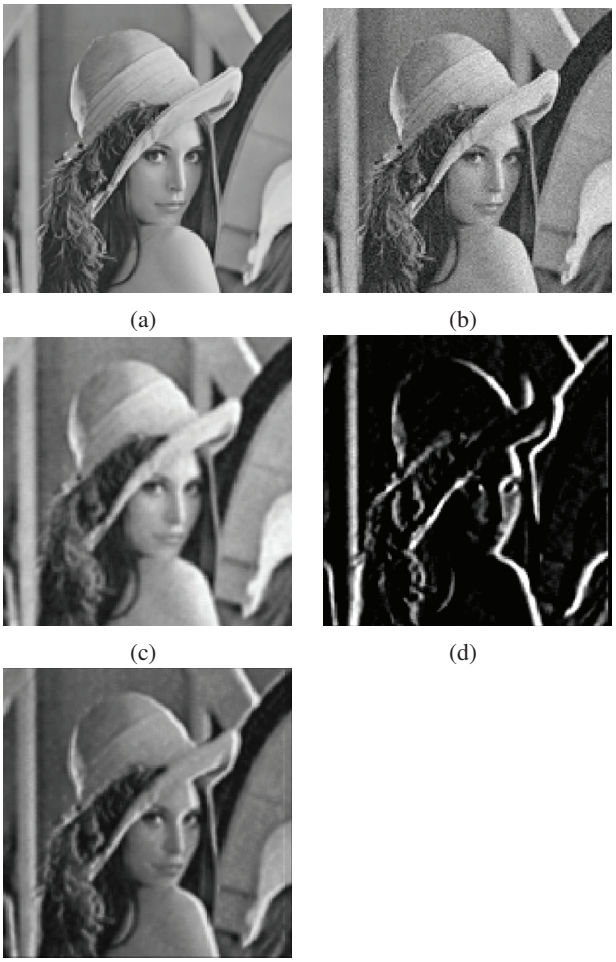
## IV. Results

The proposed algorithm was implemented in MATLAB the results obtained are shown in Fig. 2. The input grayscale image having 150 X 150 resolution is shown in Fig. 2(a). Fig. 2(b) depicts the image obtained after addition of white Gaussian noise having zero mean and variance of 0.001. In Fig. 2(c) the denoised image using Ramanujan sums is shown. Fig. 2(d) depicts the vertical edges obtained using $R_5$ operator on Fig. 2(c). Finally, Fig. 2(e) depicts the denoised image obtained using the iterative algorithm after noise suppression and image sharpening.

For comparision, the results of denoising using a 3X3 average filter are shown in Fig. 3. The input grayscale image having 150 X 150 resolution is shown in Fig. 3(a). Fig.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:59:55 UTC from IEEE Xplore. Restrictions apply.



(a) (b) (c) (d) (e)

Fig. 3: (a) Input Grayscale image (b) Image with AWGN (mean = 0, variance = 0.001) (c) Denoised Image using 3X3 Average Filter (d) Vertical Edges of the Input Image (e) Final denoised image with edges retained

3(b) is the image obtained after addition of white Gaussian noise having zero mean and variance of 0.001 to Fig. 3(a). In Fig. 3(c) denoised image using 3X3 Average filter is shown. Fig. 3(d) depicts the vertical edges of Fig. 3(c). In Fig. 3(e) the final output image of the 3X3 Average Filter after noise reduction and image sharpening is shown. On comparing the outputs obtained from the proposed iterative denoised algorithm and 3X3 average filter, it is inferred that the algorithm has performed denoising more effectively with increased SSI over the filter.

Let $f_k$ represents the denoised image for $k^{th}$ iteration ($k \geq 1$). The residual error $R_k$ is defined as:

$$R_k = \frac{\|f_{k+1} - f_k\|^2}{\|f_k\|^2} \tag{8}$$

The plot of the residual error for each iteration of the denoising algorithm using Ramanujan sums is shown in Fig. 4. It can be inferred from the figure that the residual error decreases with the increase in the number of iterations.

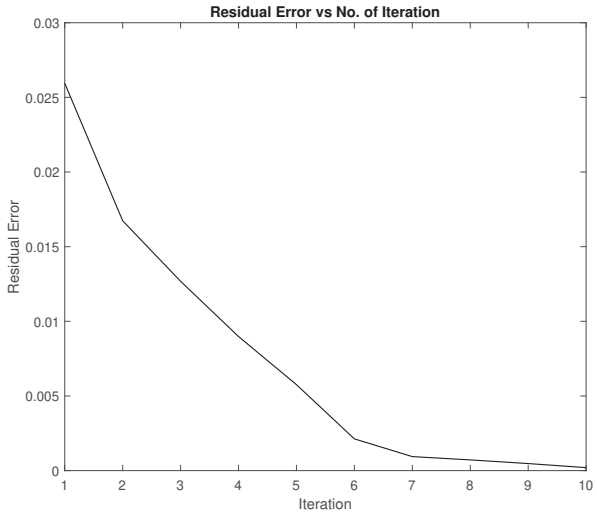



Fig. 4: Residual Error

TABLE II: COMPARISON OF STRUCTURAL SIMILARITY INDEX (SSI) FOR 3X3 FILTER AND PROPOSED ALGORITHM

| S.No | Image type | Structural Similarity Index (SSI) | |
|---|---|---|---|
| | | *3X3 Average Filter* | *Proposed Algorithm* |
| 1 | Noisy Image | 0.7381 | 0.7381 |
| 2 | Denoised | 0.7529 | 0.8003 |
| 3 | Denoised with edges restored | 0.7938 | 0.8319 |

In Table II we tabulate the structural similarity index (SSI) for the noisy image, denoised image and denoised image with edges restored of 3X3 average filter and proposed denoising algorithm [15]. It is observed that the proposed iterative denoising algorithm has better SSI in comparison to the 3X3 averaging filter.

## V. Conclusion

Ramanujan Sums are extensively being used in the areas of noise estimation and edge detection. The proposed iterative image denoising algorithm incorporating the derivative properties of Ramanujan sums was applied to noisy image. Simulation results show that the denoised ionage with proposed denoising algorithm has better SSI in comparision to averaging filter and the residual error decreases for each iteration.

## References

[1] V. Karnati, M. Uliyar and S. Dey, "Fast Non-Local algorithm for image denoising," 2009 16th IEEE International Conference on Image Processing (ICIP), 2009, pp. 3873-3876.

[2] D. Li, S. Simske and R. M. Mersereau, "Image Denoising Through Support Vector Regression," 2007 IEEE International Conference on Image Processing, 2007, pp. IV - 425-IV - 428.

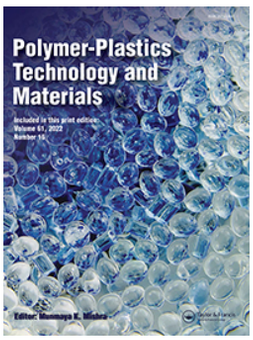


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:59:55 UTC from IEEE Xplore. Restrictions apply.

[3] A. Vyas and J. Paik, "Applications of multiscale transforms to image denoising: Survey," 2018 International Conference on Electronics, Information, and Communication (ICEIC), 2018, pp. 1-3
[4] X. Wang, G. Shi and L. Liang, "Image denoising based on translation invariant Directional Lifting," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, 2010, pp. 1446-1449,
[5] B. K. Thote and K. C. Jondhale, "Improved denoising technique for natural and synthetic images," 2016 International Conference on Signal and Information Processing (IConSIP), 2016, pp. 1-4
[6] P. P. Vaidyanathan, "Ramanujan Sums in the Context of Signal Processing—Part I: Fundamentals," in IEEE Transactions on Signal Processing, vol. 62, no. 16, pp. 4145-4157, Aug.15, 2014.
[7] P. P. Vaidyanathan, "Ramanujan Sums in the Context of Signal Processing—Part II: FIR Representations and Applications," in IEEE Transactions on Signal Processing, vol. 62, no. 16, pp. 4158-4172, Aug.15, 2014.
[8] P. P. Vaidyanathan and Srikanth Tenneti, "Srinivasa Ramanujan and signal-processing problems", Department of Electrical Engineering, California Institute of Technology, Pasadena, CA, USA, 2019.
[9] D. Abraham and M. Manuel, "Implementation of Sparse Ramanujan Sequence (SRS) based transforms in FPGA," 2021 International Conference on Communication, Control and Information Sciences (ICCISc), pp. 1-6 2021.
[10] A. Mariam Solomon, D. Abraham, A. George and M. Manuel, "Signal Denoising Using Ramanujan Periodic Transform," 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), pp. 51-56 2018
[11] D. K. Yadav, G. Kuldeep and S. D. Joshi, "Ramanujan Sums as Derivatives and Applications," in IEEE Signal Processing Letters, vol. 25, no. 3, pp. 413-416, March 2018.
[12] Somashekhar Swamy and P.K.Kulkarni ,"A Basic Overview On Image Denoising Techniques", International Research Journal of Engineering and Technology (IRJET) Volume: 07 Issue: 05 — May 2020.
[13] P. B.Alisha , K. S. Gnana , "Image Denoising Techniques-An Overview", IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), volume 11, pp 78-84, Jan. - Feb 2016
[14] C. Taswell, "The what, how, and why of wavelet shrinkage denoising," in Computing in Science Engineering, vol. 2, no. 3, pp. 12-19, May-June 2000.
[15] Zhou Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," in IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, April 2004.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:59:55 UTC from IEEE Xplore. Restrictions apply.



Polymer-Plastics Technology and Materials

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lpte21

# An overview on synthesis, properties and applications of polycaprolactone copolymers, blends & composites

Radha Sachan, Sudhir G. Warkar & Roli Purwar

**To cite this article:** Radha Sachan, Sudhir G. Warkar & Roli Purwar (2022): An overview on synthesis, properties and applications of polycaprolactone copolymers, blends & composites, Polymer-Plastics Technology and Materials, DOI: 10.1080/25740881.2022.2113890

**To link to this article:** https://doi.org/10.1080/25740881.2022.2113890

Published online: 11 Sep 2022.

Submit your article to this journal

Article views: 25

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=lpte21

POLYMER-PLASTICS TECHNOLOGY AND MATERIALS
https://doi.org/10.1080/25740881.2022.2113890

Taylor & Francis
Taylor & Francis Group

REVIEW

Check for updates

# An overview on synthesis, properties and applications of polycaprolactone copolymers, blends & composites

Radha Sachan, Sudhir G. Warkar, and Roli Purwar

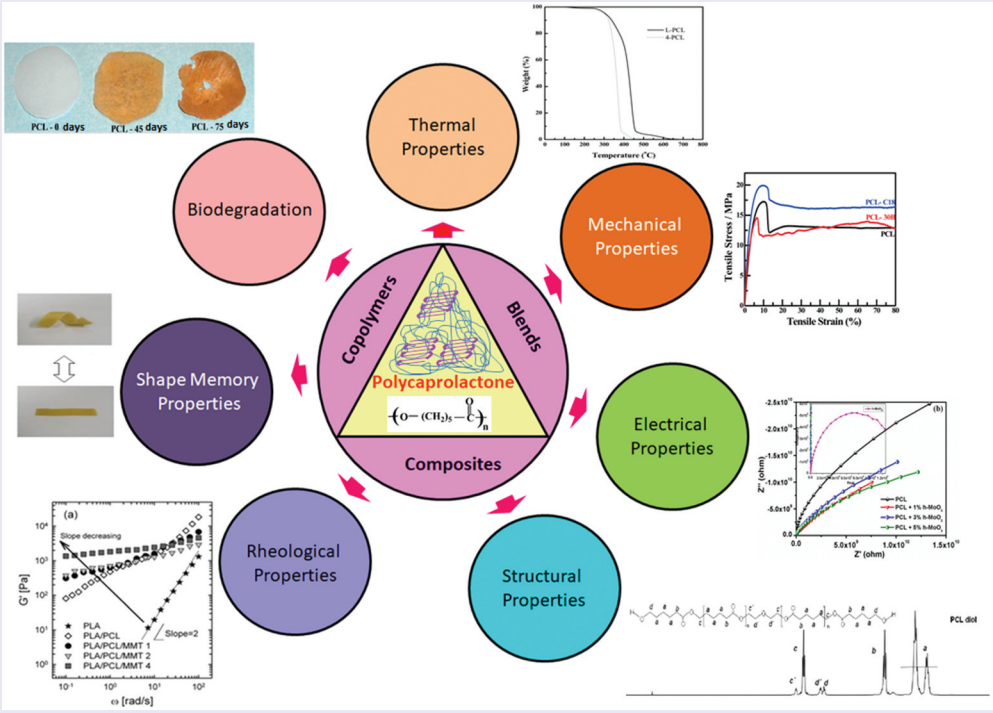
Discipline of Polymer Science and Chemical Technology, Department of Applied Chemistry, Delhi Technological University, Delhi, India

**ABSTRACT**

The formation of novel Polycaprolactone (PCL) copolymers, blends, and composites is used to broaden the versatility and applicability of the polymer. In the last decade, PCL copolymers, blends, and composites have been investigated well to tune their characteristics and properties. The review highlights thermal, mechanical, rheological, structural, electrical, shape memory, and degradation properties. Their biodegradability and modified properties make it an environment-friendly material to be used in various fields including biomedical, packaging, self-healing, coatings, oil-water separation, photo-thermal absorption and plant grafting. The commercial products based on polycaprolactone are listed in this review. Further, future recommendations for new research directions are provided.

**ARTICLE HISTORY**
Received 09 February 2022
Revised 25 July 2022
Accepted 12 August 2022

**KEYWORDS**
Polycaprolactone; copolymers; blends; composites; characterization




## 1. Introduction

Polycaprolactone (PCL) is a semi-crystalline, hydrophobic, synthetic aliphatic polyester consisting of hexanoate repeat units. It shows high solubility at the ambient conditions in organic solvents and is easily processible due to its low melting point (54–71°C). PCL's biodegradable and biocompatible properties make it highly researched material for tissue engineering, controlled drug delivery systems, wound dressing, contraceptive, dentistry, and bone engineering[1–3]. PCL has a low degradation rate in combination with high ductility and plasticity. It is certified by Food and Drug Administration, USA and CE-registered

**CONTACT** Roli Purwar roli.purwar@dtu.ac.in Discipline of Polymer Science and Chemical Technology, Department of Applied Chemistry, Delhi Technological University, Shahbad Daultapur, Delhi 110042, India

This article has been republished with minor changes. These changes do not impact the academic content of the article.

© 2022 Taylor & Francis

mark by European Community for application in numerous biomedical devices.[4] Molecular weight and degree of crystallinity affect its physical, thermal, mechanical, and rheological properties. The superior viscoelastic and mechanical properties make it suitable for food packaging, coating, adhesive, and textiles.

Natta *et al.* in 1934, first synthesized PCL by ring-opening polymerization technique.[4] Three well-known synthesis routes of polycaprolactone are as follows: (i) ionic and metal-catalyzed ring-opening polymerization (ROP) of ε-caprolactone, (ii) radical ring-opening polymerization (RROP) of 2-methylene −1,3 dioxepane (MDO), and (iii) condensation of 6-hydroxycaproic acid.[5] A good comprehensive study about the synthesis of polycaprolactone is given by Thielemans *et al.*[6]

A decent number of review articles are published on PCL and its applications, limited to biomedical and tissue engineering.[1–4, 7–17] A review on PCL is published by Woodruff *et al.* in 2010, which gives a detailed analysis of biodegradability, biocompatibility, and application in the biomedical field.[4] In 2019, Janmohanmmadi *et al.* published a review paper on biomedical applications of polycaprolactone scaffolds prepared using electrospinning technique.[3] Bartnikowski *et al.* have published a comprehensive review on degradation mechanism of polycaprolactone, its copolymer and composite.[18] Recently, reviews have been published on polycaprolactone-triol,[16] *in-vivo* studies of 3D model of polycaprolactone,[1] bone scaffolds [2] and biodegradable conducting polymers.[19]

More recently, PCL has been widely explored for thermo-responsive shape memory (TSM) behavior. Using the ROP method, the first TSM polyurethane based on PCL synthesized by Kim *et al.* in 1996.[20] The polymer exhibits temperature responsive shape memory behavior due to its semi-crystalline nature and low melting range. PCL shows exceptional blend compatibility.[16] TSM behavior of PCL copolymers and blends near-normal human body temperature makes it an interesting material for biomedical applications. The TSM behavior increases its application exponentially, including robotics, textiles, self-healing coatings, corrosion-resistant coatings, and anti-counterfeit materials.

Major published review papers focus on synthesis, biomedical application and degradability of PCL. Since 2010, a plethora of research has been reported on PCL copolymers, blends, and composites in different fields. Copolymers of PCL are synthesized to tune the properties on molecular level. PCL blends and composites have been synthesized in single & multi-phase with easier fabrication methods than copolymerization to modify their properties and broaden their applicability. Based on our knowledge, no review paper has been published on fabrication of PCL's copolymers, blends and composites with their correlation on thermal, mechanical, structural, rheological, electrical, degradation, and shape memory properties and applications.

## 2. Synthesis of polycaprolactone copolymers, blends and composites

Copolymers are composed of more than one monomer repeat unit provides an additional degree of freedom over homopolymers. Polymer properties can easily be tuned by adjusting the ratio of the monomers. PCL copolymers are synthesized by ROP and RROP methods, as shown in Figure 1. ROP is the most preferred route for the synthesis of PCL block copolymers, graft copolymers or random copolymers due to high yield (above 90%). For di-block, tri-block, or multi-block copolymers, in the presence of catalyst stannous octoate ROP method is used where other monomers are L-lactide,[21,22] dimethyl siloxane,[23,24] polyethylene glycol,[25] and *N*-2-hydroxypropyl methacrylamide.[26] ROP is used to functionalize monomers and further condensation[27] or crosslinking[28–30] is carried out to get block copolymer. Bulk polymerization of ε-CL with other monomers with different weight percent gives random copolymer [31–33] by ROP mechanism. The graft copolymers with cellulose,[34] hydroxyethyl methyl acrylate,[35] ethyl vinyl acetate[36] is synthesized by ROP followed by grafting from[35] grating through[35] and transesterification method.[36] Vinyl monomers are copolymerized with MDO(2-methylene-1,3dioxepane) for the cognizance of functional biodegradable polyesters, include vinyl bromobutanoate,[37] methyl methacrylate,[38] hydroxyethylacrylate,[39] vinylcyclopropane,[40] vinyl ethers[19] and glycidyl methacrylate[41,42] by radical ring-opening polymerization(RROP). Another outstanding feature is that PCL's structure obtained with RROP method changes when compared to ionic or metal-catalyzed ROP of ε-caprolactone. These copolymers have been suggested to comprehend materials integrating the elasticity of PCL, shape memory behavior, and alter degradation times. Table 1. represents the copolymers of polycaprolactone with different monomers and their synthesis conditions.

Polymer blends and composite are easier method than copolymerization to improve properties of polymeric system by incorporating other polymer, copolymer or filler. There are number of method available for blending of polymers such as melt mixing, solution mixing, latex mixing and preparation of

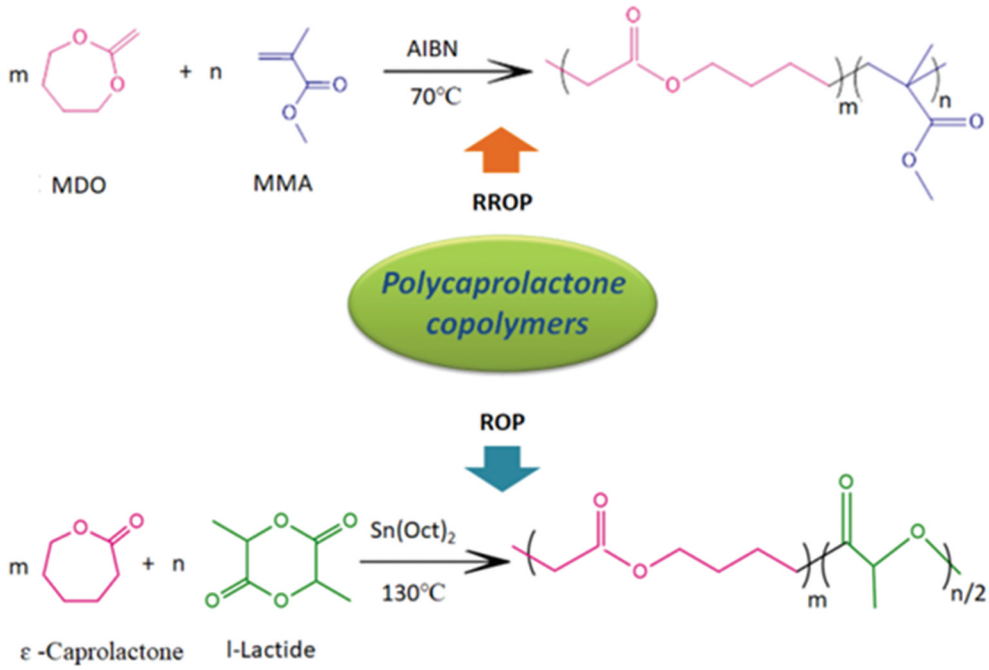



**Figure 1.** Synthesis routes of copolymerization of polycaprolactone.

**Table 1.** Copolymers of polycaprolactone.

| S.No. | Co-monomer | Method | Copolymer type | Catalyst/ Initiator | Reaction Condition | Ref |
|---|---|---|---|---|---|---|
| 1. | L-lactide | ROP | Block | Stannous Octoate | 140°C, 3 hrs | [21, 22] |
| 2. | Dimethyl Siloxane | ROP | Block | Stannous 2-ethylhexanoate | 145°C, 24 hrs and 120°C, 48 hrs | [23, 24] |
| 3. | Polyethylene glycol | ROP | Block | Stannous 2-ethylhexanoate | 140°C, 24 hrs | [25] |
| 4. | *N*-2-hydroxypropyl methacrylamide | ROP | Block | Stannous 2-ethylhexanoate | 65°C, 24 hrs | [26] |
| 5. | ω- pentadecalactone | ROP | Block | Ethylene glycol | Condensation of diols and reaction with 4-methylhexamethylene diisocynate at 83°C | [82] |
| 6. | 4-hydroxycinnamic acid | ROP | Block | Stannous 2-ethylhexanoate | Thermal polycondensation of diols at 200°C | [27] |
| 7. | *p*-dioxanone | ROP | Random | Stannous 2-ethylhexanoate | 130°C, 48 hrs | [31] |
| 8. | D.L-lactide | ROP | Random | Stannous 2-ethylhexanoate | 110°C, 24 hrs | [32] |
| 9. | Butyl acrylate | ROP | Random | Stannous 2-ethylhexanoate | Melt mixing of oligo caprolcatone diol with butyl acrylate at 80°C | [33] |
| 10. | Cellulose | ROP | Graft | Stannous 2-ethylhexanoate | 120°C, 24 hours, grafting from | [34] |
| 11. | Hydroxyethyl methyl acrylate | ROP | Graft | | Grafting though and grafting from | [35] |
| 12. | Ethyl vinyl acetate | ROP | Graft | Titanium propoxide | Transesterfication | [36] |
| 13. | Vinyl bromobutanoate | RROP | Random | azobis(isobutyronitrile) (AIBN) | 60°C, 9 hrs | [37] |
| 14. | Methyl methacrylate | RROP | Partial random | AIBN | 75°C, 24 hrs | [38] |
| 15. | Hydroxyethyl acrylate | RROP | Graft | 2,20-azobis (4-methoxy-2,4-dimethylvaleronitrile) (V-70) | 30°C, 24hrs | [39] |
| 16. | 2-vinylcyclopropane 1,1-dicarboxylate | RROP | Random | di-tert-butyl peroxide | 120°C, 48 hrs | [40] |
| 17. | Functional vinyl ethers | RROP | Graft | Diethylazobisisobutyrate | 70°C, 8hrs | [19] |
| 18. | Glycidyl methacrylate | RROP | Random | AIBN | 60°C, 3 hrs | [41,42] |

interpenetrating polymer networks (IPN). Synthesis of PCL blends is mostly carried out by melt blending, solution blending method due to low melting point and solubility in organic solvents as shown in Figure 2. Several polymers such as poly(L-lactide)/poly(lactic acid),[43–47] pine resin,[48] poly(glycolic acid),[49] lignosulfate,[50] chitosan,[51] and hydroxybutyrate[52,53] are melt blended by mixing at different temperature.

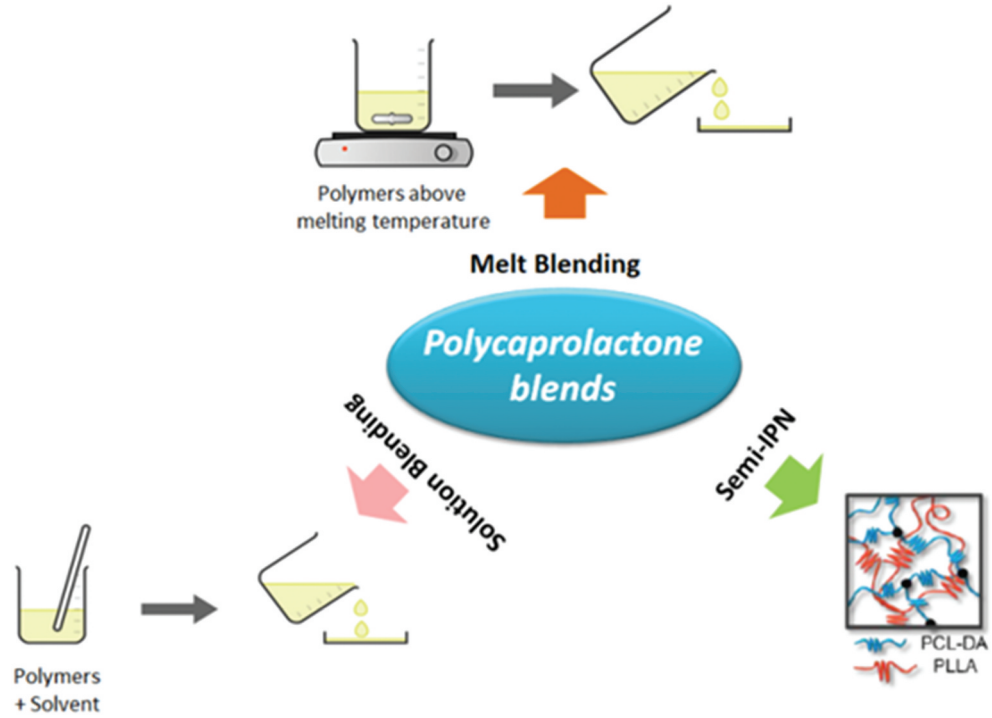



**Figure 2.** Synthesis methods of blends of Polycaprolactone.

**Table 2.** Blends of polycaprolactone.

| S.No. | Polymers | Method | Blending conditions | Ref |
|---|---|---|---|---|
| 1. | Poly (L-lactide) | Melt blended | Twin screw extruder, 150 rpm, 180°C or Haake polylab rheometer, 50 rpm, 170°C, 6 mins. or Brabender Plasticorder, 60 rpm, 10 min at 180°C | [43–47] |
| 2. | Pine resin | Melt blended | Mini max molder, 50 rpm, 15 min at 80°C. | [48] |
| 3. | Poly (glycolic acid) | Melt blended | | [49] |
| 4. | Lignosulphate | Melt blended | Rheomix 600 OS, 60 rpm, 10 min at 120°C | [50] |
| 5. | Chitosan | Melt blended | HAPRO RM-200A, 20 rpm, 20 min at 110°C | [51] |
| 6. | poly(3-hydroxybutyrate) | Melt blended | Brabender mixer, 100 rpm, 8 min at 170°C | [52, 53] |
| 7. | Poly(L-lactide) | IPN | Photo-crosslinking of acrylate terminated PCL and PLLA mixture | [54] |
| 8. | Poly(L-lactide) | Solution blended | Chloroform or Chloroform/ *N, N*-dimethylformamide mixture or Chloroform/ 1,4-dioxane mixture | [55–57] |
| 9. | Poly(glycolic acid) | Solution blended | 1,1,1,3,3,3,-hexafluoroisopropanol | [58] |
| 10. | Poly(3- hydroxybutyrate-co -3-hydroxyhexanoate) | Solution blended | Chloroform | [59] |
| 11. | Poly(hexylene adipate) | Solution blended | Chloroform | [60] |
| 12. | Poly(methylmethacrylate) | Solution blended | Chloroform | [61] |

Table 2 represents different polymer used to preprare blend with PCL along with their blending conditions. The melting temperature is chosen on the basis of polymer of higher melting point and crystallinity. To improve the compatibility of blends, free radicals like benzoyl peroxide are used for PCL/poly(3-hydroxybutyrate) in melt blending method by Przybysz *et al.*[53] and Woodard *et al.* are chosen different method i.e. semi-IPN network for PCL/PLLA blend.[54] Solution blend of PCL and PLA/PLLA,[55–57] poly (glycolic acid),[58] poly(3-hydroxybutyrate-co-3-hydroxyhexanoate),[59] poly(hexylene adipate)[60] and poly(methylmethacrylate)[61] are prepared either by chloroform or mixture of chloroform with dimethylformamide, 1–4, dioxane, acetone or ethanol because of ease of solubility at ambient conditions.

Composites are synthesized by various methods such as intercalation, in-situ polymerization, and direct

mixing of polymer and fillers. The preparation of PCL composite is reported by all the above techniques as shown in Figure 3. Different fillers used to prepare composite of PCL are shown in Table 3. PCL easily melts above 60°C and becomes soluble in chloroform or its mixture; that is why melt mixing is the preferred route to synthesize composite followed by the solvent method. Composite of PCL with alumina,[62] nibonium pentaoxide,[62] MWCNT,[63] leaf sheath date palm fibre,[64] hydroxyapatite,[65] silica carbide,[66] soy protein isolate,[67] halloysite,[68] babassu,[69] rice straw fiber,[70] clam shell powder,[71] and calcium phosphate[72] are processed by melt compounding/mixing method. A significant hurdle in the production of PCL composites with natural waste by products' fillers is that PCL is relatively hydrophobic, in contrast most natural fillers are hydrophilic, which leads to poor compatibility between the two phases of the composites. To overcome this problem, compatibilizers in the form of reactive functional groups are incorporated to PCL polymer backbone for miscibility. Melt intercalation method is used for clay composites.[73,74] The solvent method is used for films, nano-fibrous and 3D printed composites of PCL with fillers such as magnetite,[75] casein,[75] cellulose nano-whiskers,[76] hexagonal molybdenum oxide,[77] CNT,[78] MWCNT,[79] carbonyl iron powder[80] chitosan [81] and gold nanoparticles (AuNPs).[82] *In situ* method is used for CNT[83] and graphene oxide[84] where nanofillers are added in ε-caprolactone before polymerization. On a different approach, nanocomposites of PCL/gold nanoparticles are prepared by solution casting of AuNPs over PCL nanofibers.[85] The sandwich-type of composite is fabricated by compression molding method, where reinforcement layers of electrospun cellulose nanofibers [86] and gelatin [87] are placed between PCL matrix layers.

All copolymers are reported by ROP and RROP methods with a good number of synthetic polymers and cellulose as natural polymer. Here natural polymers like polysaccharides (e.g. natural gums) and proteins (e.g. silk) still need to work upon. Copolymers of polycaprolactone can also be synthesized by condensation of 6-hexanoic acid. PCL blends are prepared by melt, solution, and IPN methods, the but latex mixing method is untouched. Latex mixing is used for blending two latexes. This method can lead to preparing an emulsion of polycaprolactone and blend it with water-based polymers. Such emulsions can be used for low volatile organic content (VOC) systems. A large number of PCL composites are reported with natural and synthetic fillers. Composites based on nanoparticles of copper and silver are not reported, yet their antibacterial properties make a mark in biomedical field.

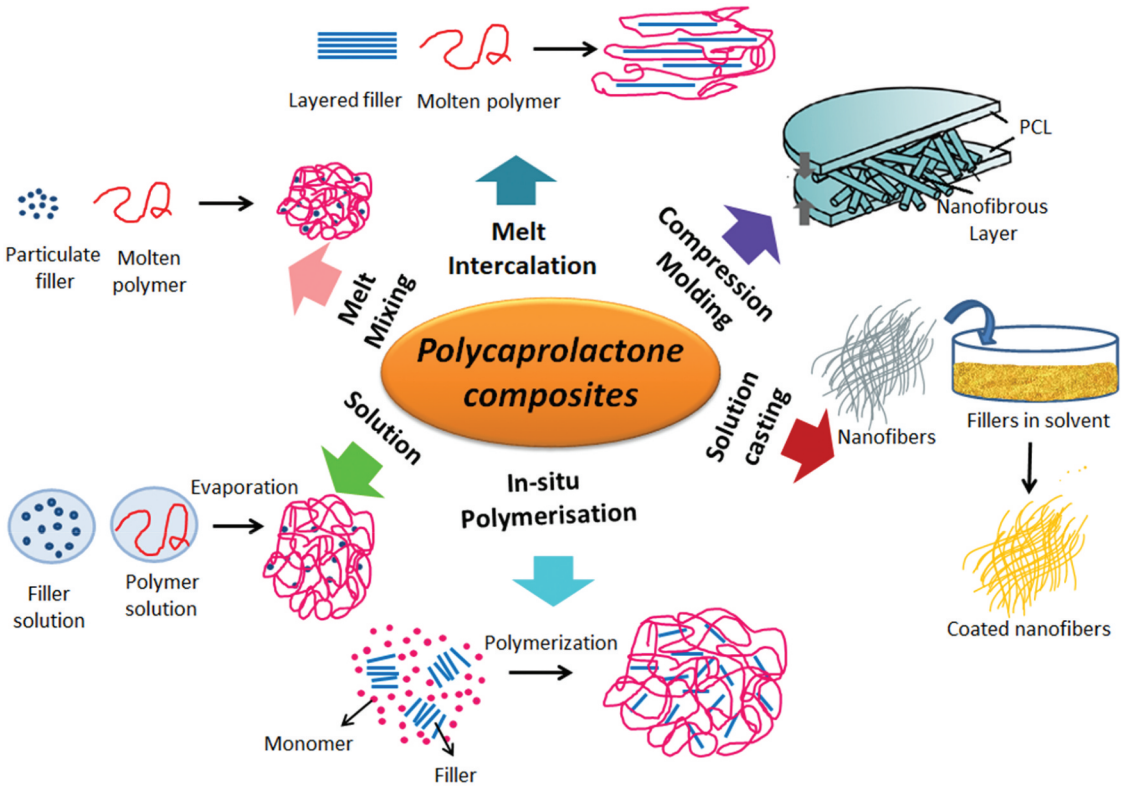



**Figure 3.** Synthesis methods of composites of polycaprolactone.

**Table 3.** Composites of polycaprolactone.

| S. No. | Fillers | Method | Conditions | Ref |
|---|---|---|---|---|
| 1. | Alumina | Melt mixing | Haake Rheomix 3000 laboratory internal mixer, 150°C | [62] |
| 2. | Nibonium pentaoxide | Melt mixing | Haake Rheomix 3000 laboratory internal mixer, 150°C | [62] |
| 3. | MWCNT | Melt mixing | Twin-screw Extruder, 400 rpm, 90°C, 2 min | [63] |
| 4. | Leaf sheath date palm fibre | Melt mixing | Twin-screw co-rotating extruder, feed rate 3.5 kg/h, screw speed 100 and 400 rpm, 120°C, | [64] |
| 5. | Hydroxyapatite | Melt mixing | Twin screw Compounder, 100°C, 100 rpm, 6 mins. | [65] |
| 6. | Silicon carbide | Melt mixing | Twin screw Thermo Scientific Haake Rheomixer, 100°C, 50 rpm, 10 min. | [66] |
| 7. | Soy protein isolate | Melt mixing | Intensive mixer, 100°C, 30 rpm, 10 min | [67] |
| 8. | Halloysite | Melt mixing | Haake Rheocorder, 100°C, 60 rpm, 10 min | [68] |
| 9. | Babassu | Melt mixing | Haake Rheomix laboratory internal mixer, 150°C, 60 rpm, 15 min | [69] |
| 10. | Rice straw fiber | Melt mixing | Plastograph 200-Nm mixer W50EHT 70–80°C, 20 min, 50 rpm | [70] |
| 11. | Clam shell powder | Melt mixing | Intensive mixer, 100°C 30 rpm, 10 min | [71] |
| 12. | Calcium Phosphate | Melt mixing | Extruder, room tempera- ture to 160°C | [72] |
| 13. | Clay ($Na^+$ MMT, Cloisite 30B, Cloisite 20A, bentonite) | Melt Intercalation | Double-screw mini-extruder DSM Xplore 5&15 microcompounder 60–90–120°C, 150 rpm, 1 min. | [73,74] |
| 14. | Magnetite | Solution method | THF, dispersed for 8 minutes,15,000 rpm | [75] |
| 15. | Casein | Solution method | THF, dispersed for 8 minutes,15,000 rpm | [75] |
| 16. | Cellulose nanowhiskers | Solution method | DMF, 70°C | [76] |
| 17. | Hexagonal molybdenum oxide | Solution method | Chloroform, 27°C | [77] |
| 18. | CNT | Solution method | Chloroform, sonication for 5 min. | [78] |
| 19. | MWCNT | Solution method | MWCNT in DMF & PCL in THF | [79] |
| 20. | Carbonyl iron powder | Solution method | | [80] |
| 21. | Chitosan | Solution method | DMF, Chloroform, Ultrasonication | [81] |
| 22. | CNT | In-situ polymerization | Addition before polymerization | [82] |
| 23. | Graphene oxide | In-situ polymerization, solvent mixing. Covalent attachment | Addition before polymerization, DMF dispersed GO mixed with PCL, Addition of filler in the presence of N,N′-dicyclohexylcarbodiimide as coupling agent and 4-dimethylaminopyridine as catalyst | [84,97] |
| 24. | Gold Naoparticles | Solution casting | Electrospun PCL nanofiber dipped in AuNPs methanol mixture | [85] |
| 25. | Cellulose nanofibre | Compression molding | Hydraulic press at 80°C | [86] |
| 26. | Gelatin | Compression molding | Hydraulic press at 80°C | [87] |

## 3. Properties of polycaprolactone copolymers, blends, and composites

PCL has melting point between 54°C and 71°C and a glass transition temperature ($T_g$) of −60°C.[88] It is a well-known semi-crystalline polymer having a maximum degree of crystallinity of 69%.[89] The physical, thermal, and mechanical properties of PCL are influenced by its molecular weight and degree of crystallinity. Extensive research has been carried out to improve or modify the chemical, mechanical, and thermal properties of PCL by copolymerization, blending, and formation of composites. They also modify degradation time and rheological properties, which changes its processing parameters. Different factors that influence the properties of polycaprolactone copolymers, blends, and composites are summarized in Figure 4. The Following sections summarize the thermal, mechanical, electrical, structural, rheological, shape memory and degradation properties of PCL copolymers, blends, and composites.

### 3.1. Thermal properties

The thermal properties of PCL are highly reliant on molecular weight and degree of crystallinity.[90] The thermal degradation of PCL is reported as a single step in the range of 358–400°C, as shown in Figure 5. Copolymerization, blending and formation of composites influence the $T_g$, degradation temperature, crystallization behavior and percentage crystallinity of PCL.

The thermal properties of PCL copolymers with other monomer-like ω-pentadecalactone,[91] L-lactide,[21,29,32] dimethyl siloxane,[23] butyl acrylate,[33] p-dioxanone,[31] ethyl vinyl acetate,[36] hydroxymethyl acrylate,[35,92] glycidyl methacrylate,[93] hydroxycinnamic acid[27] and

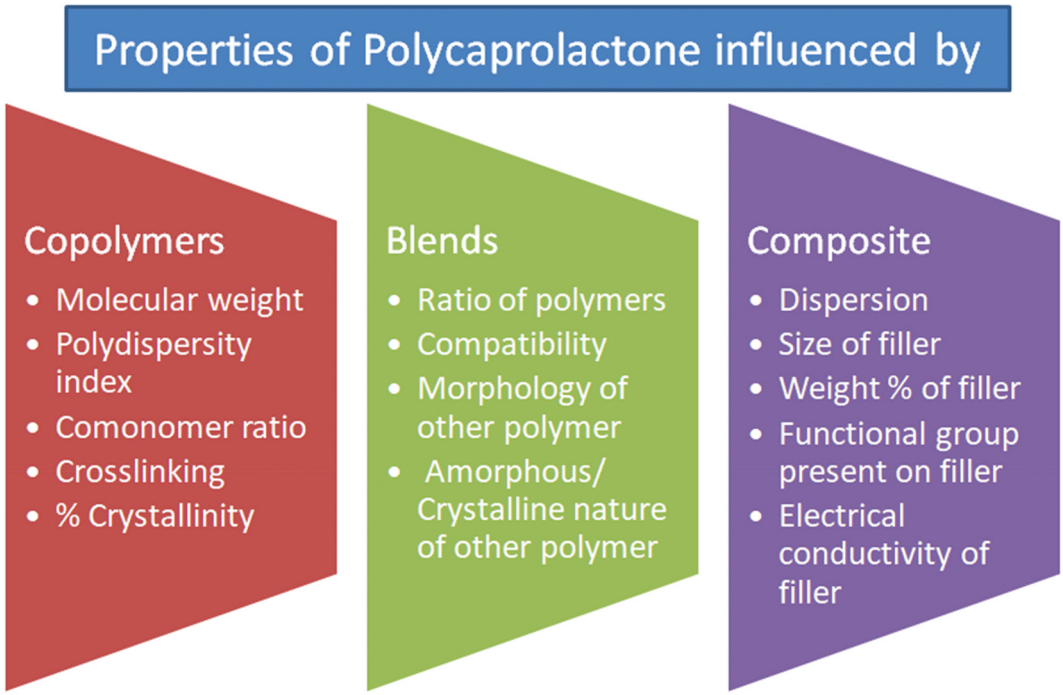



**Figure 4.** Different factors which influence the properties of copolymers, blends and composites of polycaprolactone.

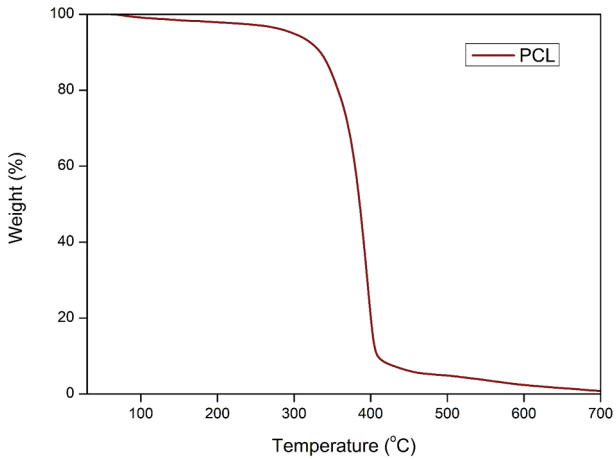



**Figure 5.** Thermal gravimetric analysis of PCL.

ethylene glycol[30] are studied by researchers. Block copolymer of PCL with ω-pentadecalactone (PDL) shows two distinct peaks of melting and crystallization owing to PCL and PDL since both have their own melting and crystallization temperature range. There is an increase in melting temperature ($T_m$) and crystalline temperature ($T_c$) of PCL with a rise in PCL content due to bigger block formation of PCL, as shown in Figure 6.1[91]. Similar observations are found in the case of PCL/L-Lactide block copolymer.[21] With increasing content of hydroxycinnamic acid (HCA) in copolymer[27] thermal stability, $T_m$ and crystallinity are decreased because the HCA chain disturbed the crystal growth of PCL content. Copolymers based on completely amorphous dimethylsiloxane ($T_g$ = −125°C) and n-butyl acrylate($T_g$ = −60°C) are showing a decrease in $T_m$, $T_c$ and crystallinity with decreasing PCL content due to shorter chain length of crystalline polymer. [23,33] PCL and p-dioxane copolymer showed no significant change in thermal stability. Still, there is a huge decrement in melting & crystallization peak and percentage crystallinity with an increase in dioxane content due to increased disorder along the polymer chains and imperfect packing of the polymeric segment in the crystalline lattice.[31] The graft copolymer of EVA/PCL showed improved thermal stability, melting temperature, and crystallinity dependent on the molar mass of both and structural differences between the two[36].

Thermal properties of PCL blends with Poly(L-lactide) acid, [43,44] polyglycolic acid (PGA), [49] thermoplastic starch (TPS), [49] chitosan, [49,94] PMMA,[61] polyhydroxy butyrate (PHB), [52,53,59] lignosulphonate[50] are reported. Blends of PCL with PLLA, PHB, lignosulphonate, and PGA are immiscible and miscible blends are with pine resin, TPS, chitosan, PHA and PMMA. The miscibility is reported by single $T_m$ and $T_g$ peaks mostly. PLLA has a higher melting point than PCL and blends showed a gradual increase in melting point with a higher content of PLLA. [43,44] The ternary blend of TPS/PGA with PCL showed a decrease in melting peak and multi-melting peak because these blends are not able to co-crystallize.

A blend of PCL with pine resin becomes thermally unstable because pine resin itself is thermally unstable (onset temperature ~130°C).[48] PCL/pine resin blends are showed a single $T_m$ and $T_c$ peaks indicating their miscibility and reduction in both with increased content of pine resin, as shown in Figure 6.2. Though, an improvement in thermal properties is reported with chitosan. An immiscible blend of PLA with PCL increases its crystallization temperature by 10°C due to discrete PLA solid phase having a strong nucleating effect. [45,56] The immiscible blends of PCL/PHB showed two distinct peaks in DSC thermogram; however, the addition of compatibilizer i.e. di-(2-tert-butyl-peroxyisopropyl)-benzene (BIB) has shown improvement in blend miscibility due to partial crosslinking[52,53].

The thermal properties of PCL composites with a variety of filler are studied extensively. Some of them include montmorillonite nanoclays,[73] bentonite,[74] carbon nanotubes,[63,78,79,83,95] alumina,[62] nibonium pentaoxide,[62] natural filler,[64,69,76,86] halloysite,[68] hydroxyapatite(HA),[65] silicon carbide,[66] zinc oxide nanoparticles[96] and graphene. [84,97] CNT/MWCNT-based nanocomposite showed improvement in thermal stability, melting & crystallization peaks, and crystallinity, as shown in Figure 6.3.[79] Functionalization of CNT showed an improvement in interfacial interaction

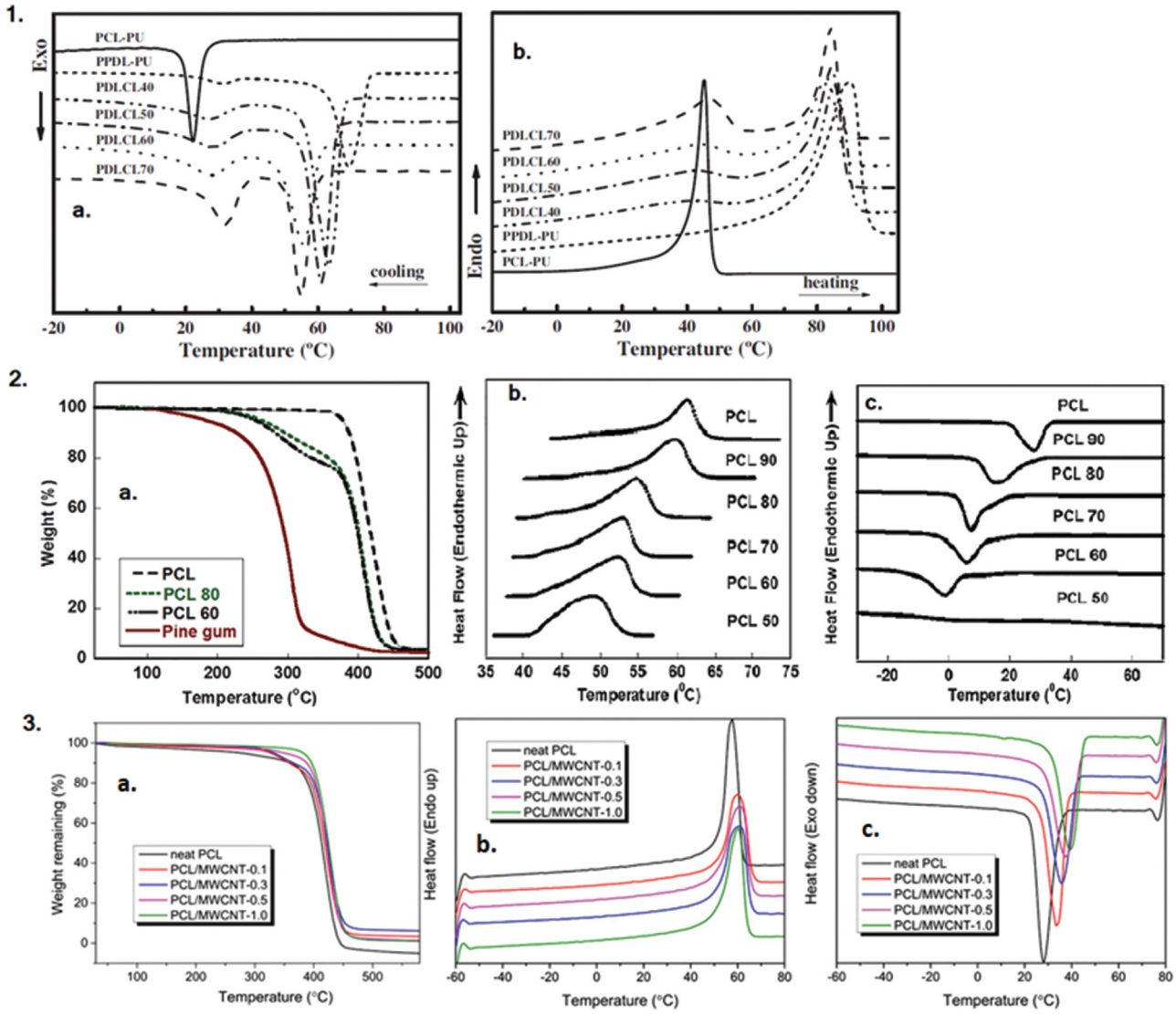



**Figure 6.** Melting 1(a) and cooling 1(b) DSC curves of PCL-co-ω-pentadecalactone reproduced with permission [91] 2019, Wiley; TGA curves 2(a); DSC curves melting 2(b) and cooling 2(c) of PCL/pine resin blends reproduced with permission [48] 2014, Wiley; TGA curves 3(a), DSC curves melting 3(b) and cooling 3(c) of PCL/MWCNT nanocomposite reproduced with permission [79] 2019, TnF.

between polymer and CNT[63,78,79,83] which resulted in improved thermal properties significantly. Chemically modified graphene, i.e. reduced graphene oxide showed better improvement in thermal properties of PCL as compared to graphene oxide. [84,97] Natural fillers like babassu fiber,[69] leaf sheath date palm fiber waste biomass,[64] cellulose nanofiber[86] & whiskers[76] and gelatin films[87] are incorporated in PCL to tailor thermal properties. There is a common observation that thermal stability, melting peak, crystallization peak, and crystallinity improved with natural fiber nanocomposite due to their good compatibility and they behave as a nucleating agent. Nanocomposite prepared with silicon carbide does not affect thermal stability and melting point in a significant way, but there is an improvement in crystallization temperature and crystallinity due to heterogeneous nucleation and an increase in the lamellar thickness of the crystallite formed.[66] When citric acid modified ZnO (ZnO-CA) nanoparticles are added to the PCL matrix, with increase in ZnO-CA content, the crystallization temperature ($T_c$) increases because, here, ZnO-CA acts as a nucleating point.[96] Thermal stability of PCL/clay nanocomposites is decreased with addition of clay due to its hydrophilicity and organically modified ions.[73,74,98] Addition of halloysite nanotubes increases the $T_g$ because of the restricted segmental motion of polymeric chains.[68] A very good study by Sousa *et al.* demonstrated that thermal properties are not only affected by filler percentage and type but also depend on the rate of heat flow[62]. The thermal properties of PCL copolymers, blends and composites are reported in Table 4.

**Table 4.** Thermal and mechanical properties of copolymers, blends, and composite of PCL.

| S. no. | Polymers | Synthesis mode | Composition | $T_m$ [°C] | %$X_c$ | Young's Modulus [Mpa] | Tensile strength [Mpa] | Elongation [%] | Ref |
|---|---|---|---|---|---|---|---|---|---|
| 1 | PCL-PDMS-PCL | Triblock copolymer | $PCL_{50}$-$PDMS_{37}$-$PCL_{50}$ | 51.0 | 30.0 | 60.5 | 15.9 | 1197.0 | [23] |
| 2 | *n* BA-PCL | Block Copolymer | PCL-BA (75:25) | 28.0 | _ | _ | 38.0 | 55.0 | [33] |
| 3 | PCL-PLLA | Block Copolymer | PCL-PLLA (50:50) | 49.9&159.9 | 27.5 & 55.1 | _ | _ | _ | [28] |
| 4 | CL- *p* dioxanone | Random copolymer | CL-DA (87:13) | 43.5 | 29.5 | 325.0 | _ | 450.0 | [31] |
| 5 | PCL-EVA | Graft copolymer | PCL-EVA (59.5/38.6) | 50.6 | _ | _ | 2.9 | 409.0 | [36] |
| 6 | PCL-HCA | Block copolymer | PCL-HCA (87:33) | 56.6 | 28.0 | _ | _ | _ | [27] |
| 7 | PCL-PEG | Random copolymer | PCL-PEG (50:50) | 40.0 | 45.2 | 0.023 | _ | _ | [25] |
| 8 | PCL-PLLA | Block Copolymer | PCL-PLLA (80:20) | 35.0 | 8.1 | 3.6 | 2.9 | 3200.0 | [32] |
| 9 | PCL/PHB | Melt blending | 75/25 | 61.2 & 154.0 | 39.1 | _ | 11.4 | 125.0 | [53] |
| 10 | PCL/PLLA | Semi IPN | 75/25 | 49.1 & 157.9 | 25.4 & 15.4 | 93.3 | 18.8 | 593.0 | [54] |
| 11 | PCL/ Pine resin | Melt blending | 70/30 | 53 | | _ | 15.2 | 2600.0 | [48] |
| 12 | PCL/PLA | Melt blending | 70/30 | 56.7 | 38.3 | 732.5 | 25.3 | _ | [47] |
| 13 | PCL/lignosulpahte | Melt blending | 70/30 | 56.0 | 36.1 | 410.8 | 14.7 | 110.9 | [50] |
| 14 | PCL/Chitosan | Melt blending | 80/20 | _ | _ | _ | 15.6 | 151.8 | [49] |
| 15 | PCL/poly(3-hydroxybutyrate-co-3-hydroxyhexanoate) | Solution blending | 20/80 P(3HB-co -7 mol% 3HH) | 59.9 | 38.9 | 477.0 | 11.3 | _ | [59] |
| 16 | PCL/PHA | Solution blending | 60/40 | 56.9 | 55.2 | | _ | _ | [60] |
| 17 | PCL/PLA | Solution blending (Electrospun) | 20/80 | 54.0 & 153.0 | 18.9 & 2.9 | | 1.7 | 94.4 | [57] |
| 18 | PCL/cellulose nanofiber | Compression composite | 100/17 | 59.4 | 56 | 744.7 | 32.8 | 9.0 | [86] |
| 19 | PCL/gelatin | Compression composite | 90/10 | 58.2 | 50.3 | 535.7 | 13.2 | 175.3 | [87] |
| 20 | PCL/CNT | In-situ composite | 100/1 | 58.0 | 54.0 | 126.7 | _ | 3.5 | [82] |
| 21 | PCL/Graphene oxide | In-situ composite | 100/0.5 | _ | _ | 199.0 | 13 | 788.0 | [84] |
| 22 | PCL/Alumina | Melt composite | 100/3 | _ | _ | 239.8 | 21.9 | 413.0 | [62] |
| 23 | PCL/ Nibonium pentaoxide | Melt composite | 100/5 | _ | _ | 243.9 | 23.5 | 485.0 | [62] |
| 24 | PCL/ Leaf sheath date palm fibre waste | Melt composite | 100/20 | _ | _ | 284.0 | 24 | 21.0 | [64] |
| 25 | PCL/SiC | Melt composite | 96/4 | 56.5 | 35.2 | 102.8 | 26.8 | 3063.0 | [66] |
| 26 | PCL/HNT | Melt composite | 90/10 | 53.1 | 42.1 | 919.0 | 84.7 | 743.0 | [62] |
| 27 | PCL/Rice straw fibre | Melt composite | 90/10 | 59.9 | 41.7 | _ | 36.1 | 655.0 | [70] |
| 28 | PCL/Na+ MMT | Melt intercalation | 100/5 | _ | 68.0 | 240.0 | 15.8 | 1345.0 | [73] |
| 29 | PCL/C30B | Melt intercalation | 100/5 | _ | 72.0 | 303.0 | 14 | 1212.0 | [73] |
| 30 | PCL/C20A | Melt intercalation | 100/5 | _ | 71.0 | 331.0 | 14.1 | 1430.0 | [73] |
| 31 | PCL/celluose nanowhiskers | Solutiom composite (electrospun) | 100/1 | 64.1 | 25.4 | 5.0 | 1.1 | 115.0 | [76] |
| 32 | PCL/MWCNT | Solution composite | 100/1 | 60.0 | 45.3 | _ | 30.4 | 1286.0 | [79] |
| 33 | PCL/chitosan | Solution composite (electrospun) | 100/50 | _ | _ | 26.5 | 4.4 | 133.0 | [81] |

Two different approaches are used to describe PCL copolymers' thermal properties. In the first approach, comonomer content is kept constant and varied the PCL chain length. The other approach is constant PCL chain length and varied comonomer content. The $T_m$, $T_c$ and percentage crystallinity increase when PCL chain length increases in copolymers. However for copolymers of fixed PCL length and varying comonomer content, thermal properties are shifting to lower degrees with increasing comonomer ratio. The blends exhibit their properties based on their miscibility. In general, miscibility or immiscibility is reported by single peak of $T_g$ and $T_m$, while in case of PCL blends, only $T_m$ and $T_c$ are discussed. In some studies, compatibility of blends is not even examined. Some immiscible blends, which are studied by researchers, are not explored for the methods to improve the miscibility. Almost negligible data is available on effect on $T_g$ for PCL copolymers, blends and composites.

Formation of crystals in PCL is reported via optical microscope or polarized optical microscope (POM).[21, 99–111] Copolymers of PCL with L-lactide acid[21] poly (2-hydroxyethylmethacrylate)[111] are used to study the effect of chain length in crystallization. In double crystalline PCL-co-PLLA copolymers, PCL segments are trapped in PLLA spherulites. As PCL chain length increases, a non-banded structure is formed, as shown in Figure 7 .1.[21] Grafting of poly(2-hydroxyethyl methyl acrylate) branches on PCL decreases the overall crystallization growth of crystals.[111] Blends of PCL/PLLA[100,110] and PCL/polyvinyl butyrate (PVB)[109] and PCL/polyvinyl acetaldehyde (PVAC)[109] are studied for crystallization behavior. The immiscible blends of PCL/PLLA showed an exceptional morphology when heated above $T_m$ of PLLA and quenched to $T_c$ of PCL. PCL is inserted into concentric ring band of PLLA during melting and a wrinkle pattern, along with a small concentric ring band, is induced in blends as shown in Figure 7.2.[110]

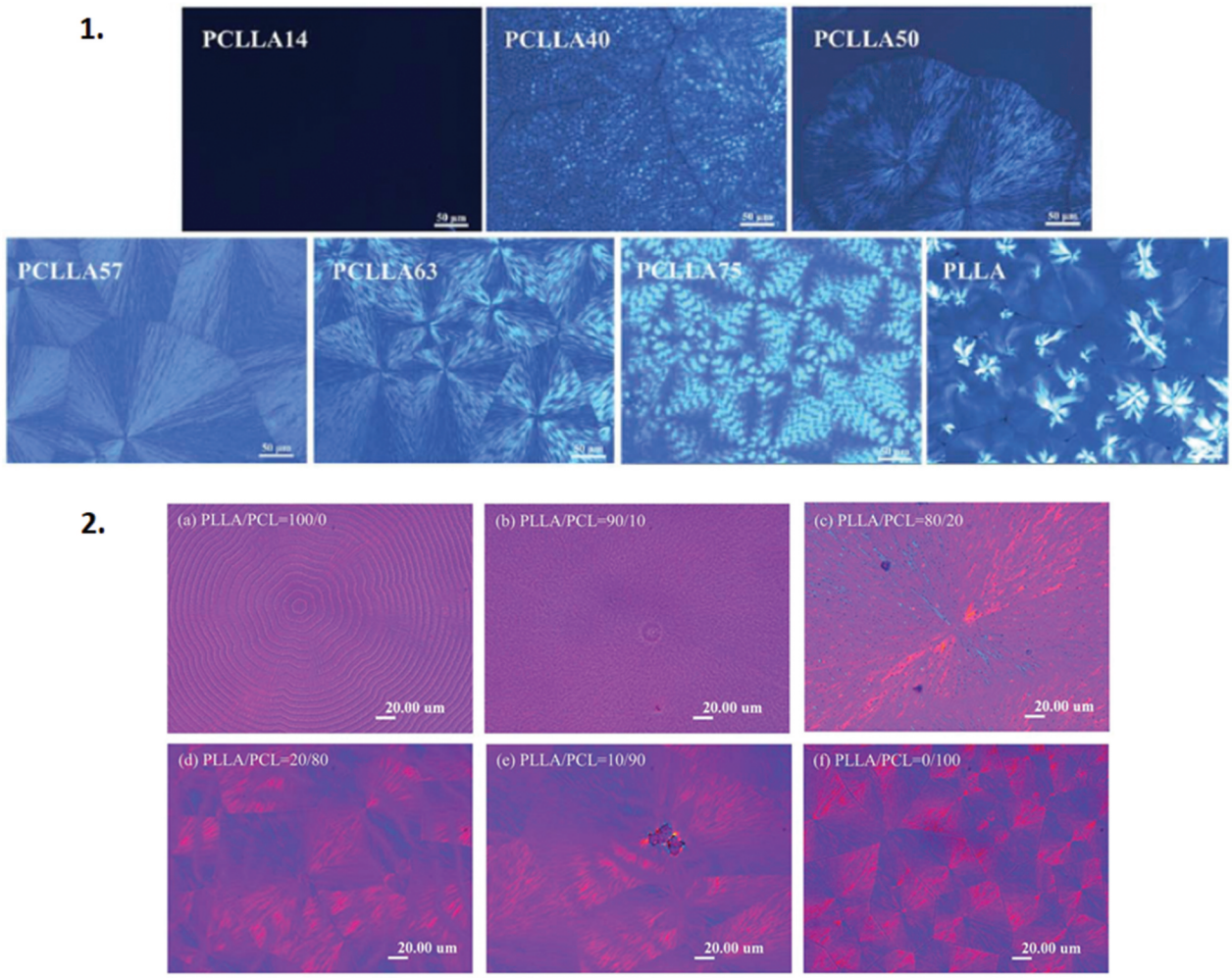



**Figure 7.** POM of crystals growth in PCL/PLLA copolymers (1) reproduced with permission [21] 2017, RSC; PCL/PLLA blends (2) reproduced with permission [106] 2014, RSC.

PVB and PVAC are functioned as nucleating agents in blends with PCL. Their addition increases the spherulite size and reduces the crystallization temperature of PCL.[109] PCL composites with ZnO,[105] Oleic acid,[105] glycerol monooleate(GMO),[105] hydroxy apatite,[104] clays,[101,104] and babassu[99] are reported for crystallization kinetics. In small amounts, organic additives, oleic acid and GMO are acted as nucleating agents and increase the crystallinity, while higher amounts works as plasticizer. Inorganic additives such as ZnO, hydroxyapatite and clay increases the crystallinity and crystal growth rate due to nucleating effect.[104] In PCL/babassu composite system, significant effect on crystallization parameters is not observed with addition of filler.[99]

### 3.2. Mechanical properties

PCL having number average molecular weight ($M_n$) of 80,000 g/mol, showed extremely ductile behavior with an average elongation of 450%.[50] The ultimate tensile strength, Young's modulus and yield strength of PCL is reported as 20 MPa, 374 MPa, and 16 MPa, respectively.[88] The copolymers, blends and composites of PCL in the form of films, porous structures, 3D structures and electrospun fibers are widely studied for mechanical properties. Table 4 represents data of mechanical properties with the addition of co-monomers, polymers and fillers in PCL.

Crosslink density and crystallinity are the major factors that affect mechanical properties of copolymers. Crystalline PCL is served as physical cross-linked points, which contribute to mechanical strength, thus, higher crystallinity results in higher tensile strength. Covalently crosslinked networks of PCL-co-BA[33] and PCL-EVA graft copolymers[36] showed a decrement in mechanical properties. Copolymers with LLA, [28,29,32] dimethyl siloxane, [23,112] cellulose fiber[113] and dioxanone[31] are reported with improved mechanical properties. The copolymers with *p*-dioxanone show peculiar behavior, and their E and TS are reduced with increased content of dioxanone, but elongation initially decreased then suddenly increased drastically from 930% for pure PCL to 450% for 87:13 ratio and 1690% for 80:20 ratio copolymers.[31]

The mechanical properties of polymer blends are dependent on the miscibility or compatibility between components along with crystallinity. Good impact properties of PCL make it a favorite polymer to be blended with other polymers to enhance their mechanical properties. PCL blends with PLA showed an improvement in mechanical properties despite their incompatibility due to the ductility of PCL, whereas PLA is very brittle (impact energy < 2.5 kJ $m^{-2}$).[46] Crystallinity has a great impact on mechanical properties that is studied by Bai *et al.* in PLA/PCL blend.[43] An improvement in mechanical properties are observed when compatibility is improved by functionalization or addition of compatibilizer for incompatible blends like Lignosulphate,[50] PLA,[47] PHB. A compatible blend of PCL/pine resin showed improved mechanical properties with increased content of PCL.[48] The mechanical properties of porous scaffolds prepared by PLLA/PCL blend are highly dependent on pore size. The scaffolds with larger pore sizes showed lower mechanical properties due to reduced polymer content.[55] Gelatin/PCL blend films showed compatibility and improved mechanical properties.[87]

Polymer composite is a renowned method for enhancing the mechanical property of any polymeric system. Factors like dispersion, morphology, and aspect ratio of fillers play a major role in mechanical properties. Composites of PCL with different types of carbon base fillers, [79,83,84,97] silica base fillers, [66,68,73] natural waste byproduct, [64,67,69,70,76,86] minerals, [49,71] alumina, and nibonium pentaoxide[62] are studied for their mechanical behavior. Carbon filler in the form of MWCNT/CNT and graphene is incorporated in PCL matrix. The carbon base fillers have improved the mechanical strength of composites due to reinforcing effect of nanofillers. Thye addition of different types of clays such as Cloisite $Na^+$, Cloisite 30B and Cloisite 20B in PCL matrix improved Young´s modulus (E) while the tensile strength (TS) remained almost constant and the elongation at break decreased in the range of 25–36%. PCL/SiC composite showed increased E, yield stress, TS, and elongation when added up to 4%.[66] Incorporation of alumina and niobium pentoxide is not able to enhance any mechanical properties. Tensile strength, elongation at break and impact strength decreased with filler addition due to agglomeration, while Young's modulus is unaffected due to inefficient chemical interaction at the filler–matrix interface.[62] Composites based on natural waste by products and minerals showed reduced mechanical properties, including TS, % elongation and impact strength due to poor interaction with PCL matrix. Whereas palm fiber showed reinforcement effects due to fiber-matrix interfacial compatibility.[64] The naturally occurring clay mineral halloysite nanotube (HNT) is, improved the TS and E of the PCL/HNT composite due to uniform dispersion of HNT, its hollow tubular structure, low hydroxyl group density on the surface, high aspect ratio and good compatibility of the HNT with PCL. Increased elongation at break results is uncommon behavior for PCL/HNT composite than pure PCL due to strong strain hardening of the PCL matrix endorsed by stiff HNTs.[68] A good study is

available for micro and nanocomposite of hydroxyapatite indicates that mechanical properties are improved and identical for both nano and micro composite till they are dispersed well in the matrix.[65]

The incorporation of PCL-grafted cellulose, even in small amounts, has increased the mechanical properties of PCL nanofiber mats. The improvements in mechanical behavior are attributed to the higher molecular weight (longer chains) of the grafted PCL chains and respective decrease in the diameter of the fiber.[34] Improved mechanical properties are observed for electrospun PLA/PCL blends.[57] Enhancement in E and TS of electrospun PCL mats after adding natural filler nanofibrilliate chitosan is due to the reinforcing effect of embedded chitosan nanofibrils oriented along the fiber axis, uniform dispersion and potential intermolecular interactions. Lower elasticity of the nanocomposite is in correlation with stiffening of chitosan.[81] When electrospun regenerated cellulose nanofibers are used in composite, E and TS increase with addition of nanofiber up to 17%, due to good compatibility.[86]

The copolymers, blends and composites of PCL in the form of films, porous structures, 3D structures and electrospun fibers are widely studied for mechanical properties.

### 3.3. Electrical properties

PCL is considered an insulating polymer having an electrical conductivity of $4.96 \times 10^{-13}$ S/cm. A copolymer of PCL with aniline induce the conductivity of 0.62 S/cm and its nanofiber showed conductivity of 0.03 S/cm.[114] Electrical conductivity data for blends are not available because blends of PCL are not intended to have an application area of electrical conductivity. Incorporation of some conductive filler improved the conductivity of PCL composites. When MWCNT contents added up to 0.3 wt%, the electrical conductivity of composites dramatically increased to $1.67x10^{-[5]}$ S/m. The fine spherulite and decreased crystallinity of PCL caused by MWCNT-induced effect favored the conductive properties of composites.[79] The conductivity of PCL is elevated by around 12-fold magnitude by inclusion of 5 wt. % chemically modified graphene in matrix.[97] Saravanamoorthy *et al.* have electrical properties of PCL/ hexagonal molybdenum oxide (h-$MoO_3$) nanocomposite and reported that the conductivity of the nanocomposites increased with addition of h-$MoO_3$ .[77] No change in the dielectric constant value is observed at low frequency as well as on the high frequency for pure PCL, whereas dielectric loss of the PCL/h-$MoO_3$ nanocomposites decreases with the increase in frequency (100 Hz to 1 MHz) and is showed better dielectric loss than pure PCL.

### 3.4. Structural properties

A well-defined semi-crystalline PCL crystal structure is studied by the X-ray diffraction method. PCL shows three strong peaks at the 2θ angles of about 21.4°, 22.0° and 23.7°, corresponding to the of diffraction planes (110), (111) and (200) of the orthorhombic unit cell, whereas lattice constants are a = 7.45 Å, b = 4.98 Å, and c = 17.05 Å.[88] Copolymers of PCL-co-PEG[115] showed a combined peak of both. Similarly, PCL-co-PLLA showed peaks of both PCL and PLLA.[116] With the increased content of PCL, its peak dominated other polymers' peaks, as shown in Figure 8.1.[115] When analysis is carried out at 60°C, its peaks get diminished due to loss of crystallinity above melting point.[116] The characteristics peaks are diminished with increasing content of PDMS in PCL-PDMS-PCL triblock copolymer.[117]

The XRD scans of the polymer blend are used to determine its homogeneity and overlap. Mixing PCL with miscible amorphous polymer doesn't influence its crystalline structure. If two crystalline polymers have low compatibility, then each polymer would have its individual crystal region in the blend. In this case, XRD showed distinct peaks of individual polymers with different lattice constants. In other cases if crystalline polymers are compatible or miscible, they will show enhanced degree of crystallinity of blend than pure polymers. XRD patterns of PCL/PMMA showed miscibility by merging the amorphous hump of PMMA and distinct crystalline peaks of PCL, as shown in Figure 8.2. The incorporation of PMMA did not influence crystalline structure of PCL.[61] A similar peak pattern is observed in PCL/TPU blend, where broad diffraction peaks ranging from 18° to 24° are replaced by sharper peaks with addition of PCL.[118]

Degree of crystallinity of PCL is increased by addition of silica carbide as observed in wide-angle XRD diffractogram as shown in Figure 8.3.[66] Incorporation of citric acid modified zinc oxide (ZnO-CA) nanoparticle reduces the intensity of PCL diffraction peaks due to weak hydrogen bond interaction. This weak interaction restricted the movement of PCL chain to some extent, making it difficult for PCL to crystallize.[96] Seyrek *et al.* have claimed that PCL/MMT composites exhibit diminishing peaks with inclusion of clay, indicating exfoliation of filler in polymeric system due to absence of

characteristic peak of MMT around 5.00° with distinct crystal structure of PCL.[98]

There is a limited study available for structural properties of copolymers, blends, and composites of PCL. The change in structural properties of PCL can be easily studied by an XRD diffractogram. Although researchers have prepared copolymers of PCL with amorphous monomers such as acrylates, acetates and semicrystalline monomers such as PDL, p-dioxanone, HCA and cellulose, their structural properties have not been evaluated. XRD is method to evaluate compatibility in blends along with thermal properties. XRD will provide better insight of PCL blends with pine resin, PGA, lignosulphonate, and PHBA for their compatibility and structural stability. XRD, along with TEM, is a renowned method to analyze dispersion of filler in composites. To attain a better understanding of PCL composites with natural waste by-products, minerals, carbon fillers need to analyze for structural properties as well.

### 3.5. Rheological properties

Rheological studies give an idea about processing parameters, mechanical and thermal behavior, and degradation conditions. Rheological properties of polycaprolactone are completely dependent on molecular weight, molecular weight distribution, molecular structure (linear/cyclic/star/branched) . [119–121] Pure PCL shows Newtonian behavior at lower frequency (less than 10 Hz) and shear thinning at higher frequency range (above 10 Hz). A very beautiful analysis of stress relaxation behavior on different molecular weight based films and scaffolds of PCL was carried out by Sethuraman *et al.* .[122] The higher molecular weight samples exhibit higher elasticity than lower ones. Films showed higher stress relaxation than scaffolds. The viscoelastic properties of scaffolds depend on pore size, structure and type of processing used. Copolymers of PCL and HPMA showed Newtonian behavior with complex viscosity of 130–140 Pa.s at 160°C.[36] Copolymer of PCL and EVA showed intermediate values of complex viscosity with respect to their neat counterparts.[36] Aqueous solution of PCL- co-*N*-2-Hydroxypropyl methacrylamide showed shear thinning behavior as studied by Rani *et al.*[26] Melt rheology of PCL-r-PLLA copolymer showed zero shear viscosity and relaxation time in between the neat polymers.[28]

The blends of PCL with $M_n$ 84,000 g/mol with PLA of different molecular weights ranging from 55,350, 92,500, and 106,900 g/mol showed interesting results for rheology. In the high molecular weight PLA (106,900 g/mol) system, storage modulus (G') and loss modulus (G") decreased with higher PCL content, as PCL is a less viscous component. In contrast, for the low molecular weight PLA (55,350) system, it is noticed that G' and G" of blends increased with an increase of PCL phase.[56]

The sandwich composite of gelatin film with PCL matrix shows higher storage modulus as compared with pure PCL throughout investigated temperature range. This is primarily attributed to an increase in stiffness of the matrix due to their enforcing effect imparted by PCL-gelatin films composite that permitted a higher degree of stress transfer from the matrix to the gelatin.[87] The sandwich structure of PCL and cellulose nanofiber showed that storage modulus is increased throughout all temperatures between −110 and 55°C in respect of pristine PCL, compensating for decrease in storage moduli at glass transition temperature and melting temperature in respect of PCL.[86]

The PCL/HNT nanocomposites are studied for melt rheological behavior to get an idea of microstructure in the melted state and its behavior at the processing time.[68] The storage modulus and loss modulus values of nanocomposites showed an increment when compared with neat PCL due to the homogeneous dispersion of HNT and good interphase attraction between PCL matrix and HNT. Viscoelastic properties of hydroxyapatite micro and nanoparticle-based PCL composite are examined at dry and wet conditions at 37°C. The stiffness of materials is increased under a dynamic load with HA concentration.[65] Melt rheology of nanocomposites with graphene, [123,124] nanoclays[123] is suggested that with increase in size and percentage of filler, viscosity increases until size of nanofiller (graphene) is less than radius of gyration of PCL molecule. Detailed analysis of PCL/cellulose nanocrystal is carried out by Wang *et al.* for linear rheology, non-linear rheology, creep behavior and transient rheology. Here particle–particle interaction of cellulose nanocrystals in PCL matrix affects the viscosity, creep behavior, and network formation.[125]

The processing parameters are perceived by flow properties like flow curve and complex viscosity. Detailed rheological analysis such as linear viscoelastic region, frequency sweep, creep recovery, structure recovery, stress relaxation, time–temperature superposition is a tool to understand behavior of materials during application and utilization. There are very few and limited studies available for rheological properties. A huge gap needs to fill by study of detailed rheological properties of all PCL copolymers, blends and composites that are synthesized.

### 3.6. Shape memory properties

Shape memory polymers (SMP) have ability to memorize a permanent shape and fixed to temporary shape under specific triggers. These triggers include temperature, pH, electric and magnetic field and light. Thermoresponsive shape memory polymers (TSMP) exhibit changes in shape from a deformed position to their original shape induced by temperature. TSMP changes into another shape or temporary shape when heated above transition temperature, deformed under load and cooling conditions. The polymer maintains this temporary shape until it is heated again without any load and regains its original shape. Strain fixity rate describes the ability to switch segments to fix/hold the temporary mechanical deformation.

$$R_f(N) = \frac{\varepsilon u(N)}{} \quad (1)$$

The strain recovery rate describes the ability of the shape memory material to recover its permanent shape:

$$R_r(N) = \frac{-\varepsilon p(N)}{-\varepsilon p(N-1)} \quad (2)$$

where N is cycle number, $\varepsilon_u(N)$ is strain in the stress-free state after the withdrawal of tensile stress in the $N^{th}$ cycle, $\varepsilon_m$ is the maximum strain executed on the polymer and $\varepsilon_p(N)$ and $\varepsilon_p(N\text{-}1)$ are the strains of the sample in two successive cycles in the stress-free state before yield stress is applied.

The semi-crystalline nature of PCL makes it a promising candidate for thermoresponsive shape memory applications. Along with crystallinity, the crosslinking network in PCL helps to store elastic energy and provides driving force to recover shape when heated above crystalline melting temperature ($T_{cm}$). The shape fixing ability of shape memory polycaprolactone is found to be influenced by the PCL-diol molecular weight. At the same time, the $R_r$ reaches 99% when heated above melting point irrespective of the length of PCL diol.[126] It is due to the stored energy in crystalline region, which is released faster than stimulated in their $T_{cm}$ range and allows the chains to recover their original shape with less time.

Shape fixity and shape recovery properties of copolymers, blends and composites are listed in Table 5. The shape memory behavior of PCL is tailored by copolymerizing it with L-lactides[127] butyl acrylates,[33] PDMS,[23,112] cellulose and its derivatives. [113,128,129] PCL-PDMS-PCL[23] and PCL-BA[33] copolymers showed increased shape fixity with an increased chain length of PCL. Copolymers composed of PCL with a molecular weight of 3000 g/mol or lower are exhibited no transition temperature, no crystallinity and act as amorphous phase. In this case, $T_g$ and crosslinking of copolymer help in shape recovery and shape fixity. [112,113,129] A 4D printed drug-loaded vascular stent prepared by grafting PCL on $\beta$-cyclodextrin showed good shape memory properties, which made it self-expandable.[130]

The blends of PCL for change in thermoresponsive behavior include thermoplastic polyurethanes,[118] diglycidyl ether of bisphenol A (DGEBA), [131] polyvinyl chloride,[132] olefin block copolymers,[133] polymethylvinylsiloxane, and natural rubber.[134] The shape memory properties of blends depend on PCL content, its crystallinity and compatibility. The strain recovery ratios of the NR/PCL blends are high, nearly 100% regardless of PCL content, due to stress release after sample molding, though shape fixity increased with increased PCL content.[134] DGEBA/PCL blend showed a different pattern. Shape fixity decreases with increase in PCL content with each consecutive cycle.[131] In PCL/

**Table 5.** Shape memory properties of copolymers, blends, composites of PCL.

| S.no. | Polymers | Synthesis mode | Composition | Shape fixity | Shape recovery | Ref |
|---|---|---|---|---|---|---|
| 1 | PCL-PDMS-PCL | Triblock copolymer | $PCL_{50}$-$PDMS_{37}$-$PCL_{50}$ | 99.7 | 98.3 | [23] |
| 2 | PCL-PDMS-PCL | Triblock copolymer | $PCL_{40}$-$PDMS_{20}$-$PCL_{40}$ | 99.9 | 98.6 | [112] |
| 3 | EC-PCL | Graft copolymer | EC M70-PCL1400 mixed with EC M70-PCL 7600 | 97.3 | 94.8 | [129] |
| 4 | MC-PCL | Graft copolymer | MC30K – PCL3K | 91.8 | 92.6 | [113] |
| 5 | *n* BA-PCL | Block Copolymer | PCL-BA (75:25) | 77.0 | - | [33] |
| 6 | PCL/OBC | Melt blend | 40/60/ 1 phr DCP | 89.0 | 99.8 | [133] |
| 7 | PCL/NR | Melt blend | 50/50/ 0.5 phr | 98.9 | 100.0 | [134] |
| 8 | PCL/TPU | Melt blend | 75/25 | 90.0 | 100.0 | [118] |
| 9 | PCL/DGEBA | Solution blend | 50/50 | 88.0 | 99.2 | [131] |
| 10 | PCL/PVC | Solution blend | 30/70 | 98.0 | 92.0 | [132] |
| 11 | PCL/MWCNT | Solution composite | 99.4/0.6 | - | 98.0 | [138] |
| 12 | PCL/HA | Solution composite | 100/5 | 90.0 | 90.0 | [135] |
| 13 | PCL/cellulose | Solution composite | 100/3 | 90.0 | 93.3 | [137] |

TPU blend high recovery ratio with increase in TPU content is observed. TPU domain stored the deformation energy and contributed to the strain recovery, while the ductile and crystalline PCL contributed to the strain fixing.[118] A physically crosslinked OBC/PCL blend compatibilized with dicumyl oxide showed two distinct peaks of $T_m$ (~ 55°C and ~120°C) and $T_c$ (~27°C and ~96°C). The blend showed good shape memory properties at 65°C.[133]

Shape memory properties of the composite of PCL studied for hydroxyapatite,[135] titanium nitride,[136] nanocrystalline cellulose,[137] calcium carbonate,[127] MWCNT[138] and POSS.[139] A common observation is obtained from strain fixity and strain recovery values of composites that they are independent of filler content and dependent on PCL chain length. Qu *et al.* coated MWCNT over PCL/TPU wet spun fiber. The composite fibers showed a self-healable thermo-electrical double response shape memory effect.[140]

The copolymers or blends with sharp transition temperature ($T_{trans}$) peaks are desirable for good shape memory properties. Broad $T_{trans}$ peak affects the SMP properties negatively, as seen in case of cellulose-g-PCL, which showed inferior strain fixity and strain recovery.[113] The structural and thermal properties are altered while copolymerization, blending or composite preparation of PCL. This further affects the transition temperature and crystallinity of the material, accordingly shape memory properties and its testing parameters vary.

### 3.7. Biodegradation

Biodegradation of aliphatic polyesters is associated to their physical and chemical properties, e.g. hydrophilicity, molar mass, $T_g$, $T_m$, degree of crystallinity, chemical structure and surface area.[36] In biodegradability, degree of crystallinity plays a significant role, as amorphous domains of a polymer are mainly attacked by enzymes. The enzymatic degradation takes place on surface of the PCL. Hydrophobic nature of PCL makes it hydrolytic degradation very slow.[141] Feng *et al.* have carried out enzymatic depolymerization of polycaprolactone involves two elementary steps: triad assisted nucleophilic attack and C-O bond cleavage.[142] The C-O bond cleavage is considered as rate determining step of degradation with average energy barrier of 15 kcal/mol. Almeida *et al.* have studied quantum mechanics and molecular mechanics molecular dynamics for enzymatic hydrolysis of PCL by thermophilic esterase.[143] The degradation of PCL copolymers, blends and composites can be divided into two parts: 1) soil burial and 2) simulated environment.

Soil burial degradation of PCL composites with silicon carbide,[66] clay,[73] gelatin[87] are studied. It is observed that cloisite $Na^+$ and cloisite 20A-based nanocomposites showed higher weight loss in comparison to pure PCL though incorporation of cloisite 30B slowed down the rate of degradation of the polymer attributing the fact that the presence of reinforcement hindered the access of the microorganisms attack on ester groups of PCL.[73] Presence of SiC nanoparticles in PCL matrix increases the crystallinity biodegradation process becomes slow.[66] PCL/gelatin sandwich composite degrades faster than pure PCL due to high water affinity of gelatin.[87]

Hydrolytic biodegradation of PCL copolymers with *p*-dioxanone,[31] 4-hydroxycinnamic,[27] EVA,[36] aniline,[114] blends with PLLA[54,144] and composites with bioactive glass,[145] cellulose nano-whiskers[76] are studied. As hydrophilicity is effective factor in biodegradability, it is observed in CL-co-dioxanone copolymer showed higher mass loss and reduced molecular weight with increase in content of *p*-dioxanone.[31] The crosslinked structure induces hydrophobicity, resulting in slower degradation rate as seen in poly(ε-caprolactone)-co-poly(4-hydroxycinnamic acid).[27] PCL-co-polyaniline nanofibers underwent rapid mass loss of almost 45% in 35 days in phosphate buffer solution (PBS).[114] Degradation rate is much higher in alkaline conditions than normal PBS of pH 7.4, as observed in semi-IPN blends of PCL with PLLA. Increased content of PLLA is responsible for more water uptake and showed more hydrolytic degradation.[54] PCL blend with EVA, a conventional non-biodegradable polymer showed decreased biodegradability.[36] The biodegradation rate of PCL composites with natural filler cellulose whiskers is increased because of higher hydrophilicity of the latter than pure PCL.[76] The PCL/bioactive glass nanocomposite coating on magnesium substrate showed more degradation than pure PCL due to dissolution of bioactive glass in simulated body fluid .[145]

In-vivo degradation of P(CL-r-LLA) copolymer is investigated with varying content of LLA. These copolymers are hypodermically implanted in the backs of rats. The degradation of P(CL-r-LLA) is determined as a function of the implantation time and monitored by weight loss, changes in molecular weight and macroscopic observation. The results showed the degradation and absorption of copolymers eliminate the necessity for removal of the materials implanted *in vivo* when used as a drug matrix.[32] Enzymatic degradation of PCL blends with pine resin,[48] epoxidized soyabean oil (ESO)[146] and poly(β-hydroxybutyrate)[52] are studied. The pine resin mainly consists of triglycerides that is easily degraded by lipase. High enzymatic hydrolysis rate of

pine resin is enhanced by the overall hydrolysis of blend in the presence of lipase from *porcine pancreatic*.[48] PCL/ESO blends are immersed in the mixture of lipase from *Pseudomonas cepasia* and buffer solution of PBS. Almost 40% weight loss is observed due to degradation of PCL content only and no mass change is observed for ESO in 5 days.

High Ductility and plasticity of PCL with a slow degradation rate help to counterbalance the expeditious degradation of natural polymers.[18] Addition of hydrophilic polymers/fillers via copolymerization or blending increases the degradation rate. The addition of crystalline filler of semicrystalline polymers increases the overall crystallinity of material and reduces the weight loss during degradation.

## 4. Applications

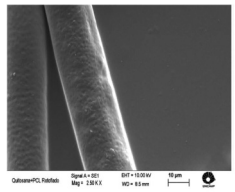
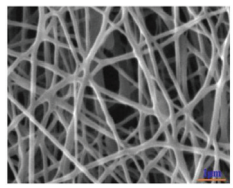
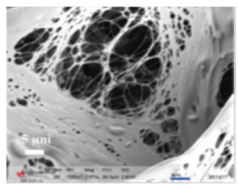
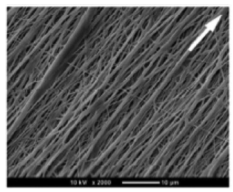
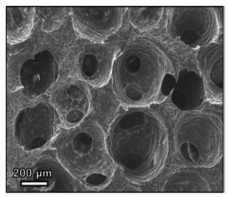
### 4.1. Biomedical applications

PCL is one of the most preferred and well-explored thermoplastics, which received much attention because of its adaptability, biodegradability with outstanding biocompatibility *in vitro* and *in vivo* for biomedical applications. This includes skin tissue engineering, bone tissue engineering and drug delivery. Scaffolds are integral to their regenerative process, assist adhesion and spreading of cells, support their growth, and successively activate their development in specific tissues.[147] Review articles emphasizing PCL-based scaffolds prepared by electrospinning for tissue engineering, drug delivery and wound dressing materials have been reported recently.[3,13,17] Suwantong has discussed about factors affecting electrospinning, such as polymer concentration, molecular weight of polymer, solution conductivity, solvent volatility, applied voltage, flow rate and ambient parameters.[13] The major scaffold fabrication techniques for PCL are summarized by Dwivedi *et al.* including solvent casting, porogen leaching, phase separation, 3D printing and electrospinning for bone scaffolds.[2] Prasad *et al.* have discussed about fused deposition technique used in fabrication of scaffolds for PCL composites using 3D printing.[15] Hajiali *et al.* have focused only on PCL/ calcium phosphate-based ceramic and PCL/bioactive glasses composites for bone tissue engineering.[14] Chan *et al.* have provided a systematic review about tracheal stents and splints patches to be used in tracheal surgeries.[1] Salehi *et al.* have provided an insight in the use of PCL in corneal tissue engineering.[10] In this review, they discussed PCL-based scaffolds prepared by different techniques to be used in epithelium, stroma and endothelium of cornea. Different techniques used for scaffolds fabrication such as salt leaching,[147] rotary jet spinning,[148] electrospinnig,[149] aligned electrospinning,[150] supercritical $CO_2$ batch foaming,[151] porogen leaching,[152] sol-gel method,[153] 3D printing,[154] controlled humidity,[155] unidirectional freeze drying[156] and hot embossing[157] are presented in Table 6.

Apart from tissue engineering, drug delivery, and wound healing, other bio-medical applications of PCL copolymers, blends and composites are bio-imaging and coatings for bio-implants. Bio-imaging is a useful tool to detect cancer and tumor. Various nanoparticles of PCL copolymers are synthesized to act as a carrier for bio-imaging. PCL-co-polyamidoamine linear dendritic block copolymer nanoparticles as a small-molecule carrier for Rhodamine B, curcumin and indolizine cyanine dye in NIR imaging and photothermal therapy are synthesized by Chandrasiri *et al.* .[158] Block copolymers of PEG and PCL[159–161] and four-arm PEG-PCL copolymers[162] are synthesized and fabricated with dyes,[159] cyclo(Arg-Gly-Asp),[160] polythiophene[161] and porphyrin[162] for bio-imaging applications. Organic fluorophore synthesized by incorporating di (thiophene-2-yl)-diketopyrrolopyrrole in the middle of polymer chain of PCL by Huang *et al.*[163] A fluorescence backbone, macroinitiator conjugated polymer poly(fluorene-alt-(4,7-bis(hexylthien)-2,1,3-benzothiadiazole)) (PFTB) grafted with PCL, poly oligo (ethylene glycol) methyl ether methacrylate] (POEGMA) blocks[164] and carborane (PmCbA) containing triblock copolymer POEGMA-PmCbA-PCL conjugated with a NIR fluorescence probe[165] are fabricated to be used in bio-imaging for effective treatment of cancer cells.

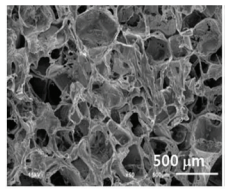
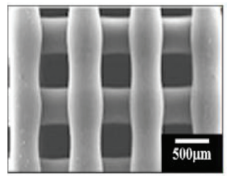
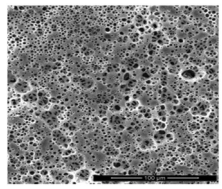
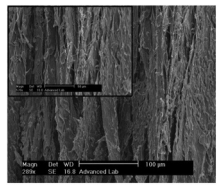
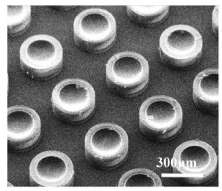
Nowadays nanocomposite coatings based on PCL are preferred to overcome inferior abrasion and wear resistance of bioimplants made-up of metal alloys and stainless steel. PCL/Laponite nanocomposite coating by electrophoretic deposition method showed great improvement in corrosion resistance than uncoated stainless steel and makes it applicable for bone implant.[166] Jokar *et al.*[167] and Shafiee *et al.*[168] have used the dip-coating method to coat stainless steel with PCL/gelatin and PCL/forsterite ($Mg_2SiO_4$) nanocomposites for improved bioactivity and corrosion resistance. Titanium alloy coated with PCL nanofibers using electrospinning. Further, it is treated with $TiO_2$ nanotubes to give cell biocompatibility to orthopedic and dental implants.[169] PCL/ HA,[170] PCL/ bioactive glass[145] nanocomposite coating and PCL

**Table 6.** Scaffold types of PCL copolymer, blends, and composites for bio-medical applications.

| S. No. | Polymers | Synthesis mode | Fabrication method | Scaffold | Ref |
|---|---|---|---|---|---|
| 1. | PCL/chitosan | Hybrid membrane, Chitosan coating over PCL mesh. | Rotary jet spinning |  | [148] |
| 2. | PCL / Curcumin | Composite | Electrospinning |  | [149] |
| 3. | PCL/PLA | Blend | Supercritical $CO_2$ batch foaming |  | [151] |
| 4. | PCL-PTHF-PCL | Triblock copolymer | Aligned electro-spinning |  | [150] |
| 5. | PCL/CNT/ $TiO_2$ | Composite | Porogen leaching | | [152] |
| 6. | PCL/Bioglass | Composite | Sol-gel method |  | [153] |

(*Continued*)

**Table 6.** (Continued).

| S. No. | Polymers | Synthesis mode | Fabrication method | Scaffold | Ref |
|---|---|---|---|---|---|
| 7. | PCL/Magnetite nanoparticles | Nanocomposite | Salt (NaCl) leaching |   | [147] |
| 8. | PCL/silanted silica nanoparticles | Nanocomposite | 3D printing |   | [154] |
| 9. | PCL/calcium silicate | Composite | Controlled humidity |  | [155] |
| 10. | PCL/Zein | Composite | Unidirectional freeze drying |  | [156] |
| 11. | PCL/Furosemide | Composite | Hot embossing |  | [157] |

coating on $MgCO_3$ coated Mg[171] by dip-coating showed a decreased rate of degradation and improved bioactivity of Mg and Mg alloys substrate. Electrospinning is also used for coating PCL/zinc oxide[172] and PCL/MgO-Ag[173] nanocomposite on Mg substrate to improve biocompatibility and corrosion resistance. Huang *et al.* have coated Mg alloy with PCL/amino acid to enhance cytocompatibility using electrografting and dip coating method.[174]

### 4.2. Non-biomedical applications

Copolymers, blends, and composites of PCL are entering into non-biomedical applications, which includes packaging, self-healing & corrosion resistance coatings, flame retardant coatings, oil-water separation, plant grafting and photothermal absorption. Figure 9. shows different non-medical applications of PCL copolymers, blends and composites.

### 4.2.1. Packaging

Hydrophobicity of PCL is an advantage in packaging because it reduces water vapor transmission rate; however, it shows weak oxygen barrier properties.[175] This limitation is overcome by inclusion of other polymers or nanoparticles. Blend and copolymers of PCL/PLA improved the oxygen barrier and mechanical properties. Further, addition of zinc oxide, clove oil, cinnamaldehyde and impregnation of supercritical $CO_2$ enhances antimicrobial properties and makes it suitable for the packaging of button mushroom and scrambled egg. [176–178] Blends of PCL with Polyhydroxybutyrate (PHB),[179] PVC[180] and starch[181,182] showed improvement in oxygen barrier properties for food packaging. PCL mixed with other material such as pomegranate rind,[183] chitosan/grapefruit seed extract[184] (Figure 8), sodium metabisulpahte,[185] silver-kaolinite,[186] hydroxytyrosol/cloisite 30b,[179] titanium dioxide nanoparticles[187] and organically modified montmorillonite[98] to form active packaging material. These materials added not only antimicrobial properties but also enhanced mechanical, thermal, and biodegradation properties. Cai *et al.* have developed novel PCL composite films by mixing curcumin loaded zeolitic imidazolate framework, which showed pH and light responsive antibacterial activity.[188] Electrospinning, electrospraying[189] and microfluidic spinning technology (MST) are recent methods to fabricate improved active packaging material. It is found that CS/ OEO/PCL blended electrospun nanofibers have lower water vapor permeability than cast film and exhibit distinctive antibacterial activity toward Gram-positive/Gram-negative bacteria.[190] Mathiazhagan *et al.* have fabricated a food packaging film by soaking PCL nanofibrous mat in leaf extract of *Acalypha indica*.[191] These antibacterial nanofibrous mats show good microbial inhibition as compared to commercial polyethylene for carrot piece packaging. Films composed of fibers of nano-sized functional particles of silver loaded konjac glucomannan (KGM) and PCL blend prepared by MST exhibit excellent antibacterial properties for active packaging material.[192]

### 4.2.2. Self-healing

PCL is a well-known polymer for its shape memory properties that make it capable of self-healing.[193] Excellent shape memory and self-healing are achieved when linear PCL is incorporated in crosslinked PCL network and forms a simultaneous IPN network. Partial fractures generated in the film can heal up by this IPN network.[194] Blends of natural rubber and PCL

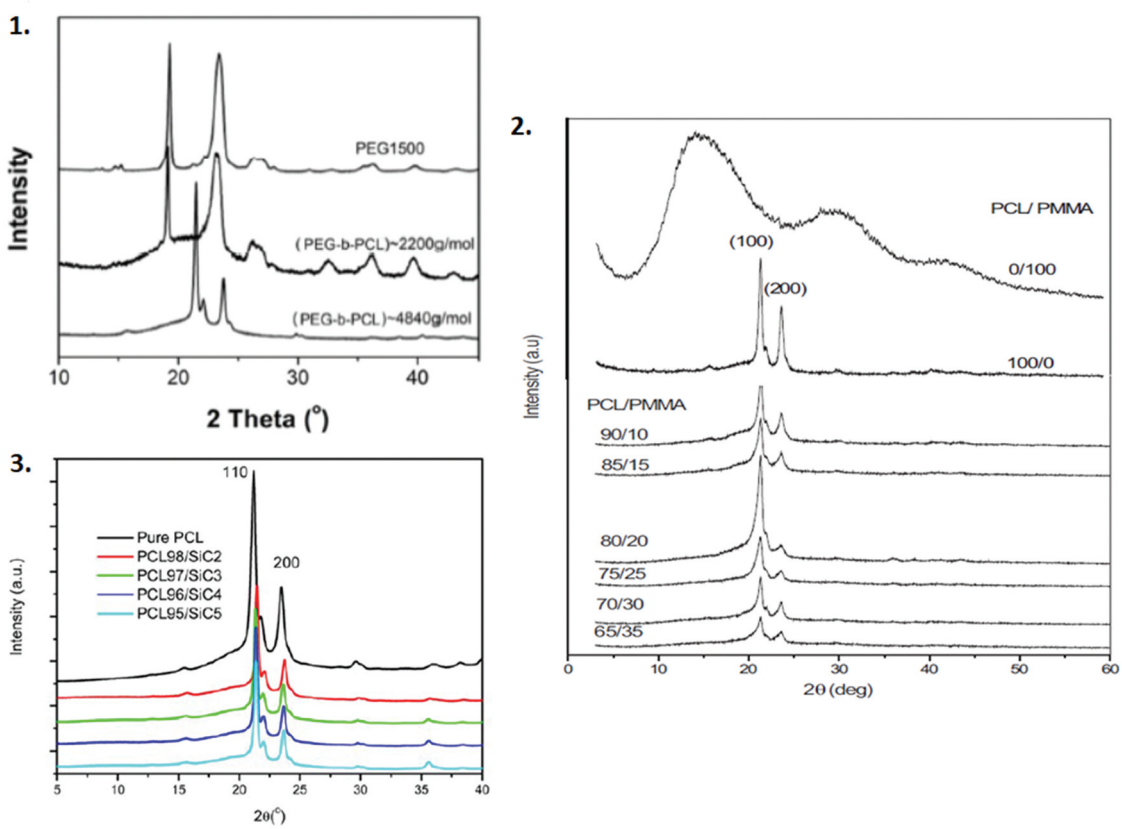



**Figure 8.** Structural analysis by XRD of PEG-PCL copolymer (1) reproduced with permission [115] 2015, RSC; PCL/PMMA blend (2) reproduced with permission [61] 2016, Elsevier; PCL/SiC nanocomposite (3) reproduced with permission [66], 2015, Wiley.

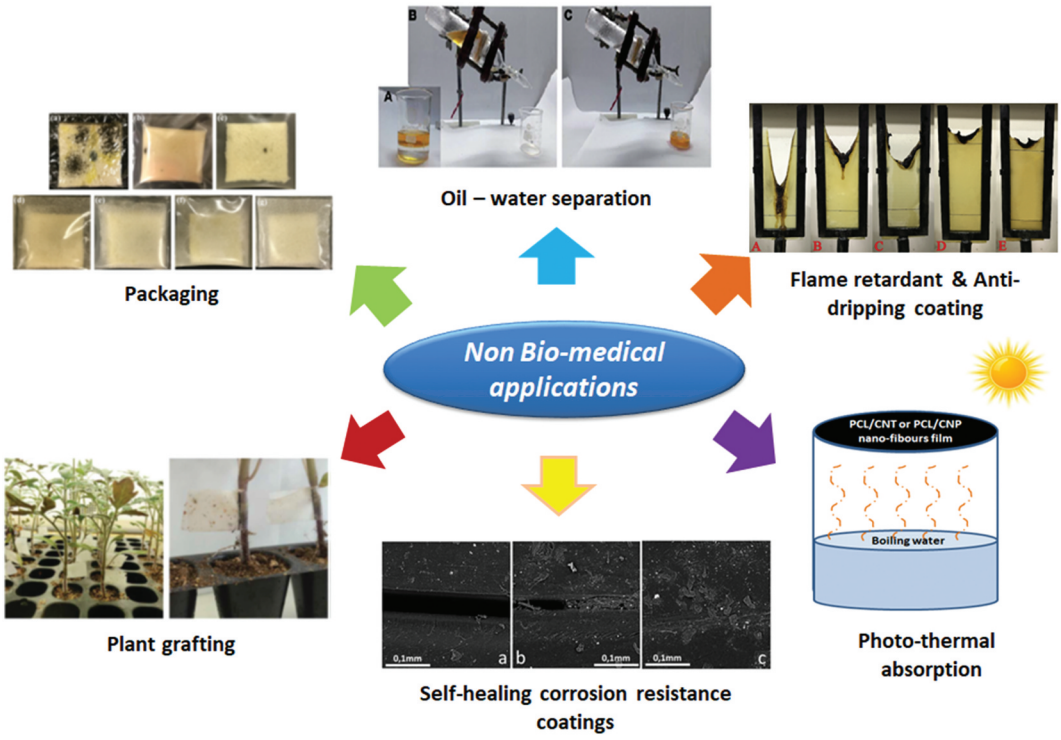



**Figure 9.** Non-biomedical applications include packaging reproduced with permission [185] 2020, Elsevier; self-healing coatings corrosion resistance coatings reproduced with permission [198] 2018, ACS; flame retardant coatings reproduced with permission [201] 2019, Elsevier; oil-water separation reproduced with permission [204] 2016, RSC; plant grafting reproduced with permission [207] 2017, TnF; photo-thermal absorption reproduced with permission [208] 2021, ACS of PCL copolymers, blends and composites.

are used for the application of shape memory-assisted self-healing properties.[195] A high-performance and self-healable PCL/polydopamine nanocomposites are fabricated.

The photothermal conversion nature of polydopamine provides self-healing functionality to the nanocomposite. On exposure to near-infrared (NIR) light, the temperature of nanocomposites swiftly goes above the polymer $T_{cm}$. This allows rapid NIR light-induced self-healing properties.[84] The nanocomposites comprised of PCL/polydopamine-capped reduced graphene oxide nanofiller[84] and TPU/PCL blend with multi-walled carbon nanotubes (MWCNTs) exhibit excellent self-healing properties in the presence of near-infrared (NIR) irradiation compared with the traditional heat-induced self-healing shape memory composites. NIR irradiation not only reduces the healing time (3 min.) but also selectively repairs the exposed damaged regions without distinct interference to the performance of surrounding parts.[196]

### 4.2.3. Coatings

A self-healing UV curable corrosion-resistant coating is prepared by acrylated polycaprolactone polyurethane, where soft segment is comprised of PCL and the hard segment is comprised of isobornyl acrylate (IBOA). Partial self-healing, corrosion protection and barrier properties are achieved at the melting temperature of PCL (at 60°C) and complete healing is observed at the $T_g$ of IBOA (115°C).[197] Microsphere of PCL and 8-hydroxyquinoline corrosion inhibitor are added to the epoxy matrix to improve self-healing and corrosion resistance.[198] Ranjitha *et al.* have studied PCL/functionalized graphene oxide (FGO) nanocomposite coating and found superior barrier and anti-corrosion properties by improving the dispersion and exfoliation of FGO in the PCL matrix.[199]

An elastomeric coating of hyperbranched-PCL/siloxane is applied on stainless steel panels and showed good antifouling property. The PCL segment inhibits marine biofouling and siloxane provides fouling release property due to good surface elasticity and low surface energy.[200] Low gloss organic montmorillonite/waterborne polyurethane nanocomposite (OWPU) coatings with polycaprolactone (PCL) and hydroxyl-terminated polybutadiene (HTPB) are prepared by Ding *et al.* and applied to polyvinyl chloride leather. Nanosheets of MMT have increased thermal stability of hard segments and soft segments derived from PCL and promoted cross-linking. Further, cyclization of HTPB at high temperature imparted flame retardant and anti-dripping properties to OWPU coatings.[201] For coating application Arya *et al.* have studied the effects of solvents, methyl chloride and

toluene on the basis of morphological properties, drying properties and cost of system.[202] They have reported that methyl chloride provides a smooth, dense and cheaper coating system in comparison to toluene. In case of PCL, methyl chloride and toluene are the most used solvents. So here, a detailed analysis made remarkable addition for the recommendation of solvent.

#### 4.2.4. Purification

For purification/filtration of water, PCL becomes a favorite polymer due to environmental concerns. A thiol-ended polycaprolactone is synthesized by Korpinar *et al*. to remove lead Pb(II) and cadmium Cd(II) ions.[203] Reshmi *et al*. fabricated a superhydrophobic, superoleophilic electrospun nanofibrous membrane of PCL/beeswax blend to separate oil from polluted water with great separation efficiency and good recyclability.[204] In another approach, Zhang, He and coworkers have developed composites based on superhydrophobic and superoleophilic membrane with a hierarchical microsphere structure by alternate electrospinning and electrospraying technology. This PCL/$SiO_2$ composite membrane showed superhydrophobicity, extreme oil-water separation efficiency (>99.8%), oil flux, stability and excellent oil adsorption performance.[205,206]

#### 4.2.5. Plant grafting

Eksiler *et al*. have developed films made of PCL/cis-1,4-polyisoprene blend as a biodegradable joining tool for plant grafting on fresh tomato plants. Soybean lecithin and acrylated-epoxidized soybean oil are used as a non-reactive compatibilizer due to poor compatibility of PCL/PI blend. These compatibilizers also accelerated the weight loss of the films in soil and increased the microorganism growth on the film. It is observed that all the spliced plants are grown without any defect at the end of two weeks.[207]

#### 4.2.6. Photothermal absorption

Chen *et al*. have fabricated PCL nanofiber composite thin films with carbon nanotubes and carbon nanoparticles to use it as photothermal absorbers. From neat PCL nanofiber thin films to composite thin films, the solar absorbance is recorded from 0.04 to 0.94 along with its good salt rejection properties while working in high salinity conditions.[208]

**Table 7.** Commercial products based on polycaprolactone.

| S. No. | Manufacturer | Polymer | Product name | Application |
|---|---|---|---|---|
| 1. | BASF, Germany | PCL | Capromer PD1-10<br>Capromer PD1-20<br>Capromer PD1-20A<br>Capromer PD4-05<br>Capromer PT1-05 | Textile and leather coatings. |
| 2. | Perstorp, Sweden | PCL/PCL copolymers | CAPA 1301<br>CAPA 2043–2803<br>CAPA 2077A- 2403D<br>CAPA 3022–3301<br>CAPA 4101–4801<br>CAPA 6100–6800<br>CAPA 7201A-7203 | Coating, adhesive, foam, orthopedic splints, dental impressions, oncology immobilization, food packaging, laminates, color master batch |
| 3. | Daicel Corporation, Tokyo, Japan | PCL | PLACCEL 200 series<br>PLACCEL 300 series<br>PLACCEL 400 series<br>PLACCEL CD series<br>PLACCEL F series<br>PLACCEL H1P | Resins |
| 4. | Haihang Industry co. ltd, Jinan City, China | PCL | Polycaprolactone | Controlled release drug carrier, cell and tissue culture medium, fully degradable plastic surgical suture line, High strength film filamentous molding, plastics modifier and plasticizer, medical modeling materials, industrial, art modeling materials, toys, organic colorants, thermal carbon ink adhesives, hot melt adhesive |
| 5. | Biogeneral, San Diego, USA | PCL/PCL-co-PDLL | PLC 70 | |
| 6. | Durect, Birmingham, USA | PCL-co-PDLL | | Fiber, films and tubing |
| 7. | Foster, Putnam USA | PCL | Polymedex | drug delivery, orthopedic, dental, and maxillofacial applications |
| 8. | Innovate Chemicals, Guangdong, China | PCL | INV PCL 6500<br>INV PCL 6800 | Surgical sutures, orthopedic splints radiation board, resin bandage, dental model, Hot melt adhesive, coating, ink, non-woven adhesive, shoes material, structural adhesive, blown films, laminated material, manual models, organic colorants, powder coating |

## 5. Commercial products

PCL is manufactured mainly by BASF, Perstrop, Daicel, Haihang, Foster and Innovate chemicals with brand names Capromer, CAPA, PLACCEL, Polycaprolactone and INV, respectively, with various molecular weight and functionality. Biogeneral and Durect provide PCL-co-PDLL copolymer with different monomer compositions. These products have suggested areas of application for dental adhesive, control release drug carrier, cell and tissue culture, surgical sutures, orthopedic splint, radiation board, resin bandage, plastic modifier & plasticizer, textiles and leather coating and ink. They are available in the form of powder, films, fibers, and tubes. Table 7 summarizes some of the commercial products based on PCL. The commercial PCL-co-PEG copolymers are reported by Dabbaghi *et al.*[8]

## 6. Future scope and conclusion

PCL polyester is easily available and relatively inexpensive. It can be modified to adjust its chemical and biological properties, physiochemical state, degradability, and mechanical properties by copolymerization, blending and composite formation. The synthesis methods of PCL copolymers, blends and composites are discussed in this review. PCL copolymers with natural polymers except cellulose are not reported yet. Blends by latex mixing and composites with filler like copper and silver are not visible in the literature. PCL-based materials are profoundly characterized for thermal, mechanical, structural and biodegradable properties. Though mechanical and thermal properties are studied in length; however, less attention has been paid to crystallization kinetics/behavior and structural properties. The detailed analysis of rheological behavior of PCL copolymers, blends and composites is also required consideration by researchers. The shape memory properties of PCL make it a promising material for robotics and anti-counterfeit application, which can further investigate for copolymers, blends, and composites. The data analyzed and discussed in this review about fabrication and properties allowed us to conclude that PCL copolymers, blends, and composites collectively provide an encouraging polymer platform to produce different materials.

Numerous researches available on applications have shown that PCL widens its horizon from biomedical to packaging, self-healing, coatings, separation membrane and plant grafting, still need to investigate more. Scientists try to develop films for photothermal absorption though their sorbent capacity is quite lower than previously reported adsorbents. Quite a good number of PCL copolymers and blends and composites are synthesized for various applications still; these are not commercialized much. Very few commercial products are reported for PCL homopolymers and copolymers, though commercialization of these extensively researched copolymers, blends, and composites is desirable. We hope that this review will prompt researchers to fill some of these gaps in the near future.

## Disclosure statement

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

## Notes on contributors

*Radha Sachan* did her Bachelor in Paint technology. After working in Chemical Industries for 6+ years she returned to academia. She did her Masters in Polymer Technology from Delhi Technology University and currently pursuing Ph. D. from Delhi Technological University in Discipline of Polymer Science and Technology. Her research interest includes structure-property relationship of polymeric material.

*Dr. S. G. Warkar* is a Professor & Head of Department of Applied Chemistry, Delhi Technological University. His areas of interest are biopolymer-based superabsorbent hydrogels and its applications in the field of agriculture, water enrichment, drug delivery, biodegradable polymers and benziporphyrins. He has 28 publications in reputed and high impact journals in last five years, and also presented his work in several national and international conferences. He is a life member of Indian Society of technical Education and reviewer of many journals of International repute.

*Dr. Roli Purwar* is currently working as Professor in the discipline of Polymer Science and Chemical Technology, Department of Applied Chemistry, Delhi Technological University. She has developed two technologies namely "Acrylic based formulations for floor" & "Formulation for improving the impact properties of recycled polypropylene and transferred to industries. She has three patents (3 Indian and 1 US) and published more than 60 papers in Journal of International repute and in Proceeding of national/ International Conferences.

## List of abbrevations

| | |
|---|---|
| PCL | Polycaprolactone |
| ROP | Ring opening polymerization |
| RROP | Radical ring opening polymerization |

| | |
|---|---|
| RAFT | Reversible Addition Fragmentation Chain-Transfer |
| MDO | 2-Methylene-1,3dioxepane |
| IPN | Interpenetrating polymer network |
| PLA | Polylactic acid |
| PLLA | Poly L-lactide acid |
| HCA | Hydroxycinnamic acid |
| CNT | Carbon nano-tubes |
| MWCNT | Multi walled carbon nano-tubes |
| $T_m$ | Meting temperature |
| $T_c$ | Crystallization temperature |
| $T_g$ | Glass transition temperature |
| TSMP | Thermoresponsive shape memory polymers |
| HNT | Halloysite nanotubes |
| $M_n$ | Number average moleculer weight |
| E | Youngs modulus |
| TS | Tensile strength |
| h-$MoO_3$ | Hexagonal molybdenum oxide |
| $T_{cm}$ | Crystalline melting temperature |
| HA | Hydroxyapatite |
| SiC | Silicon carbide |
| PVB | Polyvinyl butyrate |
| PVAC | Polyvinyl acetaldehyde |
| PGA | Polyglycolic acid |
| $CO_2$ | Carbon dioxide |
| PDS | Pamidronate |
| ZnO | Zinc Oxide |
| ZnO-CA | Citric acid modified ZnO |
| NIR | Near infrared |
| FDA | Food and Drug Administration |
| POEGMA | Poly[oligo(ethylene glycol) methyl ether methacrylate] |
| PmCbA | Carborane |
| ADH | Adult Human Dermal |
| $TiO_2$ | Titanium dioxide |
| PGA | Polyglycolic acid |
| TPU | Thermoplastic polyurethane |
| MMT | Montmorillonite |
| PHB | Polyhydroxybutyrate |
| IBOA | Isobornyl acrylate |
| FGO | Functionalised graphene oxide |
| OWPU | Waterborne polyurethane nanocomposite |
| HTPB | hydroxyl-terminated polybutadiene |
| PTHF | Polytetrahydrofuran |
| NaCl | Sodium Chloride |

## References


[1] Chan, D. S.; Fnais, N.; Ibrahim, I.; Daniel, S.; Manoukian, J. Exploring Polycaprolactone in Tracheal Surgery: A Scoping Review of in-Vivo Studies. *Int. J. Pediatr. Otorhinolaryngol.* 2019, *123*April, 38–42. DOI:10.1016/j.ijporl.2019.04.039.

[2] Dwivedi, R.; Kumar, S.; Pandey, R.; Mahajan, A.; Nandana, D.; Katti, D. S.; Mehrotra, D. Polycaprolactone as Biomaterial for Bone Scaffolds: Review of Literature. *J. Oral Biol. Craniofacial Res.* 2020, *10*(1), 381–388. DOI: 10.1016/j.jobcr.2019.10.003.

[3] Janmohammadi, M.; Nourbakhsh, M. S. Electrospun Polycaprolactone Scaffolds for Tissue Engineering: A Review. *Int. J. Polym. Mater. Polym. Biomater.* 2019, *68*(9), 527–539. DOI: 10.1080/00914037.2018.1466139.

[4] Woodruff, M. A.; Hutmacher, D. W. The Return of a Forgotten Polymer - Polycaprolactone in the 21st Century. *Prog. Polym. Sci.* 2010, *35*(10), 1217–1256. DOI: 10.1016/j.progpolymsci.2010.04.002.

[5] Guarino, V.; Gentile, G.; Sorrentino, L.; Ambrosio, L. Polycaprolactone: Synthesis. *Prop. Appl.* 2017. DOI: 10.1002/0471440264.pst658.

[6] Labet, M.; Thielemans, W. Synthesis of Polycaprolactone: A Review. *Chem. Soc. Rev.* 2009, *38*(12), 3484–3504. DOI: 10.1039/b820162p.

[7] Mandal, P.; Shunmugam, R. Polycaprolactone: A Biodegradable Polymer with Its Application in the Field of Self-Assembly Study. *J. Macromol. Sci. Part A Pure Appl. Chem.* 2020, *58*(2), 111–129. DOI: 10.1080/10601325.2020.1831392.

[8] Dabbaghi, A.; Ramazani, A.; Farshchi, N.; Rezaei, A.; Bodaghi, A.; Rezayati, S. S. Synthesis, Physical and Mechanical Properties of Amphiphilic Hydrogels Based on Polycaprolactone and Polyethylene Glycol for Bioapplications: A Review. *journal of Industrial and Engineering Chemistry*. 2021, *101*, 307–323. DOI: 10.1080/00222348.2013.860304.

[9] Raina, N.; Pahwa, R.; Khosla, J. K.; Gupta, P. N.; Gupta, M. Polycaprolactone-Based Materials in Wound Healing Applications. *Polym. Bull.* 2021, online. DOI: 10.1007/s00289-021-03865-w.

[10] Mahmoud Salehi, A. O.; Heidari Keshel, S.; Sefat, F.; Tayebi, L. Use of Polycaprolactone in Corneal Tissue Engineering: A Review. *Mater. Today Commun.* 2021, *27*(May), 102402. DOI: 10.1016/j.mtcomm.2021.102402.

[11] Ilyas, R. A.; Zuhri, M. Y. M.; Nor, M.; Misenan, M. S. M.; Jenol, M.; Samsudin, S. A.; Nurazzi, N. M.; Asyraf, M. R. M.; Supian, A. B. M.; Bangar, S. P., et al. Natural Fiber-Reinforced Polycaprolactone Green and Hybrid Biocomposites for Various Advanced Applications. *Polymers (Basel).* 2022, *14*(1), 182–211. DOI: 10.3390/polym14010182.

[12] Mondal, D. Griffith, M.; Venkatraman, S. S. Polycaprolactone-Based Biomaterials for Tissue Engineering and Drug Delivery: Current Scenario and Challenges. *Int. J. Polym. Mater. Polym. Biomater.* 2016, *65*(5), 255–265. DOI: 10.1080/00914037.2015.1103241.

[13] Suwantong, O. Biomedical Applications of Electrospun Polycaprolactone Fiber Mats. *Polym. Adv. Technol.* 2016, *27*(10), 1264–1273. DOI: 10.1002/pat.3876.

[14] Hajiali, F.; Tajbakhsh, S.; Shojaei, A. Fabrication and Properties of Polycaprolactone Composites Containing Calcium Phosphate-Based Ceramics and Bioactive Glasses in Bone Tissue Engineering: A Review. *Polym. Rev.* 2018, *58*(1), 164–207. DOI: 10.1080/15583724.2017.1332640.

[15] Prasad, A.; Kandasubramanian, B. Fused Deposition Processing Polycaprolactone of Composites for Biomedical Applications. *Polym. Technol. Mater.* 2019, *58*(13), 1365–1398. DOI: 10.1080/25740881.2018.1563117.

[16] Abrisham, M.; Noroozi, M.; Panahi-Sarmad, M.; Arjmand, M.; Goodarzi, V.; Shakeri, Y.; Golbaten-Mofrad, H.; Dehghan, P.; Seyfi Sahzabi, A.; Sadri, M.,

et al. The Role of Polycaprolactone-Triol (PCL-T) in Biomedical Applications: A State-of-the-Art Review. *Eur. Polym. J.* 2020, *131*, 109701. DOI: 10.1016/j.eurpolymj.2020.109701.

[17] Dodero, A.; Alloisio, M.; Castellano, M.; Vicini, S. Multilayer Alginate-Polycaprolactone Electrospun Membranes as Skin Wound Patches with Drug Delivery Abilities. *ACS Appl. Mater. Interfaces*. 2020, *12*(28), 31162–31171. DOI: 10.1021/acsami.0c07352.

[18] Bartnikowski, M.; Dargaville, T. R.; Ivanovski, S.; Hutmacher, D. W. Degradation Mechanisms of Polycaprolactone in the Context of Chemistry, Geometry and Environment. *Prog. Polym. Sci.* 2019, *96*, 1–20. DOI: 10.1016/j.progpolymsci.2019.05.004.

[19] Tran, J.; Pesenti, T.; Cressonnier, J.; Lefay, C.; Gigmes, D.; Guillaneuf, Y.; Nicolas, J. Degradable Copolymer Nanoparticles from Radical Ring-Opening Copolymerization between Cyclic Ketene Acetals and Vinyl Ethers. *Biomacromolecules*. 2019, *20*(1), 305–317. DOI: 10.1021/acs.biomac.8b01500.

[20] Liu, C.; Qin, H.; Mather, P. T. Review of Progress in Shape-Memory Polymers. *J. Mater. Chem. Chem.* 2007, *17*16, 1543–1558 10.1039/b615954k November 2006

[21] Han, W.; Liao, X.; Yang, Q.; Li, G.; He, B.; Zhu, W.; Hao, Z. Crystallization and Morphological Transition of Poly (l-Lactide)-Poly(ϵ-Caprolactone) Diblock Copolymers with Different Block Length Ratios. *RSC Adv.* 2017, *7* (36), 22515–22523. DOI: 10.1039/c7ra03496b.

[22] Dakshinamoorthy, D.; Peruch, F. Block and Random Copolymerization of μ-Caprolactone, L-, and Rac-Lactide Using Titanium Complex Derived from Aminodiol Ligand. *J. Polym. Sci. Part A Polym. Chem.* 2012, *50*(11), 2161–2171. DOI: 10.1002/pola.25983.

[23] Schoener, C. A.; Weyand, C. B.; Murthy, R.; Grunlan, M. A. Shape Memory Polymers with Silicon-Containing Segments. *J. Mater. Chem.* 2010, *20* (9), 1787–1793. DOI: 10.1039/b924032b.

[24] Yang, J.; Zhou, Q.; Shen, K.; Song, N.; Ni, L. Controlling Nanodomain Morphology of Epoxy Thermosets Templated by Poly(Caprolactone)-: Block -poly (dimethylsiloxane)- Block -poly(caprolactone) ABA Triblock Copolymer. *RSC Adv.* 2018, *8*(7), 3705–3715. DOI: 10.1039/c7ra12826f.

[25] Qindeel, M.; Ahmed, N.; Shah, K. U.; Ullah, N.; Asim. ur, R. New, Environment Friendly Approach for Synthesis of Amphiphilic PCL–PEG–PCL Triblock Copolymer: An Efficient Carrier for Fabrication of Nanomicelles. *J. Polym. Environ.* 2020, *28*(4), 1237–1251. DOI: 10.1007/s10924-020-01683-1.

[26] Rani, S.; Gupta, U. S. Synthesis, Morphology, and Rheological Evaluation of HPMA (N −2-hydroxypropyl Methacrylamide)-PCL (Polycaprolactone) Conjugates. *ACS Omega*. 2021, *6*(44), 29788–29803. DOI: 10.1021/acsomega.1c04243.

[27] Li, J.; Xu, H.; Hu, N.; Shi, D.; Dong, W.; Wu, C.; Chen, M. Studies on Photoreactive and Biodegradable Copolymers Composed of Poly(-Caprolactone) and 4-Hydroxycinnamic Acid. *Polym. J.* 2012, *44*(11), 1123–1130. DOI: 10.1038/pj.2012.85.

[28] Liu, M. J.; Chen, S. C.; Yang, K. K.; Wang, Y. Z. Biodegradable Polylactide Based Materials with Improved Crystallinity, Mechanical Properties and Rheological Behaviour by Introducing a Long-Chain Branched Copolymer. *RSC Adv.* 2015, *5*(52), 42162–42173. DOI: 10.1039/c5ra04742k.

[29] Jiao, M.; Yang, K.; Cao, J.; Liu, H.; Pan, W.; Gao, P. Designing and Characterization of Poly(L-Lactide)/Poly (ε-Caprolactone) Multiblock Copolymers. *J. Macromol. Sci. Part B Phys.* 2014, *53*(2), 191–204. DOI: 10.1080/00222348.2013.810058.

[30] Polo Fonseca, L.; Bergamo Trinca, R.; Isabel Felisberti, M. Thermo-Responsive Polyurethane Hydrogels Based on Poly(Ethylene Glycol) and Poly(Caprolactone): Physico-Chemical and Mechanical Properties. *J. Appl. Polym. Sci.* 2016, *133* (25), 1–10. DOI: 10.1002/app.43573.

[31] Fuoco, T.; Finne-Wistrand, A. Enhancing the Properties of Poly(ϵ-Caprolactone) by Simple and Effective Random Copolymerization of ϵ-caprolactone with p-Dioxanone. *Biomacromolecules*. 2019, *20*(8), 3171–3180. DOI: 10.1021/acs.biomac.9b00745.

[32] Yang, L. Q.; Meng, S.; Liu, D. H.; Li, M.; Yang, D.; Guan, Y. M.; Li, J. X. Potential Biodegradable Implants from ε-Caprolactone and D, L-Lactide Copolymers: Synthesis, Properties, and in Vivo Degradation. *Int. J. Polym. Anal. Charact.* 2014, *19*(5), 422–440. DOI: 10.1080/1023666X.2014.920069.

[33] Saatchi, M.; Behl, M.; Nöchel, U.; Lendlein, A. Copolymer Networks from Oligo(ε -caprolactone) and N -butyl Acrylate Enable a Reversible Bidirectional Shape-Memory Effect at Human Body Temperature. *Macromol. Rapid Commun.* 2015, *36*(10), 880–884. DOI: 10.1002/marc.201400729.

[34] Bellani, C. F.; Pollet, E.; Hebraud, A.; Pereira, F. V.; Schlatter, G.; Avérous, L.; Bretas, R. E. S.; Branciforti, M. C. M. Thermal, and Mechanical Properties of Poly(ε-Caprolactone)/Poly(ε-Caprolactone)-Grafted-Cellulose Nanocrystals Mats Produced by Electrospinning. *J. Appl. Polym. Sci.* 2016, *133*(21), 4–11. DOI: 10.1002/app.43445.

[35] Ninago, M. D.; De Freitas, A. G. O.; Hanazumi, V.; Muraro, P. I. R.; Schmidt, V.; Giacomelli, C.; Ciolino, A. E.; Villar, M. A. Synthesis of Grafted Block Copolymers Based on ε-Caprolactone: Influence of Branches on Their Thermal Behavior. *Macromol. Chem. Phys.* 2015, *216*(24), 2331–2343. DOI: 10.1002/macp.201500248.

[36] Moura, I.; Nogueira, R.; Bounor-Legare, V.; Machado, A. V. Effect of PCL and EVA Molar Mass on the Development of Sustainable Polymers. *Soft Mater.* 2014, *12*(1), 88–97. DOI: 10.1080/1539445X.2012.756818.

[37] Hedir, G. G.; Bell, C. A.; O'Reilly, R. K.; Dove, A. P. Functional Degradable Polymers by Radical Ring-Opening Copolymerization of MDO and Vinyl Bromobutanoate: Synthesis, Degradability and Post-Polymerization Modification. *Biomacromolecules*. 2015, *16*(7), 2049–2058. DOI: 10.1021/acs.biomac.5b00476.

[38] Dai, G.; Xie, Q.; Chen, S.; Ma, C.; Zhang, G. Biodegradable Poly(Ester)-Poly(Methyl Methacrylate) Copolymer for Marine Anti-Biofouling. *Prog. Org. Coatings*. 2018, *124*(August), 55–60. DOI: 10.1016/j.porgcoat.2018.08.003.

[39] Komatsu, S.; Asoh, T. A.; Ishihara, R.; Kikuchi, A. Facile Preparation of Degradable Thermoresponsive Polymers as Biomaterials: Thermoresponsive Polymers Prepared by Radical Polymerization Degrade to Water-Soluble Oligomers. *Polymer (Guildf.)*. 2017, *130*, 68–73. DOI: 10.1016/j.polymer.2017.09.073.

[40] Shi, Y.; Schmalz, H.; Agarwal, S. Designed Enzymatically Degradable Amphiphilic Conetworks by Radical Ring-Opening Polymerization. *Polym. Chem.* 2015, *6*(35), 6409–6415. DOI: 10.1039/c5py00962f.

[41] Undin, J.; Finne-Wistrand, A.; Albertsson, A. C. Adjustable Degradation Properties and Biocompatibility of Amorphous and Functional Poly(Ester-Acrylate)-Based Materials. *Biomacromolecules*. 2014, *15*(7), 2800–2807. DOI: 10.1021/bm500689g.

[42] Undin, J.; Finne-Wistrand, A.; Albertsson, A. C. Copolymerization of 2-Methylene-1,3-Dioxepane and Glycidyl Methacrylate, a Well-Defined and Efficient Process for Achieving Functionalized Polyesters for Covalent Binding of Bioactive Molecules. *Biomacromolecules*. 2013, *14*(6), 2095–2102. DOI: 10.1021/bm4004783.

[43] Bai, H.; Xiu, H.; Gao, J.; Deng, H.; Zhang, Q.; Yang, M.; Fu, Q. Tailoring Impact Toughness of Poly(L-Lactide)/Poly(ε-Caprolactone) (PLLA/PCL) Blends by Controlling Crystallization of PLLA Matrix. *ACS Appl. Mater. Interfaces*. 2012, *4*(2), 897–905. DOI: 10.1021/am201564f.

[44] Hu, M.; Deng, C.; Gu, X.; Fu, Q.; Zhang, J. Manipulating the Strength-Toughness Balance of Poly(l-Lactide) (PLLA) via Introducing Ductile Poly(ϵ-Caprolactone) (PCL) and Strong Shear Flow. *Ind. Eng. Chem. Res.* 2020, *59*(2), 1000–1009. DOI: 10.1021/acs.iecr.9b05380.

[45] Lv, Q.; Wu, D.; Xie, H.; Peng, S.; Chen, Y.; Xu, C. Crystallization of Poly(ϵ-Caprolactone) in Its Immiscible Blend with Polylactide: Insight into the Role of Annealing Histories. *RSC Adv.* 2016, *6*(44), 37721–37730. DOI: 10.1039/c6ra07752h.

[46] Ostafinska, A.; Fortelny, I.; Nevoralova, M.; Hodan, J.; Kredatusova, J.; Slouf, M. Synergistic Effects in Mechanical Properties of PLA/PCL Blends with Optimized Composition, Processing, and Morphology. *RSC Adv.* 2015, *5*(120), 98971–98982. DOI: 10.1039/c5ra21178f.

[47] Wachirahuttapong, S.; Thongpin, C.; Sombatsompop, N. Effect of PCL and Compatibility Contents on the Morphology, Crystallization and Mechanical Properties of PLA/PCL Blends. *Energy Procedia*. 2016, *89*, 198–206. DOI: 10.1016/j.egypro.2016.05.026.

[48] Chang, R.; Rohindra, D.; Lata, R.; Kuboyama, K.; Ougizawa, T. Development of Poly(ε-Caprolactone) /Pine Resin Blends: Study of Thermal, Mechanical, and Antimicrobial Properties. *Polym. Eng. Sci*. 2019, *59*(s2), E32–E41. DOI: 10.1002/pen.24950.

[49] Bou-Francis, A.; Piercey, M.; Al-Qatami, O.; Mazzanti, G.; Khattab, R.; Ghanem, A. Polycaprolactone Blends for Fracture Fixation in Low Load-Bearing Applications. *J. Appl. Polym. Sci.* 2020, *137*(32), 48940. DOI: 10.1002/app.48940.

[50] Wang, F.; Yang, X.; Zou, Y. Effect of the Maleation of Lignosulfonate on the Mechanical and Thermal Properties of Lignosulfonate/Poly(ε-Caprolactone) Blends. *J. Appl. Polym. Sci.* 2016, *133*(5), 1–7. DOI: 10.1002/app.42925.

[51] Zeng, A.; Wang, Y.; Li, D.; Guo, J.; Chen, Q. Preparation and Antibacterial Properties of Polycaprolactone/Quaternized Chitosan Blends. *Chin. J. Chem. Eng.* 2021, *32*, 462–471. DOI: 10.1016/j.cjche.2020.10.001.

[52] Vergara-Porras, B.; Gracida-Rodríguez, J. N.; Pérez-Guevara, F. Thermal Processing Influence on Mechanical, Thermal, and Biodegradation Behavior in Poly(β-Hydroxybutyrate)/Poly(ϵ-Caprolactone) Blends: A Descriptive Model. *J. Appl. Polym. Sci.* 2016, *133*(27), 1–12. DOI: 10.1002/app.43569.

[53] Przybysz, M.; Marć, M.; Klein, M.; Saeb, M. R.; Formela, K. S. Mechanical and Thermal Behavior Assessments of PCL/PHB Blends Reactively Compatibilized with Organic Peroxides. *Polym. Test.* 2018, *67*, 513–521. DOI: 10.1016/j.polymertesting.2018.03.014.

[54] Woodard, L. N.; Grunlan, M. A. Hydrolytic Degradation of PCL-PLLA Semi-IPNs Exhibiting Rapid, Tunable Degradation. *ACS Biomater. Sci. Eng.* 2019, *5*(2), 498–508. DOI: 10.1021/acsbiomaterials.8b01135.

[55] Díaz, E.; Puerto, I.; Sandonis, I.; Ibañez, I. Morphology and Mechanical Properties of PLLA and PCL Scaffolds. *Polym. - Plast. Technol. Eng.* 2014, *53*(2), 150–155. DOI: 10.1080/03602559.2013.843699.

[56] Noroozi, N.; Schafer, L. L.; Savvas, G. H. Thermorheological Properties of Poly (e-Caprolactone)/ Polylactide Blends. *Polym. Eng. Sci.* 2012, *52*(11), 2348–2359. DOI: 10.1002/pen.

[57] Sharma, D.; Satapathy, B. K. Performance Evaluation of Electrospun Nanofibrous Mats of Polylactic Acid (Pla)/poly (e-Caprolactone) (PCL) Blends. *Mater. Today Proc.* 2019, *19*(xxxx), 188–195. DOI: 10.1016/j.matpr.2019.06.698.

[58] Keridou, I.; Franco, L.; Turon, P.; Del Valle, L. J.; Puiggalí, J. Scaffolds with Tunable Properties Constituted by Electrospun Nanofibers of Polyglycolide and Poly(ε-Caprolactone). *Macromol. Mater. Eng.* 2018, *303*(7), 1–14. DOI: 10.1002/mame.201800100.

[59] Katsumata, K.; Saito, T.; Yu, F.; Nakamura, N.; Inoue, Y. The Toughening Effect of a Small Amount of Poly(ε-Caprolactone) on the Mechanical Properties of the Poly (3-Hydroxybutyrate-Co-3-Hydroxyhexanoate)/PCL Blend. *Polym. J.* 2011, *43*(5), 484–492. DOI: 10.1038/pj.2011.12.

[60] Rohindra, D.; Lata, R.; Kuboyama, K.; Ougizawa, T. Crystallization Behavior in Miscible Blends of Poly(ε-caprolactone) and Poly(Hexylene Adipate) with Similar Thermal Properties Studied by Time-resolved Fourier Transform Infrared Spectroscopy. *Polym. Cryst.* 2019, *2*(1), 1–8. DOI: 10.1002/pcr2.10037.

[61] Abdelrazek, E. M.; Hezma, A. M.; El-khodary, A.; Elzayat, A. M. Spectroscopic Studies and Thermal Properties of PCL/PMMA Biopolymer Blend. *Egypt. J. Basic Appl. Sci.* 2016, *3*(1), 10–15. DOI: 10.1016/j.ejbas.2015.06.001.

[62] Sousa, J. C.; Costa, A. R. M.; Lima, J. C.; Arruda, S. A.; Almeida, Y. M. B.; Canedo, E. L. Polycaprolactone (Pcl)/alumina and PCL/Niobium Pentoxide Composites: Rheology, Crystallization, and Mechanical

Properties. *Polym. Compos.* 2020, *41*(4), 1265–1276. DOI: 10.1002/pc.25452.
[63] Pötschke, P.; Villmow, T.; Krause, B. Melt Mixed PCL/MWCNT Composites Prepared at Different Rotation Speeds: Characterization of Rheological, Thermal, and Electrical Properties, Molecular Weight, MWCNT Macrodispersion, and MWCNT Length Distribution. *Polymer (Guildf.)*. 2013, *54*(12), 3071–3078. DOI: 10.1016/j.polymer.2013.04.012.
[64] Dhakal, H.; Bourmaud, A.; Berzin, F.; Almansour, F.; Zhang, Z.; Shah, D. U.; Beaugrand, J. Mechanical Properties of Leaf Sheath Date Palm Fibre Waste Biomass Reinforced Polycaprolactone (PCL) Biocomposites. *Ind. Crop Prod.* 2018, *126*(October), 394–402. DOI: 10.1016/j.indcrop.2018.10.044.
[65] Leung, L. H.; Naguib, H. E. Viscoelastic Properties of Poly(ε-Caprolactone) - Hydroxyapatite Micro- and Nano-Composites. *Polym. Adv. Technol.* 2013, *24*(2), 144–150. DOI: 10.1002/pat.3061.
[66] Mdletshe, T. S.; Mishra, S. B.; Mishra, A. K. Studies on the Effect of Silicon Carbide Nanoparticles on the Thermal, Mechanical, and Biodegradation Properties of Poly(Caprolactone). *J. Appl. Polym. Sci.* 2015, *132*(26), 1–9. DOI: 10.1002/app.42145.
[67] Wu, Q.; Ma, N.; Liu, T.; Koranteng, E. Properties of Compatible Soy Protein Isolate/Polycaprolactone Composite with Special Interface Structure. *Polym. Compos.* 2019, *40*(S1), E383–E391. DOI: 10.1002/pc.24694.
[68] Lee, K. S.; Chang, Y. W. T. Mechanical, and Rheological Properties of Poly(ε- Caprolactone)/Halloysite Nanotube Nanocomposites. *J. Appl. Polym. Sci.* 2013, *128*(5), 2807–2816. DOI: 10.1002/app.38457.
[69] Reul, L. T. A.; Pereira, C. A. B.; Sousa, F. M.; Santos, R. M.; Carvalho, L. H.; Canedo, E. L. Polycaprolactone/Babassu Compounds: Rheological, Thermal, and Morphological Characteristics. *Polym. Compos.* 2019, *40*(S1), E540–E549. DOI: 10.1002/pc.24861.
[70] Wu, C. S.; Liao, H. T. Polycaprolactone-Based Green Renewable Ecocomposites Made from Rice Straw Fiber: Characterization and Assessment of Mechanical and Thermal Properties. *Ind. Eng. Chem. Res.* 2012, *51*(8), 3329–3337. DOI: 10.1021/ie202002p.
[71] Weng, F.; Zhang, P.; Koranteng, E.; Zhang, Y.; Wu, Q.; Zeng, G. Effects of Shell Powder Size and Content on the Properties of Polycaprolactone Composites. *J. Appl. Polym. Sci.* 2021, *138*(43), 1–11. DOI: 10.1002/app.51264.
[72] Liu, J.; Roque, R.; Barbosa, G. F.; Malavolta, A. T. Compression Stiffness Evaluation of Polycaprolactone-Amorphous Calcium Phosphate 3D-Designed Scaffolds Oriented by Finite Element Analysis. *J. Appl. Polym. Sci.* 2021, *138*(42), 1–9. DOI: 10.1002/app.51245.
[73] Ludueña, L. N.; Vázquez, A.; Alvarez, V. A. Effect of the Type of Clay Organo-Modifier on the Morphology, Thermal/Mechanical/Impact/Barrier Properties and Biodegradation in Soil of Polycaprolactone/Clay Nanocomposites. *J. Appl. Polym. Sci.* 2013, *128*(5), 2648–2657. DOI: 10.1002/app.38425.
[74] Merino, D.; Alvarez, V. A. Thermal Degradation of Poly (ε-Caprolactone) Nanocomposites with Soy Lecithin-Modified Bentonite Fillers. *Thermochim. Acta*. 2020, *689*, 178638. DOI: 10.1016/j.addma.2020.101070.
[75] Rešček, A.; Ščetar, M.; Hrnjak-Murgić, Z.; Dimitrov, N.; Galić, K. Polyethylene/Polycaprolactone Nanocomposite Films for Food Packaging Modified with Magnetite and Casein: Oxygen Barrier, Mechanical, and Thermal Properties. *Polym. - Plast. Technol. Eng.* 2016, *55*(14), 1450–1459. DOI: 10.1080/03602559.2016.1163606.
[76] Jafari, H.; Shahrousvand, M.; Kaffashi, B. Preparation and Characterization of Reinforced Poly (ε-Caprolactone) Nanocomposites by Cellulose Nanowhiskers. *Polym. Compos.* 2020, *41*(2), 624–632. DOI: 10.1002/pc.25393.
[77] Saravanamoorthy, S.; Chandra Bose, A.; Velmathi, S. Facile Fabrication of Polycaprolactone/h-MoO3 Nanocomposites and Their Structural, Optical and Electrical Properties. *RSC Adv.* 2015, *5*(120), 99074–99083. DOI: 10.1039/c5ra17733b.
[78] Bicy, K.; Geethamma, V. G.; Kalarikkal, N.; Rouxel, D.; Thomas, S. Poly(ε-Caprolactone)/Functionalized-Carbon Nanotube Electrospun Nanocomposites: Crystallization and Thermal Properties. *Macromol. Symp.* 2018, *381*(1), 1–6. DOI: 10.1002/masy.201800140.
[79] Xie, M. M.; Wang, B. B.; Zhang, P. The Effect of Crystallization Behavior on High Conductivity, Enhanced Mechanism and Thermal Stability of Poly(ε-Caprolactone)/Multi-Walled Carbon Nanotube Composites. *J. Dispers. Sci. Technol.* 2019, *40*(1), 94–102. DOI: 10.1080/01932691.2018.1464470.
[80] Singh, J.; Pandey, P. M.; Kaur, T.; Singh, N. A Comparative Analysis of Solvent Cast 3D Printed Carbonyl Iron Powder Reinforced Polycaprolactone Polymeric Stents for Intravascular Applications. *J. Biomed. Mater. Res. Part B Appl. Biomater.* 2021, *109*(9), 1344–1359. DOI: 10.1002/jbm.b.34795.
[81] Fadaie, M.; Mirzaei, E.; Geramizadeh, B.; Asvar, Z. Incorporation of Nanofibrillated Chitosan into Electrospun PCL Nanofibers Makes Scaffolds with Enhanced Mechanical and Biological Properties. *Carbohydr. Polym.* 2018, *199*, 628–640. DOI: 10.1016/j.carbpol.2018.07.061.
[82] El-Naggar, M. E.; Shalaby, E. S.; Abd-Al-Aleem, A. H.; Abu-Saied, M. A.; Youssef, A. M. Synthesis of Environmentally Benign Antimicrobial Dressing Nanofibers Based on Polycaprolactone Blended with Gold Nanoparticles and Spearmint Oil Nanoemulsion. *J. Mater. Res. Technol.* 2021, *15*, 3447–3460. DOI: 10.1016/j.jmrt.2021.09.136.
[83] Victor, H, A.-C. S.; Go´ mez-Salazar Martin, R.; Soto, V.; Luna-Ba´rcenas, G.; Katime, I.; Sergio, M. N.-D. Comparative Study of the Thermal and Mechanical Properties of Nanocomposites Prepared by in Situ Polymerization of E-Caprolactone and Functionalized Carbon Nanotubes. *Polym. Compos.* 2012, *33*(4), 5563–5572. DOI: 10.1002/pc.
[84] Wang, Y.; Li, T.; Ma, P.; Zhang, S.; Du, M.; Dong, W.; Xie, Y.; Chen, M. Graphene-Assisted Fabrication of Poly(ϵ-Caprolactone)-Based Nanocomposites with High Mechanical Properties and Self-Healing

Functionality. *New J. Chem.* 2018, *42*(12), 10348–10356. DOI: 10.1039/c8nj01278d.

[85] Cheng, Z.; Xiuwen, W.; Guopu, C.; Feng, W.; Jianan, R. AuNPs-PCL Nanocomposite Accelerated Abdominal Wound Healing through Photothermal Effect and Improving Cell Adhesion. *J. Biomater. Sci. Polym. Educ.* 2018, *29*, 2035–2049.

[86] Inukai, S.; Kurokawa, N.; Hotta, A. Mechanical Properties of Poly(ε-Caprolactone) Composites with Electrospun Cellulose Nanofibers Surface Modified by 3-Aminopropyltriethoxysilane. *J. Appl. Polym. Sci.* 2020, *137*(17), 1–10. DOI: 10.1002/app.48599.

[87] Zaman, H. U.; Beg, M. D. H. Improvement of Physico-Mechanical, Thermomechanical, Thermal and Degradation Properties of PCL/Gelatin Biocomposites: Effect of Gamma Radiation. *Radiat. Phys. Chem.* 2015, *109*, 73–82. DOI: 10.1016/j.radphyschem.2014.12.011.

[88] Lozano-Sánchez, L. M.; Bagudanch, I.; Sustaita, A. O.; Iturbe-Ek, J.; Elizalde, L. E.; Garcia-Romeu, M. L.; Elías-Zúñiga, A. Single-Point Incremental Forming of Two Biocompatible Polymers: An Insight into Their Thermal and Structural Properties. *Polymers (Basel).* 2018, *10*, 4. DOI: 10.3390/polym10040391.

[89] Iroh, J. O.;. *Polymer Data Handbook*; ed. J. E. ACS publication, J Am Chem Soc. 1999.DOI: 10.1021/ja907879q

[90] Wurm, A.; Zhuravlev, E.; Eckstein, K.; Jehnichen, D.; Pospiech, D.; Androsch, R.; Wunderlich, B.; Schick, C. Crystallization and Homogeneous Nucleation Kinetics of Poly(ε- Caprolactone) (PCL) with Different Molar Masses. *Macromolecules*. 2012, *45*(9), 3816–3828. DOI: 10.1021/ma300363b.

[91] Jiang, Y.; Fang, L.; Kratz, K.; Lendlein, A. Crystallization Behavior of Copolyesterurethanes Containing Different Weight Contents of Crystallizable Poly(ε-Caprolactone) Segments. *Macromol. Symp*. 2014, *345*(1), 59–65. DOI: 10.1002/masy.201400138.

[92] Shen, H.; Quintard, G.; Chen, J.; Taha, M. Synthesis and Thermomechanical Properties of Allyl-Functionalized Polycaprolactone Urethane-Co-2-Hydroxylethyl Methacrylate Networks. *J. Appl. Polym. Sci.* 2015, *132* (3), 1–10. DOI: 10.1002/APP.41295.

[93] Shen, H.; Chen, J.; Taha, M. Cross-Linking and Damping Properties of Poly(Caprolactone-Co-Glycidyl Methacrylate). *Polym. J.* 2014, *46*(9), 598–608. DOI: 10.1038/pj.2014.29.

[94] Abdolmohammadi, S.; Siyamak, S.; Ibrahim, N. A.; Wan Yunus, W. M. Z.; Ab Rahman, M. Z.; Azizi, S.; Fatehi, A. Enhancement of Mechanical and Thermal Properties of Polycaprolactone/Chitosan Blend by Calcium Carbonate Nanoparticles. *Int. J. Mol. Sci.* 2012, *13*(4), 4508–4522. DOI: 10.3390/ijms13044508.

[95] Bing, Z.; Zai-Zai, T.; Jie, H.; Jun-Ting, X.; Zhi-Qiang, F. Isothermal Crystallization Kinetics of Multi-Walled Carbon Nanotubes-Graft-Poly(e-Caprolactone) with High Grafting Degrees. *CrystEngComm*. 2013, *15*(38), 7824–7832. DOI: 10.1039/c3ce40606g.

[96] Liu, L.; Zhang, Y.; Li, C.; Cao, J.; He, E.; Wu, X.; Wang, F.; Wang, L. Facile Preparation PCL/ Modified Nano ZnO Organic-Inorganic Composite and Its Application in Antibacterial Materials. *J. Polym. Res.* 2020, *27*(3), 78. DOI: 10.1007/s10965-020-02046-z.

[97] Kumar, D.; Babu, G.; Krishnan, S. Study on Mechanical & Thermal Properties of PCL Blended Graphene Biocomposites. *Polimeros*. 2019, *29*(2), 1–9. DOI: 10.1590/0104-1428.05318.

[98] Seyrek, M. E.; Okur, M.; Saraçoğlu, N. Improvement of Mechanical, Thermal and Antimicrobial Properties of Organically Modified Montmorillonite Loaded Polycaprolactone for Food Packaging. *J. Vinyl Addit. Technol.* 2021, *27*(4), 894–908. DOI: 10.1002/vnl.21860.

[99] Schäfer, H.; Reul, L. T. A.; Souza, F. M.; Wellen, R. M. R.; Carvalho, L. H.; Koschek, K.; Canedo, E. L. Crystallization Behavior of Polycaprolactone/Babassu Compounds. *J. Therm. Anal. Calorim.* 2021, *143*(4), 2963–2972. DOI: 10.1007/s10973-020-09433-0.

[100] Rizzuto, M.; Mugica, A.; Zubitur, M.; Caretti, D.; Müller, A. J. Plasticization and Anti-Plasticization Effects Caused by Poly(Lactide-Ran-Caprolactone) Addition to Double Crystalline Poly(l-Lactide)/Poly(ε-Caprolactone) Blends. *CrystEngComm*. 2016, *18*(11), 2014–2023. DOI: 10.1039/c5ce02559a.

[101] Lanfranconi, M.; Alvarez, V. A.; Ludueña, L. N. Isothermal Crystallization of Polycaprolactone/Modified Clay Biodegradable Nanocomposites. *J. Therm. Anal. Calorim.* 2016, *126*(3), 1273–1280. DOI: 10.1007/s10973-016-5734-x.

[102] Wu, C.-M.; Huang, C.-W. Melting and Crystallization Behavior of Copolymer from Cyclic Butylene Terephthalate and Polycaprolactone. *Polym. Eng. Sci.* 2011, *51*(5), 1004–1013. DOI: 10.1002/pen.

[103] Palacios, J. K.; Mugica, A.; Zubitur, M.; Iturrospe, A.; Arbe, A.; Liu, G.; Wang, D.; Zhao, J.; Hadjichristidis, N.; Müller, A. J. Sequential Crystallization and Morphology of Triple Crystalline Biodegradable PEO-b-PCL-b-PLLA Triblock Terpolymers. *RSC Adv.* 2016, *6*(6), 4739–4750. DOI: 10.1039/c5ra25812j.

[104] Cesur, S.; Alp, B.; Küçükgöksel, Y.; Kahraman, T.; Balköse, D. Crystallization Kinetics and Affecting Parameters on Polycaprolactone Composites with Inorganic and Organic Additives. *J. Vinyl Addit. Technol.* 2015, *21*(3), 174–182. DOI: 10.1002/vnl.21399.

[105] Alp, B.; Cesur, S. Isothermal Crystallization Kinetics and Mechanical Properties of Polycaprolactone Composites with Zinc Oxide, Oleic Acid, and Glycerol Monooleate. *J. Appl. Polym. Sci.* 2013, *130*(2), 1259–1275. DOI: 10.1002/app.39217.

[106] Mahalakshmi, S.; Parthasarathy, V.; Tung, K. L.; Anbarasan, R.; Alagesan, T. Non-Isothermal Crystallization and Degradation Kinetics of Fe 3 O 4 –thymolblue Functionalized Poly(ε-Caprolactone). *J. Polym. Environ.* 2019, *0*, 0. DOI: 10.1007/s10924-019-01401-6.

[107] Kratochvíl, J.; Rotrekl, J.; Kaprálková, L.; Hromádková, J.; Kelnar, I. Epoxy/Poly(ε-Caprolactone) Nanocomposites: Effect of Transformations of Structure on Crystallization. *J. Appl. Polym. Sci.* 2013, *130*(5), 3197–3204. DOI: 10.1002/app.39536.

[108] Liang, J. Z.; Zhou, L.; Tang, C. Y.; Tsui, C. P. Crystallization Properties of Polycaprolactone Composites Filled with Nanometer Calcium

Carbonate. *J. Appl. Polym. Sci.* 2013, *128*(5), 2940–2944. DOI: 10.1002/app.38359.

[109] Yang, B.; Zhang, X.; Wang, C.; Liu, R.; Fan, B.; Zhang, H.; Sun, H. Effect of Polyvinyl Acetals on Non-Isothermal Crystallization Behaviour and Mechanical Properties of Poly(ε-Caprolactone). *RSC Adv.* 2019, *9*(63), 36815–36824. DOI: 10.1039/c9ra08133j.

[110] Harmansyah, F.; Woo, E. M.; Lee, L. T.; Chien, H. R. Distorted Ring-Banded Spherulites in Poly(l-Lactic Acid)/Poly(ε-Caprolactone) Blends. *RSC Adv.* 2014, *4*(90), 49006–49015. DOI: 10.1039/c4ra08658a.

[111] Ninago, M. D.; Redondo, F. L.; De Freites, A. G. O.; Giacomelli, C.; Ciolino, E.; Villar, M. A. Effects of Branches on the Isothermal Crystallization of Copolymers Based on Poly (E -caprolactone). *Macromol. Symp.* 2019, *383*(1), 1700082. DOI: 10.1002/masy.201700082.

[112] Zhang, D.; Giese, M. L.; Prukop, S. L.; Grunlan, M. A. Poly (ε-Caprolactone) -based Shape Memory Polymers with Variable Polydimethylsiloxane Soft Segment Lengths. *J. Polym. Sci. Part A Polym. Chem.* 2011, *49*(3), 754–761. DOI: 10.1002/pola.24488.

[113] Bai, Y.; Zhang, X.; Wang, Q.; Wang, T. Shape Memory Property of Microcrystalline Cellulose – Poly (e-Caprolactone) Polymer Network with Broad Transition Temperature. *J. Mater. Sci.* 2014, *49*(5), 2252–2262. DOI: 10.1007/s10853-013-7920-6.

[114] Raana, S.; Mahyar, -A.-A.; Massoumi, B.; Younes, B.-K.; Samira, A. Conductive and Biodegradable Scaffolds Based on a five-arm and Functionalized star-like poly-aniline–polycaprolactone Copolymer with a D -glucose Core. *New J. Chem.* 2017, *41*(14), 6371–6384. DOI: 10.1039/C7NJ01063J.

[115] Yin, G.; Chen, G.; Zhou, Z.; Li, Q. Modification of PEG-b-PCL Block Copolymer with High Melting Temperature by the Enhancement of POSS Crystal and Ordered Phase Structure. *RSC Adv.* 2015, *5*(42), 33356–33363. DOI: 10.1039/c5ra01971k.

[116] Luo, S.; Peng, X.; Chen, Y.; Su, T.; Cao, J.; Li, S.; He, B. S. Characterization, and Crystallization of Biodegradable Poly(ε-Caprolactone)-Poly(L-Lactide) Diblock Copolymers. *E-Polymers.* 2015, *15*(1), 15–23. DOI: 10.1515/epoly-2014-0155.

[117] Yilgör, E.; Isik, M.; Söz, C. K.; Yilgör, I. Synthesis and Structure-Property Behavior of Polycaprolactone-Polydimethylsiloxane-Polycaprolactone Triblock Copolymers. *Polymer (Guildf.).* 2016, *83*, 138–153. DOI: 10.1016/j.polymer.2015.12.024.

[118] Jing, X.; Mi, H. Y.; Huang, H. X.; Turng, L. S. Shape Memory Thermoplastic Polyurethane (TPU)/Poly(ε-Caprolactone) (PCL) Blends as Self-Knotting Sutures. *J. Mech. Behav. Biomed. Mater.* 2016, *64*, 94–103. DOI: 10.1016/j.jmbbm.2016.07.023.

[119] Kotula, A. P.; Migler, K. B. Evaluating Models for Polycaprolactone Crystallization via Simultaneous Rheology and Raman Spectroscopy. *J. Rheol. (N. Y. N. Y).* 2018, *62*(1), 343–356. DOI: 10.1122/1.5008381.

[120] Chae, D. W.; Nam, Y.; An, S. G.; Cho, C. G.; Lee, E. J.; Kim, B. C. Effects of Molecular Architecture on the Rheological and Physical Properties of Polycaprolactone. *Korea Aust. Rheol. J.* 2017, *29*(2), 129–135. DOI: 10.1007/s13367-017-0014-2.

[121] Kelly, C. A.; Murphy, S. H.; Leeke, G. A.; Howdle, S. M.; Shakesheff, K. M.; Jenkins, M. J. Rheological Studies of Polycaprolactone in Supercritical CO2. *Eur. Polym. J.* 2013, *49*(2), 464–470. DOI: 10.1016/j.eurpolymj.2012.11.021.

[122] Sethuraman, V.; Makornkaewkeyoon, K.; Khalf, A.; Madihally, S. V. Influence of Scaffold Forming Techniques on Stress Relaxation Behavior of Polycaprolactone Scaffolds. *J. Appl. Polym. Sci.* 2013, *130*(6), 4237–4244. DOI: 10.1002/app.39599.

[123] Bouakaz, B. S.; Habi, A.; Grohens, Y.; Pillin, I. Effect of Combinations of Nanofillers on Rheology-Structure Relations in Biodegradable Poly(ε-Caprolactone) Nanocomposites. *Appl. Clay Sci.* 2018, *161*(April), 35–47. DOI: 10.1016/j.clay.2018.04.006.

[124] Nie, Y.; Zhan, H.; Kou, L.; Gu, Y. Atomistic Insights on the Rheological Property of Polycaprolactone Composites with the Addition of Graphene. *Adv. Mater. Technol.* 2021, *2100507*, 1–9. DOI: 10.1002/admt.202100507.

[125] Wang, Y.; Xu, C.; Wu, D.; Xie, W.; Wang, K.; Xia, Q.; Yang, H. Rheology of the Cellulose Nanocrystals Filled Poly(ε-Caprolactone) Biocomposites. *Polymer (Guildf.).* 2018, *140*, 167–178. DOI: 10.1016/j.polymer.2018.02.050.

[126] Tian, G.; Zhu, G.; Ren, T.; Liu, Y.; Wei, K.; Liu, Y. X. The Effects of PCL Diol Molecular Weight on Properties of Shape Memory Poly(ε-Caprolactone) Networks. *J. Appl. Polym. Sci.* 2019, *136*(6), 1–8. DOI: 10.1002/app.47055.

[127] Liang, J. Z.; Duan, D. R.; Tang, C. Y.; Tsui, C. P.; Chen, D. Z. Flexural Properties of Poly-L-Lactide and Polycaprolactone Shape Memory Composites Filled with Nanometer Calcium Carbonate. *J. Macromol. Sci. Part B Phys.* 2013, *52*(7), 964–972. DOI: 10.1080/00222348.2012.746572.

[128] Jahangiri, M.; Kalajahi, A. E.; Rezaei, M.; Bagheri, M. Shape Memory Hydroxypropyl Cellulose-g-Poly (ε-Caprolactone) Networks with Controlled Drug Release Capabilities. *J. Polym. Res.* 2019, *26*(6), 6. DOI: 10.1007/s10965-019-1798-1.

[129] Bai, Y.; Jiang, C.; Wang, Q.; Wang, T. Multi-Shape-Memory Property Study of Novel Poly (ε-Caprolactone)/ Ethyl Cellulose Polymer Networks. *Macromol. Chem. Phys.* 2013, *214*(21), 2465–2472. DOI: 10.1002/macp.201300389.

[130] Zhou, Y.; Zhou, D.; Cao, P.; Zhang, X.; Wang, Q.; Wang, T.; Li, Z.; He, W.; Ju, J.; Zhang, Y. 4D Printing of Shape Memory Vascular Stent Based on BCD-g-Polycaprolactone. *Macromol. Rapid Commun.* 2021, *42*(14), 1–9. DOI: 10.1002/marc.202100176.

[131] Iregui, Á.; Otaegi, I.; Arandia, I.; Martin, M. D.; Müller, A. J.; Irusta, L.; González, A. Fully Reversible Spherulitic Morphology in Cationically Photopolymerized DGEBA/PCL Shape-Memory Blends. *Macromolecules.* 2020, *53*(4), 1368–1379. DOI: 10.1021/acs.macromol.9b02474.

[132] Liu, W.; Zhang, R.; Huang, M.; Dong, X.; Xu, W.; Ray, N.; Zhu, J. Design and Structural Study of a

Triple-Shape Memory PCL/PVC Blend. *Polymer (Guildf.)*. 2016, *104*, 115–122. DOI: 10.1016/j.polymer.2016.09.079.

[133] Lai, S.; Wang, X. Shape Memory Properties of Olefin Block Copolymer (OBC)/ Poly (E -caprolactone) (PCL) Blends. *J. Appl. Polym. Sci.* 2017, *134*(44), 45475. DOI: 10.1002/app.45475.

[134] Chen, W. C.; Lai, S. M.; Chang, M. Y.; Liao, Z. C. Preparation and Properties of Natural Rubber (Nr)/ polycaprolactone (PCL) Bio-Based Shape Memory Polymer Blends. *J. Macromol. Sci. Part B Phys.* 2014, *53*(4), 645–661. DOI: 10.1080/00222348.2013.860304.

[135] Tian, G.; Zhu, G.; Xu, S.; Ren, T. A Novel Shape Memory Poly(ε-Caprolactone)/Hydroxyapatite Nanoparticle Networks for Potential Biomedical Applications. *J. Solid State Chem.* 2019, *272*(January), 78–86. DOI: 10.1016/j.jssc.2019.01.029.

[136] Ishii, S.; Uto, K.; Niiyama, E.; Ebara, M.; Nagao, T. Hybridizing Poly(ε -caprolactone) and Plasmonic Titanium Nitride Nanoparticles for Broadband Photoresponsive Shape Memory Films. *Appl. Mater. Interfaces*. 2016, *8*(8), 5634–5640. DOI: 10.1021/acsami.5b12658.

[137] Wang, W.; Liu, D.; Lu, L.; Chen, H.; Gong, T.; Lv, J.; Zhou, S. The Improvement of the Shape Memory Function of Poly(ϵ-Caprolactone)/Nano-Crystalline Cellulose Nanocomposites via Recrystallization under a High-Pressure Environment. *J. Mater. Chem. A*. 2016, *4*(16), 5984–5992. DOI: 10.1039/c6ta00930a.

[138] Lu, H.; Gou, J. Fabrication and Electroactive Responsive Behavior of Shape-Memory Nanocomposite Incorporated with Self-Assembled Multiwalled Carbon Nanotube Nanopaper. *Polym. Adv. Technol.* 2012, *23*(12), 1529–1535. DOI: 10.1002/pat.2074.

[139] Yang, P.; Zhu, G.; Shen, X.; Yan, X.; Nie, J. Poly (e-Caprolactone)-Based Shape Memory Polymers Crosslinked by Polyhedral Oligomeric Silsesquioxane. *RSC Adv.* 2016, *6*(93), 90212–90219. DOI: 10.1039/C6RA20431G.

[140] Qu, M.; Wang, H.; Chen, Q.; Wu, L.; Tang, P.; Fan, M.; Guo, Y.; Fan, H.; Bin, Y. A Thermally-Electrically Double-Responsive Polycaprolactone – Thermoplastic Polyurethane/Multi-Walled Carbon Nanotube Fiber Assisted with Highly Effective Shape Memory and Strain Sensing Performance. *Chem. Eng. J.* 2022, *427*, 131648. DOI: 10.1016/j.cej.2021.131648.

[141] Banerjee, A.; Chatterjee, K.; Madras, G. Enzymatic Degradation of Polycaprolactone-Gelatin Blend. *Mater. Res. Express*. 2015, *2*(4), 45303. DOI: 10.1088/2053-1591/2/4/045303.

[142] Feng, S.; Yue, Y.; Chen, J.; Zhou, J.; Li, Y.; Zhang, Q. Biodegradation Mechanism of Polycaprolactone by A Novel Esterase MGS0156: A QM/MM Approach. *Environ. Sci. Process. Impact.* 2020, *22*(12), 2332–2344. DOI: 10.1039/d0em00340a.

[143] Almeida, B. C.; Figueiredo, P.; Carvalho, A. T. P. Polycaprolactone Enzymatic Hydrolysis: A Mechanistic Study. *ACS Omega*. 2019, *4*(4), 6769–6774. DOI: 10.1021/acsomega.9b00345.

[144] Woodard, L. N.; Page, V. M.; Kmetz, K. T.; Grunlan, M. A. PCL–PLLA Semi-IPN Shape Memory Polymers (Smps): Degradation and Mechanical Properties. *Macromol. Rapid Commun.* 2016, *37*(23), 1972–1977. DOI: 10.1002/marc.201600414.

[145] Yang, Y.; Michalczyk, C.; Singer, F.; Virtanen, S.; Boccaccini, A. R. In Vitro Study of Polycaprolactone/ Bioactive Glass Composite Coatings on Corrosion and Bioactivity of Pure Mg. *Appl. Surf. Sci.* 2015, *355*, 832–841. DOI: 10.1016/j.apsusc.2015.07.053.

[146] Tsujimoto, T.; Takayama, T.; Uyama, H. Biodegradable Shape Memory Polymeric Material from Epoxidized Soybean Oil and Polycaprolactone. *Polymers (Basel)*. 2015, *7*(October), 2165–2174. DOI: 10.3390/polym7101506.

[147] Kim, J. J.; Singh, R. K.; Seo, S. J.; Kim, T. H.; Kim, J. H.; Lee, E. J.; Kim, H. W. Magnetic Scaffolds of Polycaprolactone with Functionalized Magnetite Nanoparticles: Physicochemical, Mechanical, and Biological Properties Effective for Bone Regeneration. *RSC Adv.* 2014, *4*(33), 17325–17336. DOI: 10.1039/c4ra00040d.

[148] Badrossamay, M. R.; McIlwee, H. A.; Goss, J. A.; Parker, K. K. Nanofiber Assembly by Rotary Jet-Spinning. *Nano Lett.* 2010, *10*(6), 2257–2261. DOI: 10.1021/nl101355x.

[149] Sadeghianmaryan, A.; Yazdanpanah, Z.; Soltani, Y. A.; Sardroud, H. A.; Nasirtabrizi, M. H.; Chen, X. Curcumin-loaded Electrospun Polycaprolactone/ Montmorillonite Nanocomposite: Wound Dressing Application with Anti-bacterial and Low Cell Toxicity Properties. *J. Biomater. Sci. Polym. Educ.* 2020, *31*(2), 169–187. DOI: 10.1080/09205063.2019.1680928.

[150] Mi, H. Y.; Jing, X.; Napiwocki, B. N.; Hagerty, B. S.; Chen, G.; Turng, L. S. B. Degradable Thermoplastic Polyurethane Based on Polycaprolactone-Block -polytetrahydrofuran- Block -polycaprolactone Copolymers for Soft Tissue Engineering. *J. Mater. Chem. B*. 2017, *5*(22), 4137–4151. DOI: 10.1039/c7tb00419b.

[151] Zhao, N.; Lv, Z.; Ma, J.; Zhu, C.; Li, Q. Fabrication of Hydrophilic Small Diameter Vascular Foam Scaffolds of Poly(ε-Caprolactone)/Polylactic Blend by Sodium Hydroxide Solution. *Eur. Polym. J.* 2019, *110*, 31–40. DOI: 10.1016/j.eurpolymj.2018.11.011.

[152] Świętek, M.; Brož, A.; Tarasiuk, J.; Wroński, S.; Tokarz, W.; Kozieł, A.; Błażewicz, M.; Bačáková, L. Carbon Nanotube/Iron Oxide Hybrid Particles and Their PCL-Based 3D Composites for Potential Bone Regeneration. *Mater. Sci. Eng. C*. 2019, *104*, 109913. DOI: 10.1016/j.msec.2019.109913.

[153] Collignon, A.; Granel, H.; Rochefort, G. Y.; Granel, H.; Rochefort, G. Y.; Dentaire, F. D. C.; Pathologies, L.; Descartes, P.; Dentaire, F. D. C.; Pathologies, L., et al. Bioactive Glass/Polycaprolactone Hybrid with a Dual Cortical/Trabecular Structure for Bone Regeneration. *ACS Appl. Bio Mater.* 2019, *2*(8), 3473–3483. DOI: 10.1021/acsabm.9b00407.

[154] Choi, E.; Jin, S.; Shim, J.; Yun, W.; Park, K.-H.; Choi, S.; Choi, E.; Jin, S.; Shim, J.-H.; Yun, W.-S. Fabrication and

Characterization of 3D – Printed Biocomposite Scaffolds Based on PCL and Silanated Silica Particles for Bone Tissue Regeneration. *Chem. Eng. J.* 2019, *360*, 519–530. DOI: 10.1016/j.cej.2018.11.176.

[155] Siqueira, I. A. W. B.; Koba, N.; Moura, D.; Paulo, J.; Machado, D. B.; Henrique, E.; Roberto, F.; De Sousa, E. Porous Membranes of the Polycaprolactone (PCL) Containing Calcium Silicate Fibers for Guided Bone Regeneration. *Mater. Lett.* 2017, *206*, 210–213. DOI: 10.1016/j.matlet.2017.07.011.

[156] Fereshteh, Z.; Fathi, M.; Bagri, A.; Boccaccini, A. R. Preparation and Characterization of Aligned Porous PCL/Zein Scaffolds as Drug Delivery Systems via Improved Unidirectional Freeze-Drying Method. *Mater. Sci. Eng. C.* 2016, *68*, 613–622. DOI: 10.1016/j.msec.2016.06.009.

[157] Dash, T. K.; Konkimalla, V. B. Polymeric Modification and Its Implication in Drug Delivery: Poly-ε-Caprolactone (PCL) as a Model Polymer. *Mol. Pharm.* 2012, *9*(9), 2365–2379. DOI: 10.1021/mp3001952.

[158] Chandrasiri, I.; Abebe, D. G.; Loku Yaddehige, M.; Williams, J. S. D.; Zia, M. F.; Dorris, A.; Barker, A.; Simms, B. L.; Parker, A.; Vinjamuri, B. P., et al. Self-Assembling PCL–PAMAM Linear Dendritic Block Copolymers (Ldbcs) for Bioimaging and Phototherapeutic Applications. *ACS Appl. Bio Mater.* 2020, *3*(9), 5664–5677. DOI: 10.1021/acsabm.0c00432.

[159] Yeroslavsky, G.; Umezawa, M.; Okubo, K.; Nigoghossian, K.; Thi Kim Dung, D.; Miyata, K.; Kamimura, M.; Soga, K. Stabilization of Indocyanine Green Dye in Polymeric Micelles for NIR-II Fluorescence Imaging and Cancer Treatment. *Biomater. Sci.* 2020, *8*(8), 2245–2254. DOI: 10.1039/c9bm02010a.

[160] Wu, S.; Su, F.; Magee, H. Y.; Meldrum, D. R.; Tian, Y. CRGD Functionalized 2,1,3-Benzothiadiazole (Btd)-containing Two-Photon Absorbing Red-Emitter-Conjugated Amphiphilic Poly(Ethylene Glycol)-: Block -poly(ϵ-caprolactone) for Targeted Bioimaging. *RSC Adv.* 2019, *9*(59), 34235–34243. DOI: 10.1039/c9ra06694b.

[161] Molina, B. G.; Cianga, L.; Bendrea, A. D.; Cianga, I.; Alemán, C.; Armelin, E. A. A. Heterografted Polythiophene Copolymer Containing Biocompatible/Biodegradable Side Chains for Use as an (Electro)active Surface in Biomedical Applications. *Polym. Chem.* 2019, *10*(36), 5010–5022. DOI: 10.1039/c9py00926d.

[162] Dong, X.; Wei, C.; Lu, L.; Liu, T.; Lv, F. Fluorescent Nanogel Based on Four-Arm PEG-PCL Copolymer with Porphyrin Core for Bioimaging. *Mater. Sci. Eng. C.* 2016, *61*, 214–219. DOI: 10.1016/j.msec.2015.12.037.

[163] Huang, S.; Liu, S.; Wang, K.; Yang, C.; Luo, Y.; Zhang, Y.; Cao, B.; Kang, Y.; Wang, M. Highly Fluorescent and Bioresorbable Polymeric Nanoparticles with Enhanced Photostability for Cell Imaging. *Nanoscale.* 2015, *7*(3), 889–895. DOI: 10.1039/c4nr05576d.

[164] Yang, C.; Huang, S.; Wang, X.; Wang, M. Theranostic Unimolecular Micelles of Highly Fluorescent Conjugated Polymer Bottlebrushes for Far Red/near Infrared Bioimaging and Efficient Anticancer Drug Delivery. *Polym. Chem.* 2016, *7*(48), 7455–7468. DOI: 10.1039/c6py01838f.

[165] Ruan, Z.; Liu, L.; Fu, L.; Xing, T.; Yan, L. An Amphiphilic Block Copolymer Conjugated with Carborane and a NIR Fluorescent Probe for Potential Imaging-Guided BNCT Therapy. *Polym. Chem.* 2016, *7*(26), 4411–4418. DOI: 10.1039/c6py00799f.

[166] Tabesh, E.; Kharaziha, M.; Mahmoudi, M.; Shahnam, E.; Rozbahani, M. Biological and Corrosion Evaluation of Laponite®: Poly(Caprolactone) Nanocomposite Coating for Biomedical Applications. *Colloids Surfaces A Physicochem. Eng. Asp.* June 2019, *583*, 123945. DOI: 10.1016/j.colsurfa.2019.123945.

[167] Jokar, M.; Darvishi, S.; Torkaman, R.; Kharaziha, M.; Karbasi, M. Corrosion and Bioactivity Evaluation of Nanocomposite PCL-Forsterite Coating Applied on 316L Stainless Steel. *Surf. Coatings Technol.* 2016, *307*, 324–331. DOI: 10.1016/j.surfcoat.2016.08.094.

[168] Mojarad Shafiee, B.; Torkaman, R.; Mahmoudi, M.; Emadi, R.; Karamian, E. An Improvement in Corrosion Resistance of 316L AISI Coated Using PCL-Gelatin Composite by Dip-Coating Method. *Prog. Org. Coatings.* 2019, *130*(January), 200–205. DOI: 10.1016/j.porgcoat.2019.01.057.

[169] Capellato, P.; Silva, G.; Popat, K.; Simon-Walker, R.; Alves Claro, A. P.; Zavaglia, C. Cell Investigation into the Biocompatibility of Adult Human Dermal Fibroblasts with PCL nanofibers/TiO 2 Nanotubes on the Surface of Ti–30Ta Alloy for Biomedical applications. *Artif. Organs.* 2020, *44*(8), 877–882. *Early view*. DOI: 10.1111/aor.13713.

[170] Bakhsheshi-Rad, H. R.; Hamzah, E.; Shuang, C. P.; Berto, F. Preparation of Poly(ε-caprolactone)-hydroxyapatite Composite Coating for Improvement of Corrosion Performance of Biodegradable Magnesium. *Mater. Des. Process. Commun.* 2020, *2*(4), *early view*. DOI: 10.1002/mdp2.170.

[171] Palanisamy, M. S.; Kulandaivelu, R.; Nellaiappan, S. N. T. S. Improving the Corrosion Resistance and Bioactivity of Magnesium by a Carbonate Conversion-Polycaprolactone Duplex Coating Approach. *New J. Chem.* 2020, *44*(12), 4772–4785. DOI: 10.1039/c9nj06030h.

[172] Kim, J.; Mousa, H. M.; Park, C. H.; Kim, C. S. Enhanced Corrosion Resistance and Biocompatibility of AZ31 Mg Alloy Using PCL/ZnO NPs via Electrospinning. *Appl. Surf. Sci.* 2017, *396*, 249–258. DOI: 10.1016/j.apsusc.2016.10.092.

[173] Bakhsheshi-Rad, H. R.; Ismail, A. F.; Aziz, M.; Hadisi, Z.; Omidi, M.; Chen, X. Antibacterial Activity and Corrosion Resistance of Ta 2 O 5 Thin Film and Electrospun PCL/MgO-Ag Nanofiber Coatings on Biodegradable Mg Alloy Implants. *Ceram. Int.* 2019, *45*(9), 11883–11892. DOI: 10.1016/j.ceramint.2019.03.071.

[174] Huang, W.; Mei, D.; Zhang, J.; Chen, D.; Li, J.; Wang, L.; Zhou, Y.; Zhu, S.; Guan, S. Improved Corrosion Resistance and Cytocompatibility of Mg–Zn–Y–Nd

Alloy by the Electrografted Polycaprolactone Coating. *Colloids Surfaces A Physicochem. Eng. Asp.* 2021, *629* (June), 127471. DOI: 10.1016/j.colsurfa.2021.127471.

[175] Sangroniz, A.; Sangroniz, L.; Hamzehlou, S.; Del Río, J.; Santamaria, A.; Sarasua, J. R.; Iriarte, M.; Leiza, J. R.; Etxeberria, A. Lactide-Caprolactone Copolymers with Tuneable Barrier Properties for Packaging Applications. *Polymer (Guildf.)*. 2020, *202*, 122681. DOI: 10.1016/j.polymer.2020.122681.

[176] Ahmed, J.; Mulla, M.; Jacob, H.; Luciano, G.; T.b, B.; Almusallam, A. Polylactide/Poly(ε-Caprolactone)/Zinc Oxide/Clove Essential Oil Composite Antimicrobial Films for Scrambled Egg Packaging. *Food Packag. Shelf Life*. 2019, *21*, 100355. DOI: 10.1016/j.fpsl.2019.100355.

[177] Milovanovic, S.; Hollermann, G.; Errenst, C.; Pajnik, J.; Frerich, S.; Kroll, S.; Rezwan, K.; Ivanovic, J. Supercritical CO2 Impregnation of PLA/PCL Films with Natural Substances for Bacterial Growth Control in Food Packaging. *Food Res. Int.* 2018, *107*, 486–495. DOI: 10.1016/j.foodres.2018.02.065.

[178] Qin, Y.; Liu, D.; Wu, Y.; Yuan, M.; Li, L.; Yang, J. Effect of PLA/PCL/Cinnamaldehyde Antimicrobial Packaging on Physicochemical and Microbial Quality of Button Mushroom (Agaricus Bisporus). *Postharvest Biol. Technol.* 2015, *99*, 73–79. DOI: 10.1016/j.postharvbio.2014.07.018.

[179] Correa, J. P.; Molina, V.; Sanchez, M.; Kainz, C.; Eisenberg, P.; Massani, M. B. Improving Ham Shelf Life with a Polyhydroxybutyrate/Polycaprolactone Biodegradable Film Activated with Nisin. *Food Packag. Shelf Life*. 2017, *11*, 31–39. DOI: 10.1016/j.fpsl.2016.11.004.

[180] Hadj-Hamou, A. S.; Yahiaoui, F. Performances of PCL/PVC/Organoclay Nanobioblends Films for Packaging Applications. *Macromol. Symp.* 2019, *386*(1), 1–8. DOI: 10.1002/masy.201800239.

[181] Alix, S.; Mahieu, A.; Terrie, C.; Soulestin, J.; Gerault, E.; Feuilloley, M. G. J.; Gattin, R.; Edon, V.; Ait-Younes, T.; Leblanc, N. Active Pseudo-Multilayered Films from Polycaprolactone and Starch Based Matrix for Food-Packaging Applications. *Eur. Polym. J.* 2013, *49* (6), 1234–1242. DOI: 10.1016/j.eurpolymj.2013.03.016.

[182] Ortega-Toro, R.; Contreras, J.; Talens, P.; Chiralt, A. Physical and Structural Properties and Thermal Behaviour of Starch-Poly(E{open}-Caprolactone) Blend Films for Food Packaging. *Food Packag. Shelf Life*. 2015, *5*, 10–20. DOI: 10.1016/j.fpsl.2015.04.001.

[183] Khalid, S.; Yu, L.; Feng, M.; Meng, L.; Bai, Y.; Ali, A.; Liu, H.; Chen, L. Development and Characterization of Biodegradable Antimicrobial Packaging Films Based on Polycaprolactone, Starch and Pomegranate Rind Hybrids. *Food Packag. Shelf Life*. 2018, *18*, 71–79. DOI: 10.1016/j.fpsl.2018.08.008.

[184] Wang, K.; Lim, P. N.; Tong, S. Y.; Thian, E. S. Development of Grapefruit Seed Extract-Loaded Poly (ε-Caprolactone)/Chitosan Films for Antimicrobial Food Packaging. *Food Packag. Shelf Life*. 2019, *22*, 100396. DOI: 10.1016/j.fpsl.2019.100396.

[185] Jeong, S.; Lee, H. G.; Cho, C. H.; Yoo, S. R. Characterization of Multi-Functional, Biodegradable Sodium Metabisulfite-Incorporated Films Based on Polycarprolactone for Active Food Packaging Applications. *Food Packag. Shelf Life*. 2020, *25*(April), 100512. DOI: 10.1016/j.fpsl.2020.100512.

[186] Benhacine, F.; Ouargli, A.; Hadj-Hamou, A. S. Preparation and Characterization of Novel Food Packaging Materials Based on Biodegradable PCL/Ag-Kaolinite Nanocomposites with Controlled Release Properties. *Polym. Technol. Mater.* 2019, *58*(3), 328–340. DOI: 10.1080/03602559.2018.1471714.

[187] Xie, J.; Hung, Y. C. UV-A Activated TiO2 Embedded Biodegradable Polymer Film for Antimicrobial Food Packaging Application. *LWT-Food Sci. Technol.* 2018, *96*, 307–314. DOI: 10.1016/j.lwt.2018.05.050.

[188] Cai, Y.; Guan, J.; Wang, W.; Wang, L.; Su, J.; Fang, L. PH and Light-Responsive Polycaprolactone/Curcumin@zif-8 Composite Films with Enhanced Antibacterial Activity. *J. Food Sci.* 2021, *86*(8), 3550–3562. DOI: 10.1111/1750-3841.15839.

[189] Angelica, D.; Alberto, J.; Costa, V.; Greque, M.; Morais, D. A Novel Nanocomposite for Food Packaging Developed by Electrospinning and Electrospraying. *Food Packag. Shelf Life*. 2019, *20*, 100314. DOI: 10.1016/j.fpsl.2019.100314.

[190] Hasanpour Ardekani-Zadeh, A.; Hosseini, S. F. Electrospun Essential Oil-Doped Chitosan/Poly(ε-Caprolactone) Hybrid Nanofibrous Mats for Antimicrobial Food Biopackaging Exploits. *Carbohydr. Polym.* 2019, *223*, 115108. DOI: 10.1016/j.carbpol.2019.115108.

[191] Mathiazhagan, S.; Periasamy, V.; Vadivel, A. Ecofriendly Antimicrobial Acalypha Indica Leaf Extract Immobilized Polycaprolactone Nanofibrous Mat for Food Package Applications. *J. Food Process. Preserv.* 2021, *45*(4), 1–12. DOI: 10.1111/jfpp.15302.

[192] Lin, W.; Ni, Y.; Pang, J. Size Effect-Inspired Fabrication of Konjac Glucomannan/ Polycaprolactone Fiber Films for Antibacterial Food Packaging. *Int. J. Biol. Macromol.* 2020, *149*, 853–860. DOI: 10.1016/j.ijbiomac.2020.01.242.

[193] Sugane, K.; Yoshioka, Y.; Shimasaki, T.; Teramoto, N.; Shibata, M. Self-Healing 8-Armed Star-Shaped ε-caprolactone Oligomers Dually Crosslinked by the Diels-Alder and Urethanization Reactions. *Polymer (Guildf.)*. 2018, *144*, 92–102. DOI: 10.1016/j.polymer.2018.04.045.

[194] Rodriguez, E. D.; Luo, X.; Mather, P. T. Linear/Network Poly (ε -caprolactone) Blends Exhibiting Shape Memory Assisted Self-Healing (SMASH). *ACS Appl. Mater. Interfaces.* 2011, *3*(2), 152–161. DOI: 10.1021/am101012c.

[195] Lai, S. M.; Liu, J. L.; Huang, Y. H. Preparation of Self-Healing Natural Rubber/Polycaprolactone (NR/PCL) Blends. *J. Macromol. Sci. Part B Phys.* 2020, *59* (9), 587–607. DOI: 10.1080/00222348.2020.1757218.

[196] Xu, X.; Fan, P.; Ren, J.; Cheng, Y.; Ren, J.; Zhao, J.; Song, R. Self-Healing Thermoplastic Polyurethane (Tpu)/polycaprolactone (PCL) /Multi-Wall Carbon Nanotubes (Mwcnts) Blend as Shape-Memory Composites. *Compos. Sci. Technol.* 2018, *168*, 255–262. DOI: 10.1016/j.compscitech.2018.10.003.

[197] Lutz, A.; Van Den Berg, O.; Damme Van, J.; Verheyen, K.; Bauters, E.; De Graeve, I.; Du Prez, F. E.; Terryn, H. A Shape-Recovery Polymer Coating for the Corrosion Protection of Metallic

Surfaces. *ACS Appl. Mater. Interfaces*. 2015, *7*, 175–183. DOI: 10.1021/am505621x.

[198] Huang, Y.; Deng, L.; Ju, P.; Huang, L.; Qian, H.; Zhang, D.; Li, X.; Terryn, H. A.; Mol, J. M. C. Applications of Polymer, Composite, and Coating Materials Triple-Action Self-Healing Protective Coatings Based on Shape Memory Polymer (SMP) Containing Dual-Function Microspheres Triple-Action Self-Healing Protective Coatings Based on Shape Memory Pol. *ACS Appl. Mater. Interfaces*. 2018, *10*(27), 23369–23379. DOI: 10.1021/acsami.8b06985.

[199] Rajitha, K.; Mohana, K. N. Application of Modified Graphene Oxide – Polycaprolactone Nanocomposite Coating for Corrosion Control of Mild Steel in Saline Medium. *Mater. Chem. Phys*. 2020, *241*, 122050. DOI: 10.1016/j.matchemphys.2019.122050.

[200] Ai, X.; Xie, Q.; Ma, C.; Zhang, G. Fouling Release Coating Consisting of Hyperbranched Poly(ε-Caprolactone)/Siloxane Elastomer. *ACS Appl. Polym. Mater*. 2020, *2*(4), 1429–1437. DOI: 10.1021/acsapm.9b01056.

[201] Ding, Z.; Li, J.; Xin, W.; Zhang, G.; Luo, Y. Progress in Organic Coatings Low Gloss Waterborne Polyurethane Coatings with Anti-Dripping and Flame Retardancy via Montmorillonite Nanosheets. *Prog. Org. Coatings*. 2019, *136*, 105273. DOI: 10.1016/j.porgcoat.2019.105273.

[202] Arya, R. K.; Kaur, J.; Chandra, A.; Ahuja, S.; Rawat, M.; Sharma, J. Designing of Biodegradable Polycaprolactone: Binary and Ternary Coatings to Minimize the Defects and Cost of Solvent(S). *J. Appl. Polym. Sci*. 2021, *138*(35), 1–15. DOI: 10.1002/app.50888.

[203] Akat, H.; Erdem Yayayürük, A.; Yayayürük, O.; Akat, H. Thiol – Ended Polycaprolactone : Synthesis, Preparation and Use in Pb (II) and Cd (II) Removal from Water Samples. *Mater. Today Commun*. 2021, *29*, 102908. DOI: 10.1016/j.mtcomm.2021.102908.

[204] Reshmi, C. R.; Sundaran, S. P.; Juraij, A.; Athiyanathil, S. RSC Advances Fabrication of Superhydrophobic Polycaprolactone/Beeswax Electrospun Separation †. *RSC Adv*. 2016, *7*, 2092–2102. DOI: 10.1039/C6RA26123J.

[205] He, N.; Li, L.; Chen, J.; Zhang, J.; Liang, C. Extraordinary Superhydrophobic Polycaprolactone-Based Composite Membrane with an Alternated Micro-Nano Hierarchical Structure as an Eco-Friendly Oil/Water Separator. *ACS Appl. Mater. Interfaces*. 2021, *13*(20), 24117–24129. DOI: 10.1021/acsami.1c03019.

[206] Zhang, X.; Zhao, J.; Ma, L.; Shi, X.; Li, L. Biomimetic Preparation of a Polycaprolactone Membrane with a Hierarchical Structure as a Highly E Ffi Cient Oil – Water Separator. *J. Mater. Chem. A*. 2019, *7*, 24532–24542. DOI: 10.1039/c9ta08660a.

[207] Eksiler, K.; Andou, Y.; Nakayama, N.; Yoshinaga, K.; Shirai, Y. Design of Biodegradable PCL/PI Films as a Joining Tape for Grafting Plant. *Environ. Technol. (United Kingdom)*. 2017, *38*(18), 2362–2372. DOI: 10.1080/09593330.2016.1261186.

[208] Chen, F.; Xu, L.; Tian, Y.; Caratenuto, A.; Liu, X.; Zheng, Y. Electrospun Polycaprolactone Nano Fi Ber Composites with Embedded Carbon Nanotubes/Nanoparticles for Photothermal Absorption. *ACS Appl. Nano Mater*. 2021, *4*(5), 5230–5239. DOI: 10.1021/acsanm.1c00623.

Environment, Development and Sustainability
https://doi.org/10.1007/s10668-022-02416-1

# Analysis of land use and land cover changes and their impact on temperature using landsat satellite imageries

**Rubeena Vohra**[1] · **K. C. Tiwari**[2]

Received: 1 September 2021 / Accepted: 28 April 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

**Abstract**
Urban growth and the changing scenario of Land Use and Land Cover (LULC) have been an increasing trend in both towns and cities. The higher rate of transformation from non-built up land to the impervious area becomes a warning symbol of Land surface temperature variations. In this study, an attempt has been made to determine the transition of natural land area and its impact on Land surface temperature (LST) in Vellore district, Tamil Nadu, India. According to the current statistics, the study area records the hottest climate crossing 40° mark in recent years. This is mainly due to the minimum rainfall, the ground level is 200 m just above the sea level, and the pollution caused by tanneries. Landsat imageries are collected for three different years 1994, 2002, and 2018 that map the LULC into agricultural, water bodies, built-up land, and barren land classes. The major purpose of this research is to (i) analyze changes of LULC in and around Vellore city, (ii) categorize the images into various classes like vegetative and non-vegetative land, (iii) Assessment of Spatio-temporal variations in LST and link with classes and urbanization growth using satellite images. The LULC impact on LST is analyzed with the widely used Getis–Ord statistics. The simulation result shows that the built-up area raises to 81%, vegetation land decline by about −65% for the years 1994–2018 respectively. It is observed that LST has attained the highest degree in the built-up class due to the unplanned LULC changes and the conversion of built-up areas. The overall accuracy is achieved at about 92, 89, and 91% for three different years respectively. Based on the obtained result, this can be adopted for the development of rural, and urban areas in the coming future.

**Keywords** Land surface temperature · Land use · Land cover · Remote sensing · Normalized differential vegetation index (NDVI) · Change detection

✉ Rubeena Vohra
rubeena.vohra@gmail.com

[1] Department of Electronics and Communication Engineering, Bharati Vidyapeeth's College of Engineering (BVCOE), New Delhi, India

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

## 1 Introduction

In recent decades, urbanization has witnessed a significant impact on Land Use Land Cover (LULC) which makes a rapid transmission of agricultural, water bodies to the built-up area (Fonseka et al., 2019). Due to urbanization, the urban population has completely transformed from its current statistics of 55% to 68% by 2050 to live in cities (Gohain et al., 2021). The migration of humans to cities has both negative and positive outcomes (Uttara et al., 2012). The advantages of the urban population lead to the improvement of quality of life and surplus job opportunities where the negative impacts result in causes serious environmental pollution, health problems, enhance Urban Heat Island (UHI), and burdens on the city infrastructure. UHI formation is a serious impact of urbanization in which the urban areas have significantly higher temperatures than the surrounding rural regions. Also, the changes in LULC alter the regional climate, and lower the biodiversity level, and thereby increase the LST. The higher LST results in increasing the surface temperature, disrupts the natural habitat, UHI creation which could be the greatest concern for geographers, health activists, urban climatologists, and regional planners (Maimaitiyiming et al., 2014; Huang et al., 2018). The extensive growth of urbanization and industrialization in developing or developed countries makes the situation worsen and the urban cities to be problematic. Research studies on the consequences of rising temperature and UHI are in the greatest analysis in the twenty-first century.

Traditionally, UHI is assessed based on a comparison between air temperature in rural and urban areas. Direct monitoring for the estimation of Change detection in LULC and LST is time-consuming, rigorous labor charge, and is error-prone (Hart et al., 2009; Lilly Rose et al., 2009). Currently, remote sensors are utilized in the extraction of LST due to their easy accessibility, faster acquisition of data, lesser cost, high spatial and temporal resolutions. Remote sensing (RS) is used for the LULC mapping and the change variation valuation adopted by various techniques (Balogun et al., 2017; Kafy et al., 2020). It was stated that Landsat satellite images are more appropriate in the estimation of changes in LULC and urban area analysis. For the estimation of UHI, high-resolution Landsat satellite series Thematic Mapper (TM), Enhanced TM (ETM+), and Operational Land Imager (OLI) imageries are used over the past years (Shi et al., 2018; Das et al., 2020). Also, for the LST at low-resolution to high-resolution scale, Thermal Infrared Region (TIR) captures the surface temperature, emissivity response over the surface. Landsat 7 (EHI+) and 8 (OLI, TIRS) provide high spatial resolution imageries and their impact on LST in urban thermal monitoring. This data series has the TIR of (10.4−12.5 m) calculates the brightness temperature using various algorithms provided by NASA rather than the geostationary satellite has TIR region calculates LST at regional to global scale (Wang et al., 2018). Compared to MODIS satellite data, Landsat data has the greater advantage of a longer-temporal range (Grigoraș G et al., 2019). The combination of multi-spectral bands with TIR of Landsat imageries enhances the ability of RS techniques to compute LULC, and LST simultaneously. LST can be estimated easily from the RS data at various spatial, temporal scales give a better understanding of temperature distribution compared to observation data. Studies performed on classification tasks obtained from multispectral images use supervised classification algorithms. Most of the researchers employ a maximum likelihood classifier (MLC) for the categorization into multi-classes. Also, Support Vector Machine (SVM) classifier is employed for the effective learning of the features and classification of satellite images (Vohra et al., 2020). In (Vohra et al., 2019), three spatial shape features like width, length,

and pixel shape index (PSI) are extracted for the identification of buildings and roads in urban areas and generate the classification map based on SVM model. The LULC changes and LST are the major factors in Land use management and the UHI analysis estimated by the NDVI parameter (Madanian et al., 2018). As from the relationship estimated between NDVI and LST, the change in LULC indirectly affects the LST due to the differentiation in NDVI (Zhang et al., 2009).

Other than NDVI spectral indices, Biophysical Composition Index (BCI), Normalized Difference Soil Index (NDSI) are employed in LULC mapping. BCI is designed for the active separation of bare soil, and built-up regions. An improved version of BCI is developed such as Modified Normalized Difference Soil Index (MNDSI) and Normalized Ratio Urban Index (NRUI) enhances the above separability process (Piyoosh et al., 2018; Pan et al., 2010). From the basis of NRUI, Normalized Difference Anthropogenic Impervious Surface Index is designed for semi-automatic mapping. Many algorithms like multi-channel, mono-window, and split-window are employed earlier for the analysis of LST changes in terms of different characteristics, and application scenarios (Piyoosh et al., 2018). Several researchers are conducted the analysis of long-term LULC changes over temperature using multi-temporal RS data. This study investigates the impacts on the relationship between LULC, LST, and NDVI using remote-sensing imagery for the LU management by utilizing the Landsat TM/ETM + and OLI data in Vellore district, India. Vellore, the headquarters of Vellore district is one of the eleven district that forms north region of TN. This district is one of the leading ones that grows consistently in the educational sector. Vellore connects Mumbai and Chennai and to the other important Indian cities. Vellore is also known for its leathery product account for about 37% of country exports. The higher rate of occupation is given by the leather and industrial regions around the cities. It ranks 7th place in terms of geographical area with 1.74Lac Hectare net cultivated region. Vellore is selected for the smart city project mission by the Government of India between the years 2019 and 2023. Also, VIT green Vellore project was launched in 2017 to improve the green cover and water levels caused due to the effects of delays in monsoon. This mission is launched to create the awareness of planting trees, water resources conservation, preventing pollution, and preventing drainage in Palar River. The Government of India's Digital India program predicts the assistance of location-based/GIS-based decision-making for planning. GIS platform recreated as Bharat maps is a multilayered GIS platform that provides base maps, and satellite images, and hybrid maps based on global geospatial standards.

***Contribution is summarized as follows;***

- The present study investigates the variations of the temperature and their relationship between vegetation indices for three different years (1994, 2002, 2018) using Landsat satellite imageries.
- This work highlights the assessment of change for the region in terms of four different classes.
- Hot spot areas are detected and the rate of increase in changes is provided for the effective analysis of the LST pattern.
- Experimental analysis is evaluated with LST by various land use indices, and the hotspot zones are detected for the study region.

The rest of the section is modeled as follows; Sect. 2 describes the recent works carried out in the field of LULC over LST analysis using satellite images. In Sect. 3, database collection, study area description, and the steps for the analysis of LST over the changes of LULC are discussed. The experimental analysis is conducted and their LULC change

analysis, accuracy assessment, LST measurements, and the Hot spot evaluation are discussed in Sect. 4. Finally, the overall result of the study region is explained in Sect. 5.

## 2 Related work

Recent works presented on changes in LULC and their impacts on the temperature variations are listed as follows;

**Nayak and Mandal**, (2019), presented an Observation minus Reanalysis (OMR) model that analyses the temperature trend over India and predicts the LULC for the data collected at the time of 1981–2006. Here, the data is collected based on the data availability lasts for at least 20-years long. The total period is classified into three classes to estimate the decadal point. Vegetated lands refer to the agricultural land, grassland, dense forest, open forest whereas barren lands, snow areas are considered as a non-vegetated class. For the accuracy estimation, the LULC maps are evaluated with the forest survey and use random samples for high-resolution satellite images.

**Wang et al.**, (2019), analyze the variability of LULC Spatio-temporal changes in the Pearl River Delta region for the year ranges from 2000–2015. Initially, the historical Local Climate Zone (LCZ) maps were based on World Urban Database. The images are first preprocessed and mosaicked, resampled to 100 m resolution, and clipped to fit in the range. Then the samples are digitized based on the local knowledge and the classification is trained by a random forest classifier. The classification rule is created by the random forest and obtains the unbiased error output without additional test data. Further, manual corrections are performed for the maps with the assumption of the areas remain built types without any change. Then, the change detection is conducted to reveal the transformation made during the specific periods. Moreover, the surface urban heat island zones are filtered based on the variation of the temperature calculated from the mean and standard deviation measures. The area specified with high values is classified as high zone, lesser values as low SUHI zone, and the values between low and high are termed as medium zones.

**Choudhury et al.**, (2019), presented a study that attempted to analyze the surface temperature that utilizes the multi-temporal satellite data conducted in the Asansol-Durgapur region. The land surface temperature data is collected in terms of summer, winter, and post-monsoon periods. This method utilizes the Landsat TM and OLI data and the data first performs geometric corrections by the Google Earth images. For the classification task, the maximum likelihood algorithm is applied to present the changes in LULC. In the accuracy estimation, an error matrix of the referenced data is obtained to get the producer's and user's accuracy. Also, kappa accuracy is computed for obtaining the land use maps that provide interclass discrimination in classification results. Pearson's correlation measurement is performed based on the relationship between the surface temperature, water-body, and vegetation. Further, regression analysis is performed to visualize the changing variations in temperature with spatial indices.

**Tariq et al.,** (2020), investigate satellite images acquired from the arid potohar region of Pakistan which analyses the Spatio-temporal trends in temperature acquired in the time-interval 2000–2015. At first, preprocessing is performed for the strips, haze, and gaps removal by Environment for visualizing images (ENVI). Then the land cover indices are analyzed which are obtained from the visible portion of red, green, near-infrared, and shortwave infrared images of LANDSAT 7 and 8 images. For the classification of land area, an MLC is utilized that classifies the areas into water bodies, vegetation, barren, and

built-up area. Here, the geospatial and statistical analysis is performed to obtain the active band for the classification. Then the radiance values are extracted to compute the brightness temperature expressed in kelvin.

**John et al.,** (2020), presented the classification of LULC temperature analysis using the remotely sensed satellite images captured from the Wayanad district between 2004 and 2018. For the classification, a maximum likelihood classifier is adopted that categorizes the land into six features water body, forest, dense, agricultural, built, and paddy field. The preprocessing process named geometrical correction is performed based on the selection of ground control points from the satellite images. Also, radiometric correction is applied to solve the distortion error in the satellite images. For image classification, supervised and unsupervised learning algorithms are used to determine the various features and estimate the land characteristics. In the accuracy assessment, the classified results are compared to the reference information in which the confusion matrix technique is applied as a post-comparison task. For the temperature analysis task, the digitized information is processed to convert to temperature are then rescaled and further changed to brightness temperature using thermal constants.

**Piyoosh et al.,** (2020), proposed a Normalized Ratio Urban index (NRUI) for the assessment of LULC changes using Landsat multi-sensor satellite data captured from Dehradun district, India over the period 1991–2019. In the classification task, three bands include visible, Near Infra-Red (NIR), and SWIR are selected from the satellite images based on Optimum Index Factor (OIF) criteria. The NRUI is computed based on the normalized ratio urban index metric for the multi-sensor data. Also, the existing spectral indices are calculated and then the LULC maps are generated from the MLC supervised algorithm. The classification of LULC maps is categorized into different classes like agriculture, scrub, forest, plantation, water, etc. A Jefferies-Matusita (JM) distance measure is computed for separability analysis that estimates the best spectral indices and raw data are selected. The integration of sub-classes is carried out by recoding process and accuracy estimation by false-color composite data as a reference. Finally, normalized LST difference is computed for each year for the estimation of changed areas.

**Thakur et al.,** (2021), explored the temperature changes, from LULC, utilized ETM+, Operational Land Imager images, and Landsat TM collected from the Sundarbans Reserve Forest of the south part of West Bengal. Here the LULC maps are acquired for the years 2000, 2010, and 2017 respectively. The initial corrections like radiometric and atmospheric corrections are performed for the input satellite imageries. To enhance the accuracy, and error reduction, the images are reclassified and the post-classification ground truth comparison is performed. Then the recoding process is applied for the correction of wrongly classified pixels to the corrected class. For the detection of changes between two images, the change detection technique is carried out by the matrix union function. The land surface temperature retrieval task is computed in terms of the Single window algorithm. The LST is calculated based on the indices like conversion to Top of the atmosphere radiance, NDVI calculation, proportion of vegetation calculation, and ground emissivity calculation.

The overall summary for the above-discussed work is mentioned in Table 1.

Normalized Difference Water Index –**NDWI**, Top of the Atmosphere –**TOA**, Green NDVI-**GNDVI**, Difference Vegetation Index-**DVI**, Normalized Red-**NR**, Normalized Green-**NG**, Normalized NIR-**NNIR**.

From the analysis of the recent studies, it is noted that several works investigate the urban areas for LULC and LST relationships in terms of both local and global weather conditions. The result shows the transformation of vegetated land into the built-up area rises LST which is caused due by the higher rate of the urban population. Also, the concept of

**Table 1** Summary of existing approaches in the assessment of LULC changes

| Author/year /ref number | Technique used | Features | Image type | Place and year | LC/LU indices | Advantages | Disadvantages |
|---|---|---|---|---|---|---|---|
| Nayak et al., (2019) | OMR model | Changes in temperature trend | Landsat MSS, Landsat TM, and Landsat ETM + | India 1981 – 2006 | NDVI | Classification task is better | Great impact of result is not attained |
| Wang et al., (2019) | WUDAPT protocol based LCZ mapping | Built and natural land cover are differentiated | Landsat, and MODIS | Pearl River delta region, china 2000 – 2015 | – | Influence of LCZ is studied better | Quality and resolution of input image is not good |
| Choudhury et al., (2019) | – | LST changes based on LULC | LANDSAT TM and LANDSAT OL | Asansol-Durgapur region, WB India 1990 – 2015 | NDVI, NDBI, and NDWI | Correlation analysis shows strong changes | Accessibility rate is low |
| Tariq et al., (2020) | Geospatial techniques | Analyses LST variation over LULC | Landsat-7, LANDST-8 | Arid potohor region, Pakistan 2000 – 2015 | NDVI, NDWI, NDSI, NDBI, and BI | Effective analysis of temperature variations | Change in different classes are not estimated |
| John et al., (2020) | – | Analysis of LULC and LST Assessment | LANDST ETM +, LISS III, IRS P6 | Wayanad, Kerala, India 2004 and 2018 | NDVI, MNDWI, DVI, GNDVI, VI, NR, NG, and NNIR | Changing impacts are depicted accurately | Classification performance is challenging |
| Piyoosh et al., (2020) | NRUI | Highlight the change inbuilt area | Landsat OLI | Dehradun, India 1991 – 2019 | NDVI, NDWI, NDSI, and BCI | Better seperability of changes | Low-quality images are used |
| Thakur et al., (2021) | – | Specifies the type of changes in greenness area | Landsat TM, Landsat ETM +, OLI | Mangrove forest, Sundarbans, India 2000, 2010, and 2017 | NDVI, TOA spectral radiance, ground emissivity | Better accuracy | Large area distribution of LST is not possible |

LST estimates the variations of the unpredictable changes in surface temperature and their relationship to LU/LC using RS satellite data. To understand the variations of the surface temperature, various spectral land indices have to be employed to determine the land-use patterns with the LST. Hence, an analysis is needed to estimate the significant relationship between LULC & LST in various categories, and the correlation between the LULC indices and LST of the regions in Vellore, India from the year 1994–2018. Also, the comparative analysis for the variations of LST at different land cover classes and the association of extracted land cover indices with LST are presented.

Researchers widely pay attention to the concept of LST to estimate the variations of the unpredictable changes in surface temperature and their relationship to LU/LC using RS satellite data. To understand the variations of the surface temperature, various spectral land indices are employed that determine the land-use patterns with the LST. This study attempts to determine the significant relationship between LULC & LST in various categories, and the correlation between the LULC indices and LST of the regions.

The different land cover indices like NDWI, TOA, GNDVI, DVI, NR, NG, and NNIR are obtained from visible portion (VP) Green (G), Red (R), Near Infrared (NIR), and Short Wave Infra-Red (SWIR) reflectance bands and are extracted from the LANDSAT 7 (ETM+) and LANDSAT 8 (OLI & TIRS) images. These indices are helpful in LULC mapping for the estimation of buildup area, barren land, water bodies, and vegetation cover.

## 3 Materials and methodology

The primary goal of this research work is to analyze the impact of LST based on LU/LC changes for the three different years using satellite images. This study is classified into four different stages; (a) Preprocessing,( b) LU/LC mapping, (c) Accuracy assessment, (d) change detection, and (e) LST analysis.

Initially, the input satellite imageries are preprocessed to attain more detailed, and precise RS information. Next, LULC maps are obtained from the MLC technique that classifies the satellite image into different classes like vegetative land (agricultural land, and water bodies), non-vegetative land (built-up land, and barren land). In the third phase, post-classification is performed for the correctness of the classification process using the confusion matrix. It is computed by comparing the categorized pixels of the Landsat dataset to the ground truth (Test pixels). For the assessment of accuracy, the parameters such as overall accuracy, producer accuracy, user's accuracy, and kappa coefficient are calculated using a confusion matrix. Then the change detection is computed for three different sub-periods and categorized. Finally, the LST retrieval task is performed based on the estimation of land surface emissivity. Further, the relationship between indices, and LST are calculated for each land class.

### 3.1 Study area

This paper investigates the Vellore district has the latitudes of 12° 15' to 13° 15' North and longitude of 78° 20' to 79° 50' East. For this district, the geographical cover is 6077 sq. km and the headquarters is Vellore city which is separated into four zones and classified further into sixty wards, with an area of 87.915 km $^2$. The topography of the Vellore district is almost flat with a slope from West to East and elevation is about 216 m (709 ft). The geographical map of the Vellore district is illustrated in Fig. 1. Vellore is one of the

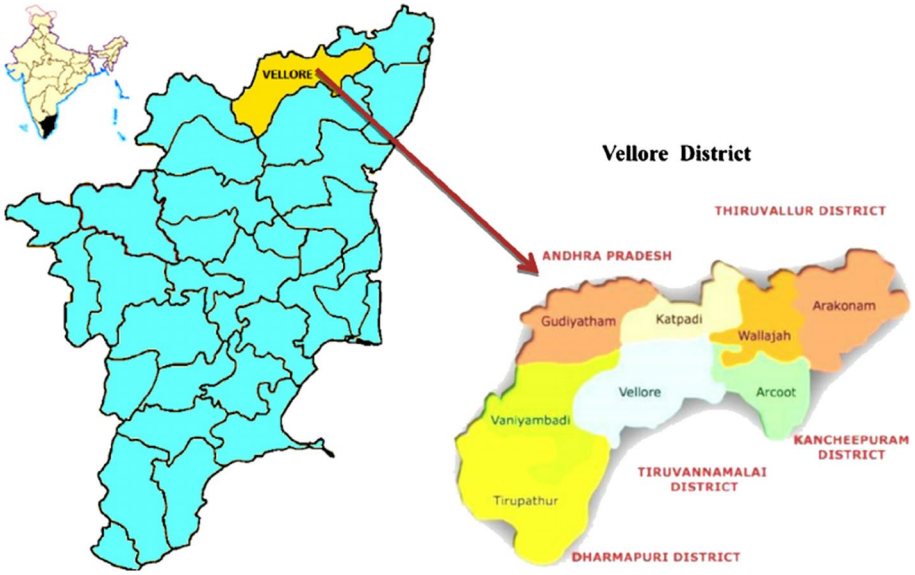



**Fig. 1** Geographical map of Vellore

38 districts of TN situated on the bank of the Palar River in the northern region. This river is the major water source that flows towards the east spanning a distance of about 295 km. The river flows in parallel ranges running to the hill ranges of Eastern Ghats. Even though the land consists of a massive flood plain, there exists a dry climate in the foremost part of the year. Vellore is well known for its arid climatic condition and during summer its reaches the highest temperature. The climatic variation for the whole year ranging from January to December is illustrated in Fig. 2. This district corresponds to a tropical savanna

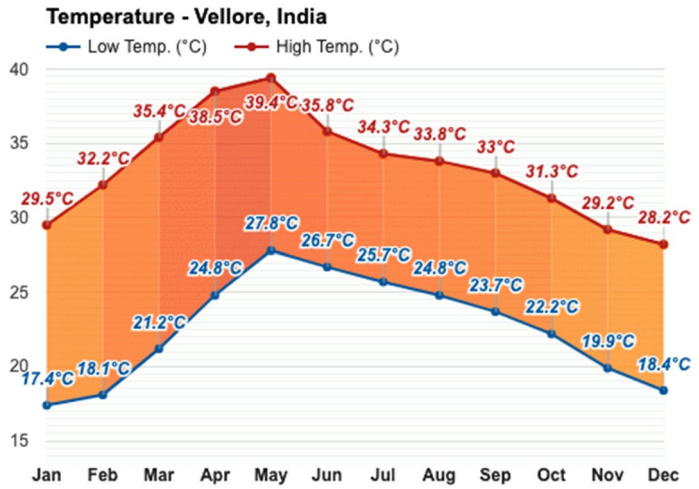



**Fig. 2** Climatic graph of Vellore district

climate where the months from April-June are the warmest, and December–January is the minor ones. It is the second-most populous district and the higher drought-hit district in TN. Urbanization has been increasing progressively due to the growth factor of industrialization, migration of population causes a great impact on the rise of temperature. The average level of temperature varies for the summer and winter seasons are 38.3 °C and 18.95 °C. During summer, and winter, the humidity ranges from 40 to 63%, and 67 to 87% respectively. Further, the southwest (June–September) and the northeastern monsoon (October-December) receive rainfall of 517.1 mm and 388.4 mm respectively.

### 3.2 Database collection

Satellite images acquired from Landsat 4–5 TM C1 Level-1, Landsat 7 ETM + C1 Level-1, and Landsat 8 OLI/TIRS C1 Level-1 are utilized for the assessment of LULC and the LST analysis. The overall details of the remote sensing data acquired from the study region are shown in Table 2. Landsat 4–5 consists of seven bands in which two SWIR, three visible, one thermal, and one NIR bands. The spatial resolution for the first five bands and the seventh band is 30 m and the sixth band corresponds to 120 m. Landsat 8 comprises two TIR bands (10&11) for the assessment of LST. Landsat 7 consists of ETM + images with nine spectral bands with a resolution of 30, 15, and 100 m. These imageries are collected from the U.S Geological Survey (USGS) GloVis website for the period of 1994, 2003, and 2018 respectively.

### 3.3 The steps for LULC change classification and LST estimation is presented as follows;

#### 3.3.1 A. preprocessing

Satellite images (Landsat TM/ETM + , and Landsat OTI) are collected from the dataset and perform preprocessing in terms of radiometric calibration, geometric corrections. The radiometric and atmospheric corrections are executed by the dark-object subtraction method and the geometric corrections by the google earth images.

Radiometric calibration: This process is performed by choosing appropriate ground control points that reduce the various types of noises and distortions.

Geometric correction**:** This correction is applied to eliminate the geometric distortion created by the sensor. Due to the differences between OLI and TM images, the correction was done based on Digital Elevation Model (DEM) data from the USGS portal of the specific area.

#### 3.3.2 B. LU/LC mapping

In LU/LU variation estimation for the three different years 1994, 2003, and 2018, a supervised classification algorithm using MLC is applied that obtains the changes on LST in terms of six different classes. To obtain different classes, several signatures are collected from three sets of images and merged to obtain the target class.

The result obtained from this MLC decision model is noisy due to the class similarity that gives misclassification outcomes and confusion arises from the harvested land, forest area, scrub forest, settlement area, water bodies, and more. To reduce the misclassification, some surface indices and accuracy parameters are computed.

**Table 2** Information of Landsat imageries used in this study

| Date of acquisition | Scene identifier | Sensor/satellite/ Thermal band | Path/row | Spectral bands resolution (m) | TIR band resolution (m) | Cloud cover (%) |
|---|---|---|---|---|---|---|
| 12/1/1994 | LT51430511994012ISP00 | Landsat 4–5 TM (Band 6) | 143/51 | 30 | 60 | < 10% |
| 29/1/2003 | LE71430512003029SGS00 | Landsat 7 ETM+ (Band 6) | 143/51 | 30 | 60 | < 10% |
| 30/1/2018 | LC81430512018030LGN00 | Landsat 8 OLI (Band 10) | 143/51 | 30 | 60 | < 10% |

The post-classification correction technique is used to reduce the errors after its initial MLC process using ancillary information (DEM, texture analysis, vegetation index (NDVI measure). This correction is performed in four steps as follows;

**Step-1**: For the forest area misclassification, reclassify the function in ArcGIS 9.3 that removes the forest class pixels from the maps obtained at the interval between 1994 and 2003. This is performed by generating the forest LULC patches by using 3 × 3variance texture images obtained from the NIR band of 1994 and 2018 individually and integrated with respective visual band data. Based on the developed imagery, the classification is performed for four LULC classes.

**Step-2:** Estimation of NDVI.

To analyze the vegetation cover from RS images, the NDVI measure is used. This measure is calculated based on the Red and NIR bands of Landsat 5 respectively (Yuan & Bauer, 2007). It is computed as follows;

$$\mathrm{NDVI} = \frac{(\mu_{\mathrm{NIR}} - \mu_{\mathrm{Red}})}{\mu_{RMNIR} + \mu_{\mathrm{Red}}} \tag{1}$$

The reclassification is performed with the threshold value with NDVI less than -0.5 is estimated as water body, else with the between -0.5–0.15 are termed as green land.

**Step-3:** Estimation of NDWI.

For the water body analysis and classification, NDWI is computed based on NIR and green band of RS imageries (Das and, Angadi, 2020).

$$\mathrm{NDWI} = frac(\eta_{\mathrm{green}} - \eta_{\mathrm{NIR}})(\eta_{\mathrm{green}} + \eta_{\mathrm{NIR}}) \tag{2}$$

The NDWI measure greater than zero is termed as water surface and the fewer values or equal to represents the non-water bodies. Here, the threshold measure is estimated based on the rule classifier by iteration variation in ENVI software.

**Step 4:** Estimation of NDBI.

To discriminate the different surface covers like water, barren region, built-up, farmland, etc., the Normalized Difference Built-up Index (NDBI) is adopted (Das and, Angadi, 2020). This measure separates the built-up area from other surface covers computed as follows;

$$\mathrm{NDBI} = frac\mathrm{SWIR} - \mathrm{NIRSWIR} + \mathrm{NIR} \tag{3}$$

In the NDBI measure, the output measure lies between -1 to + 1. Here, the higher value denotes the build-up areas and the negative values are the water surfaces.

### 3.3.3 C. Accuracy assessment

For the evaluation of the accuracy of each LULC map, a confusion matrix is employed to determine how well the data is classified or misclassified. Based on the confusion matrix, the quantitative measurement of accuracy parameters such as overall accuracy, producer, user's accuracy, and kappa statistics are computed for each year. Two widely used accuracy evaluation measures,

a) Kappa coefficient ($K_C$) estimates the probability of classified pixels compared to the ground truth ranges between 1 (perfectly accurate) else 0 (Choudhury et al., 2019).

$$\text{Kappa} = \frac{p(a) - p(c)}{1 - p(c)} \tag{4}$$

Where,
p(a)- observed accuracy.
p(c)- change agreement

$$K_C = \frac{(\text{Total sum of correct}) - \text{sum of all the (row column total)}}{\text{Total squared} - \text{sum of all the (row column total)}} \quad \text{(Das and Angadi, 2020)} \tag{5}$$

b) Overall accuracy ($O_A$) gives the result of how the individual pixels are classified against the total number of LC sample points (Pramit et al., 2020).

$$O_A = \frac{\sum C_{PP}}{N} \tag{6}$$

where, $C_{PP}$ is the total number of correctly classified pixels, and N is the total samples in the confusion matrix.

c) User accuracy or commission error is computed by dividing the matched number of pixels in individual land-use class by the total pixels in that category (Das & Angadi, 2020).

$$\text{User's accuracy}(\%) = \frac{\text{correctly classified pixels (each class)}}{\text{total number of pixels in that class}} \times 100 \tag{7}$$

d) Producer's accuracy or omission error is computed by dividing the matched number of matched observed zones by the total number of sites from sample data (Das & Angadi, 2020).

$$\text{Producers accuracy}(\%) = \frac{\text{Number of correctly classified pixels}}{\text{Total number of test pixel of eah class}} \times 100 \tag{8}$$

For each LU/LC accuracy estimation, producers' and users' accuracy are utilized. Producer's accuracy/Error of Omission specifies the correctly classified percentage of a specific class on the ground whereas the User's accuracy/Error of commission percentage of a classified pixel on the same category. The flow chart for the LULC prediction and LST analysis is depicted in Fig. 3.

### 3.3.4 D. LST Change detection

After LULC map preparation, the change detection is carried out by "from-to" change information that determines the changes of LULC between different time intervals. Here, a cross-tabulation analysis has conducted that estimates the change of a particular category to the category based on the pixel-to-pixel basis. The changes detection for three different categories like 1994–2003, 2003–2018, and 1994–2018 of obtained LULC is utilized that tabulate the matrix to find the improvement and damages occurred in each category of the study area.

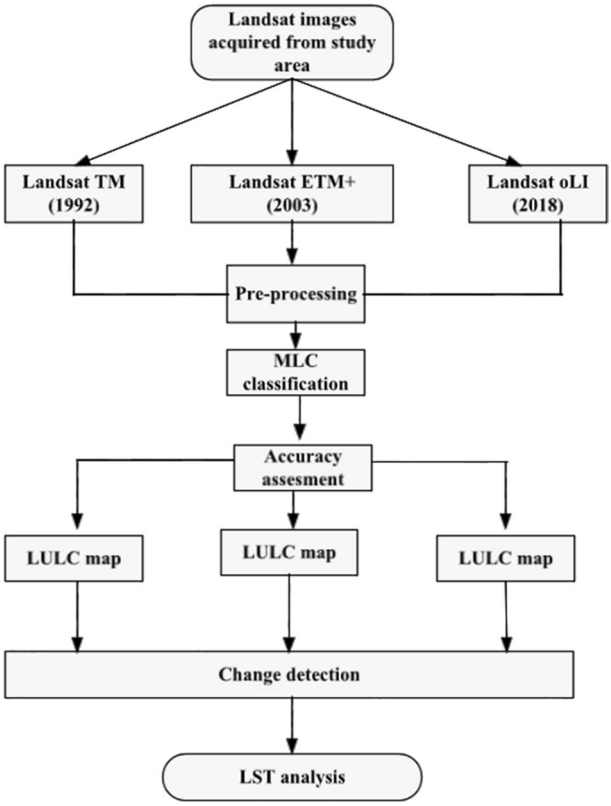



**Fig. 3** Flowchart of the proposed methodology

### 3.3.5 E. LST analysis

The thermal data of the Landsat sensors are acquired for the study area are stored in Digital Number (DN). For the extraction of LST from Landsat data, different steps are adopted.

### 3.3.6 (i) Spectral Radiance (SR) conversion from DN Landsat images

The LST is computed from geometrically improved Landsat images for the years 1994, 2003, and 2018. Firstly, the pixel-oriented digital number (DN) is changed into the radiance ($L_{\lambda}$) as follows. For LST estimation, the following set of equations is used (Thakur et al., 2021).

$$L_{\lambda} = \left(\frac{L_{MAX\lambda} - L_{MIN\lambda}}{255} DN\right) + L_{MIN\lambda} \tag{9}$$

Here, $L_{MAX\lambda}$ *and* $L_{MIN\lambda}$ represents the maximum and minimum radiance measure acquired from the metadata file.

For Landsat 8, the raw thermal band data is converted into spectral radiance using the below equation

$$L_\lambda = L_{MIN}\left(\frac{L_{MAX} - L_{MIN}}{QCAL_{MAX} - QCAL_{MIN}}\right)DN \quad \text{(Tariq et al., 2019)} \tag{10}$$

$$L_\lambda = M_L \times Q_{CAL} + A_L \quad \text{(Tariq et al., 2019} \tag{11}$$

where $QCAL_{MAX}$ is the maximum DN, $QCAL_{MIN}$ is the minimum DN value, $Q_{CAL}$ is the DN of each image, $M_L$, $A_L$ are the band-specific multiplicative and additive rescaling factors from the MTL file.

#### 3.3.7 (ii) Conversion of SR to brightness temperature

The brightness temperature ($T_S$) is calculated by the following equation;

$$T_S = \frac{C_2}{\ln\left(\frac{C_1}{L_\lambda} + 1\right)} - 273.15 \quad \text{(Tariq et al., 2019)} \tag{12}$$

where $C_1$, $C_2$ represents the constant measure obtained from USGS.

#### 3.3.8 (iii) Estimation of Land surfaces indices

The vegetation indices NDVI, NDWI, and NDBI are estimated based on the Eq. (1), (2), and (3) mentioned in the LULC mapping process above. Apart from these measures, NDMI is computed based on the following Eq. (13). This measure estimates the moisture content of the area. Here NIR (band 5 of Landsat 8), SWIR (band 6 of Landsat 8) (Das & Angadi, 2020).

$$\text{NDMI} = \frac{\text{NIR} - \text{MIR}}{\text{NIR} + \text{MIR}} \tag{13}$$

#### 3.3.9 (iv) Proportion of vegetation

($P_V$)is estimated using the maximized, and minimized NDVI values (Tariq et al.,2019)

$$P_V = \left(\frac{\text{NDVI} - \text{NDVI}_{\text{MIN}}}{\text{NDVI}_{\text{MAX}} - \text{NDVI}_{\text{MIN}}}\right)^2 \tag{14}$$

Where $NDVI_{MAX}$*and* $NDVI_{MIN}$ are the maximized and minimized NDVI values.

#### 3.3.10 (v) Derivation of land surface emissivity ($\xi$)

For the land surface temperature estimation, the knowledge of emissivity measure is vital which is calculated based on the NDVI measure (Tariq et al., 2019).

$$\xi = 0.004 \times P_v + 0.986 \tag{15}$$

### 3.3.11 (vi) LST analysis in °C

Estimated based on emissivity and brightness temperature measure and then the values are normalized to compare the data with different years. Then, the LST relationship between the spectral indices is determined by the Pearson correlation and regression analysis method.

$$\mathrm{LST} = fracT_S\left[1 + \left(\frac{\eta T_S \sigma}{hc} \ln \xi\right)\right] \quad \text{(Ullah et al., 2019)} \tag{16}$$

where, $\eta$ is the center of the wavelength, c is the velocity of light ($2.998 \times 10^8 m/s$), and $\sigma$ is the Boltzmann constant ($1.38 \times 10^{-23} J/K$), h is the Plank's constant ($6.626 \times 10^{-34} J/s$).

### 3.3.12 (vii) Hotspot analysis for LST

The study area is examined and classified into cold and hot spot areas based on Getis-Ord Gi* local statistics model in the ArcGIS software. The separation of these two areas is determined based on a comparison between each feature of LST with its neighboring measure. After comparison, the feature with minimum values is grouped into cold spots while the maximum value is termed into hot spots.

The overall outline of the LST estimation for the study region is illustrated in Fig. 4.

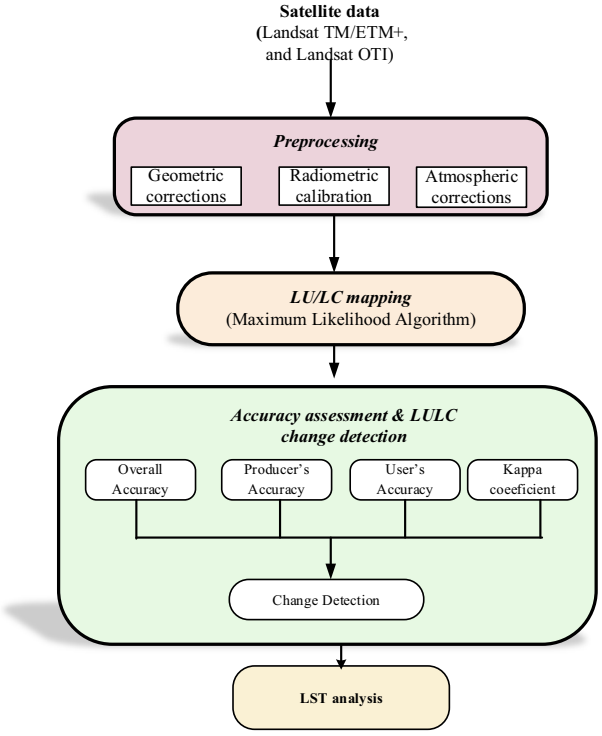



**Fig. 4** Outline of the LST analysis model

## 4 Findings and Discussion

In this study, we perform the simulation on a python tool platform conducted using multi-spectral remote sensing images collected from the USGS GloVis website in the region of Vellore district, Tamil Nadu, India during 1994, 2003, and 2018 respectively. For the map preparation, we adopt ArcGIS 10.1 application. This study also examines the LULC changes for years (1994–2018), confusion matrices, accuracy assessments, the correlation between LULC&LST, NDVI & LST.

### 4.1 LULC classification and accuracy assessment

Figure 5 represents the classification result of LULC variations for four different categories during three different years of the 1994, 2003, and 2018 time periods. It is observed that the built-up area (red areas) get increased drastically in the urban parts of the Vellore district while the vegetative land decreased due to the population growth, and urbanization in these areas.

Table 3 shows the vegetative land declined by -635.91 (-65%), built-up land expanded by 802.49 (81%) in the study period. Also, the amount of barren land improved by 2862.476 (290%), and the water bodies get reduced by -436.55% (-44%) from the year 1994–2018. LULC variations for three different time period is depicted in the bar chart as shown in Fig. 6. In this study period, the vegetative, and water land has reduced but the built-up, and barren land have increased significantly between the years 1994–2018.

The overall accuracy for LULC during 1994, 2003, and 2018 time period are 92.38%, 89.17%, and 91.34% respectively as illustrated in Table 4. Also, the result of kappa statistics, user accuracy, and producer accuracy for LULC maps by using Landsat satellite imageries are determined. The kappa value greater than 0.75 classified results is termed as compatible.

### 4.2 Result of vegetative indices

#### 4.2.1 NDVI

NDVI analysis is employed for the detection of green areas in the study region as well as the changes in vegetation cover for the three different years. From Fig. 7, dense vegetative surfaces are highlighted in a dark green area and the surface of the non-vegetated area is indicated with a negative value (light green). NDVI values achieved with lower than 0.22 are represented as the sparsely vegetated land. The NDVI values are in the ranges of 0.44–1.0, 0.32–0.44, 0.22–0.32, 0.12–0.22, and -0.81–0.12 respectively represented in Fig. 7. The lower vegetation land is observed in the urban cities of the Vellore district over the study period.

The greater change in NDVI measure has a great impact on the changes in LST where the lowering values of vegetation cover indicate the higher LST value.

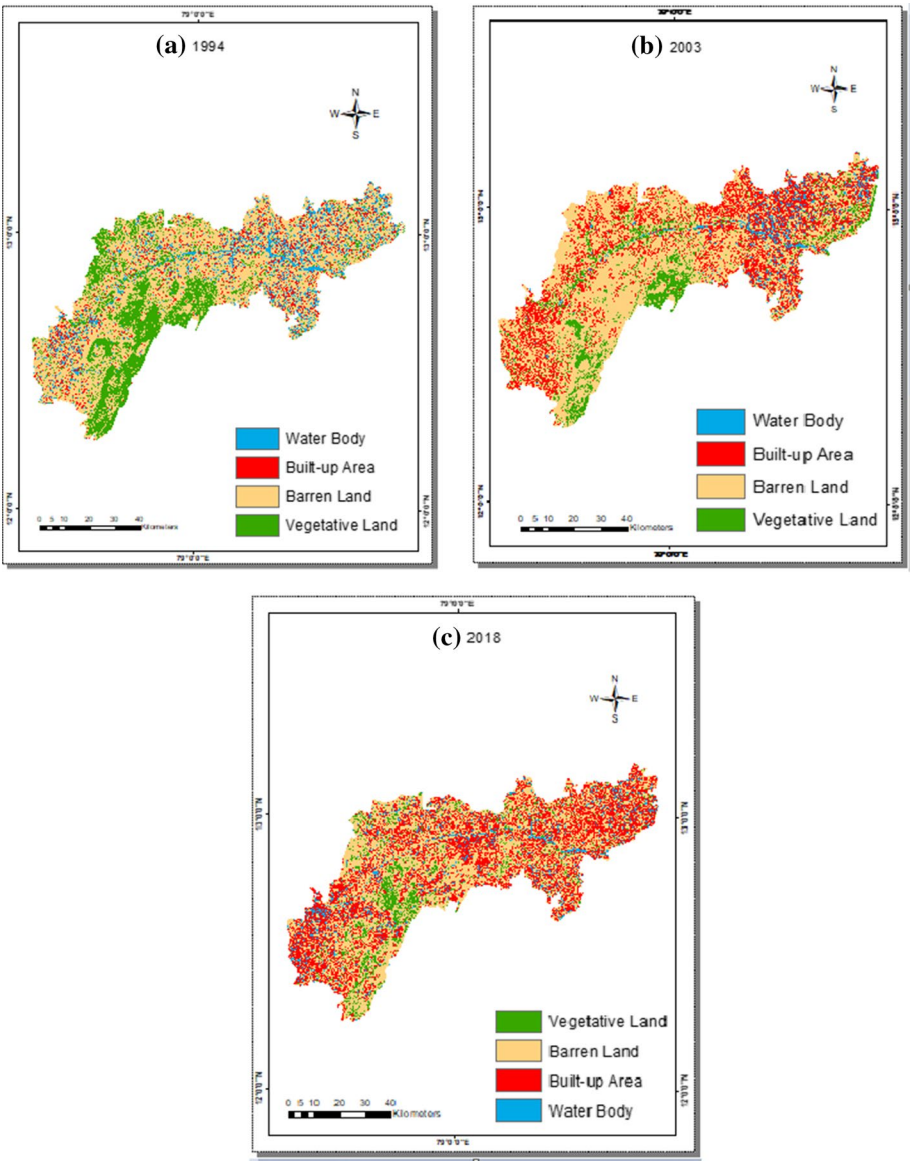



**Fig. 5** LULC maps estimated from Landsat data for years **a** 1994, **b** 2003 and **c** 2018

### 4.2.2 NDWI

To estimate the presence of water bodies, the NDWI measure is adopted over the period 1994–2018. NDWI measure ranges from the value of -1 to + 1, where the positive value indicates the higher density of the water surface. From the analysis of this study as shown in the Fig. 8, it is evident that there is a significant decrement in water bodies in the examined Vellore district. The rapid urbanization in the district replaces the

**Table 3** Area and % for each LULC classes for 1994 – 2018

| Land use types | Area in sq.km and Area in percentage | | Area in sq.km and Area in percentage | | Change in area | Change in percentage |
|---|---|---|---|---|---|---|
| | 1994 | 1994 | 2018 | 2018 | 1994 – 2018 | 1994 – 2018 |
| Barren Land | 4574.74 | 24% | 4304.54 | 72% | 2862.476 | 290% |
| Built-up Land | 303.96 | 5% | 1106.45 | 18% | 802.49 | 81% |
| Vegetative Land | 1138.11 | 19% | 501.2 | 8% | – 635.91 | – 65% |
| Water Body | 541.07 | 9% | 104.62 | 2% | – 436.55 | – 44% |

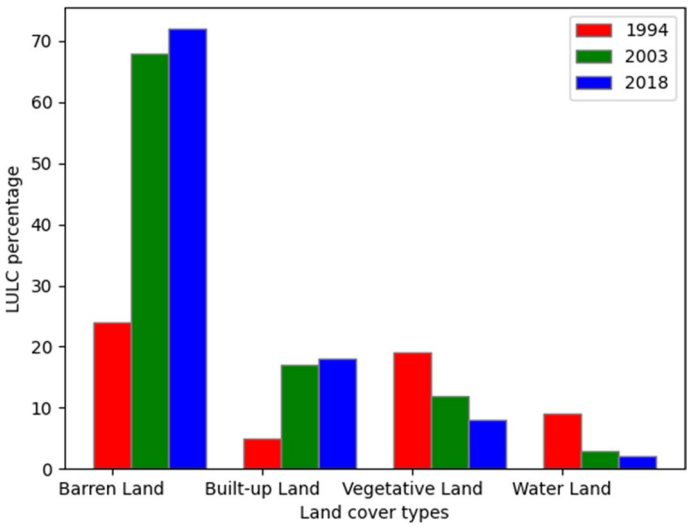



**Fig. 6** LULC change analysis for three different years

**Table 4** Assessment of accuracy for LULC for 1994, 2003, and 2018

| Year | User accuracy (%) | Producer Accuracy (%) | Overall accuracy (%) | Kappa coefficient |
|---|---|---|---|---|
| 1994 | 93.17 | 94.84 | 92.38 | 0.92 |
| 2003 | 91.23 | 95.05 | 89.17 | 0.85 |
| 2018 | 90.0 | 92.55 | 91.34 | 0.81 |

vegetative practices and reduces NDWI measures. The lower value of NDWI value lowers the air humidity and greatly increases the LST values in the study area.

### 4.2.3 NDBI

The estimation of built-up area is illustrated by the NDBI measure that ranges from -1 to + 1. The NDBI maps with the positive value denote the high density of the built-up

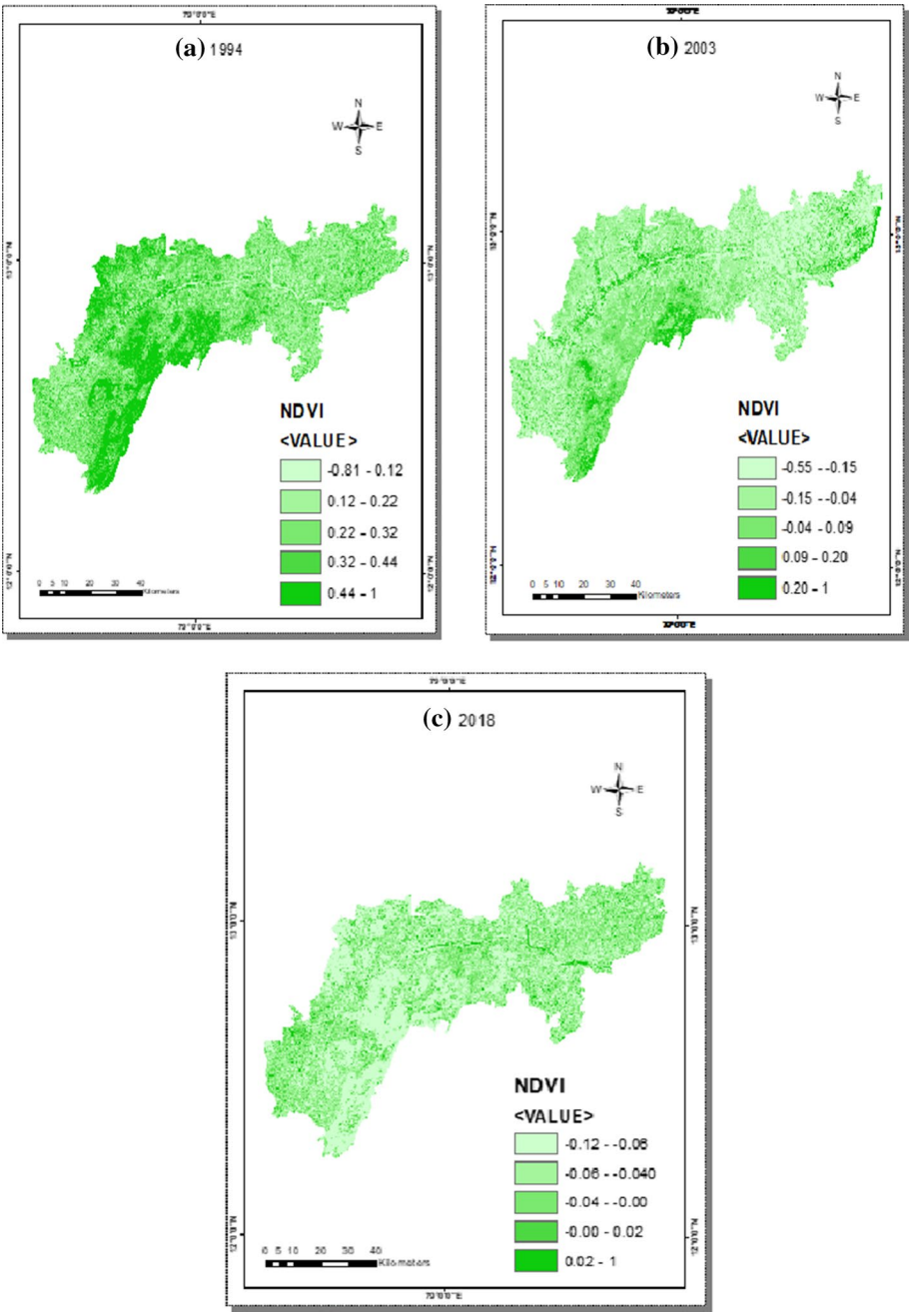



**Fig. 7** NDVI measure for the study area in terms of **a** 1994, **b** 2003, **c** 2018

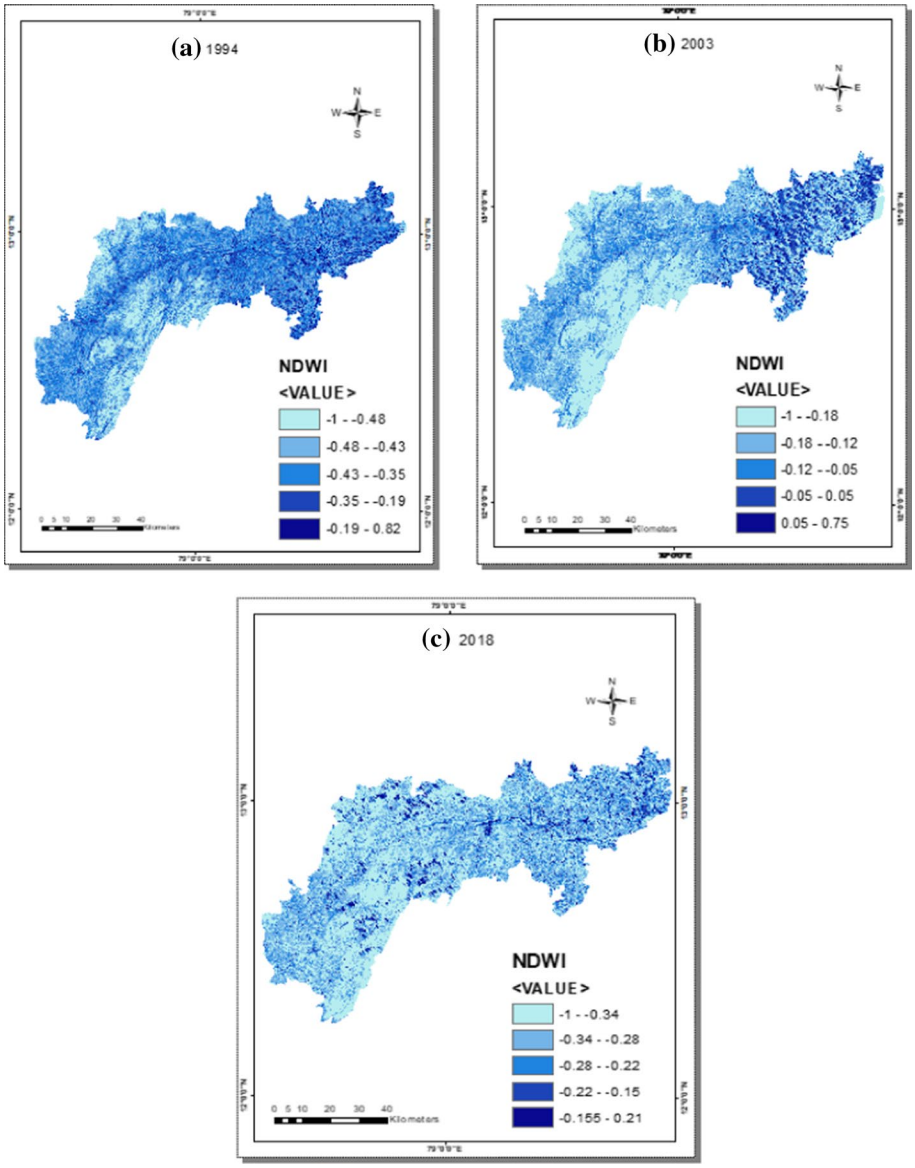



**Fig. 8** NDWI maps during the study region of **a** 1994, **b** 2003, and **c** 2018

region. From Fig. 9, it is clear that there is a higher reflectance of urban distribution increased over the study period. The rapid increase of NDBI is due to the industrial development around the district.

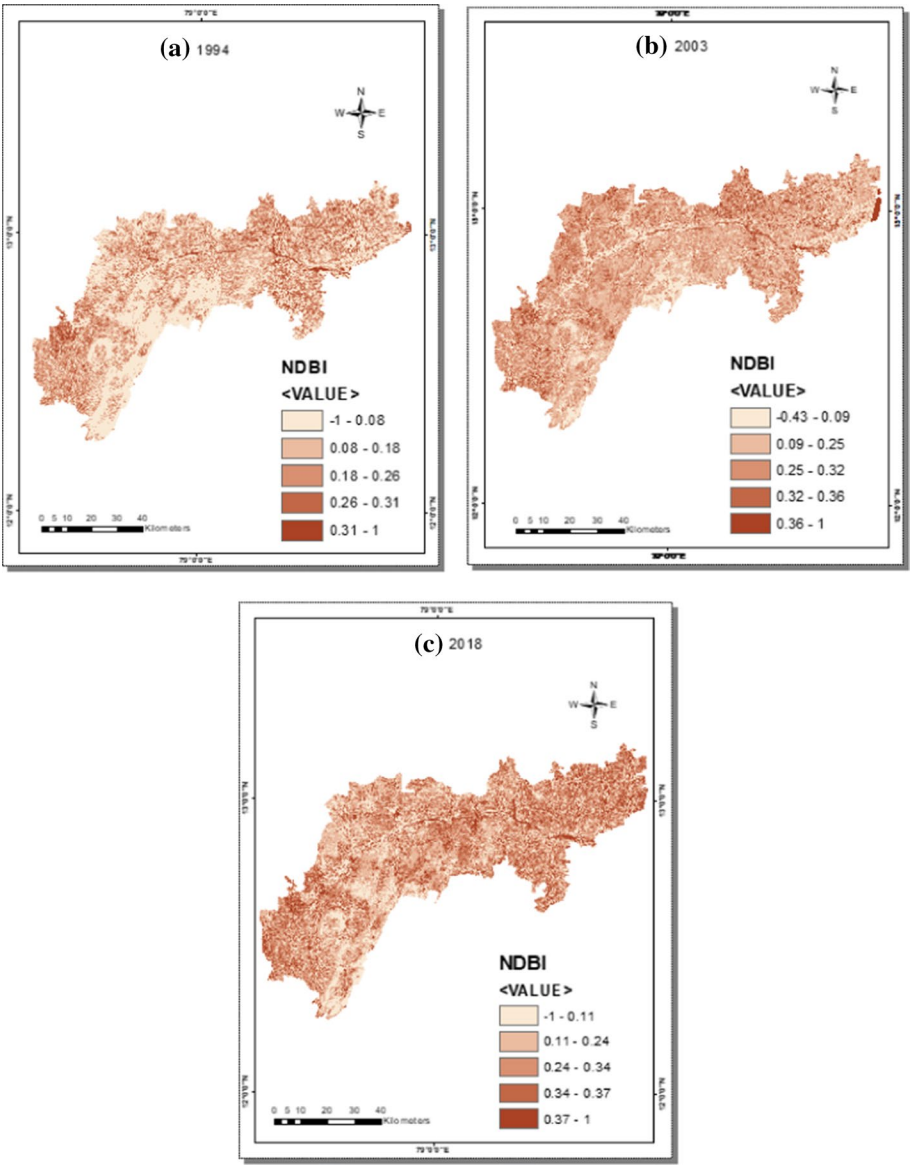



**Fig. 9** Spatial distribution of NDBI values for three different years **a** 1994, **b** 2003, and **c** 2018

### 4.3 LST change analysis

From Fig. 10, it is evident that the spatial changes in LST of 1994, 2003, and 2018 have increased in the Vellore district. The LULC changes greatly affect the variation in LST. This transformation between different LULC classes is mainly due to the urban expansion that further enhances the hot spot distribution in the area. The rapid increase of

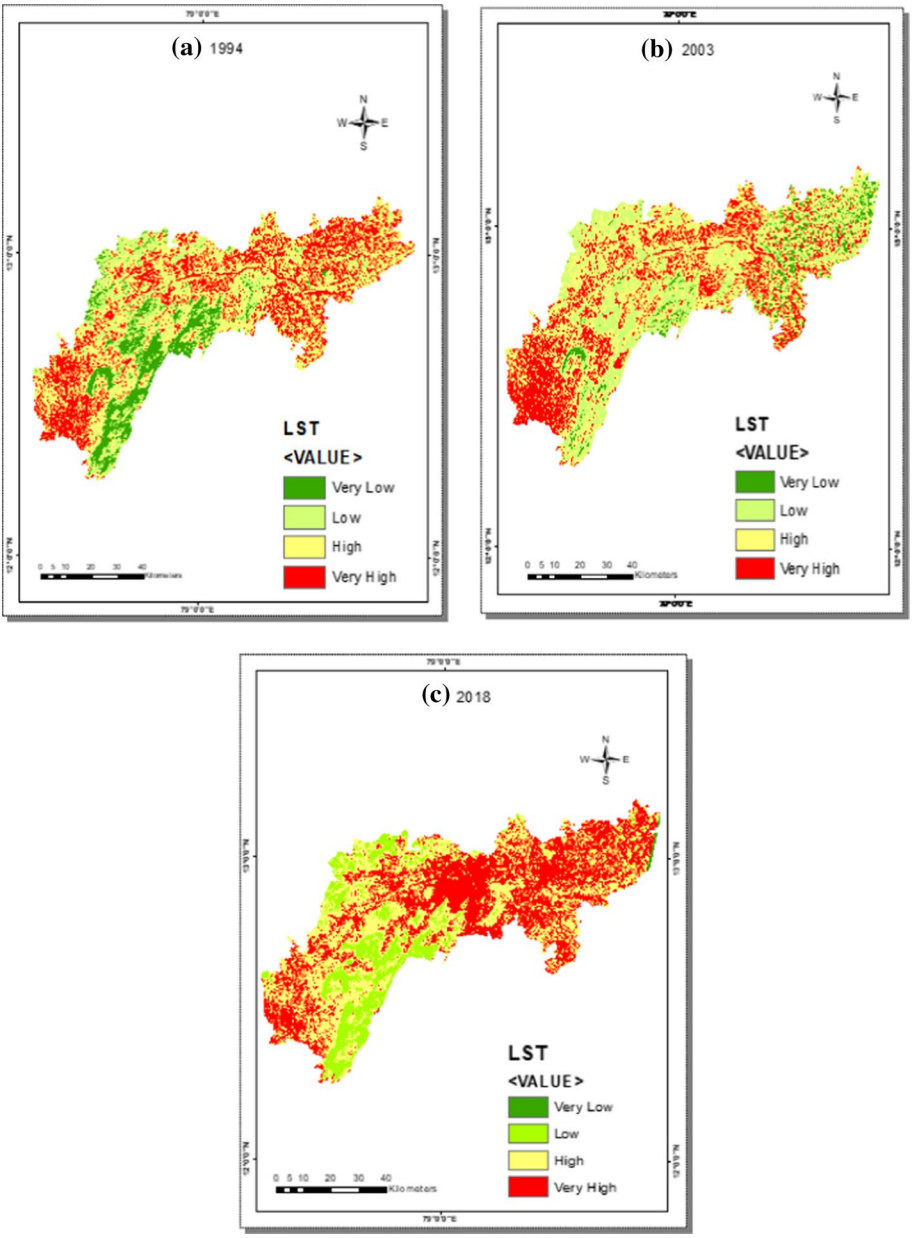



**Fig. 10** LST map (in Celsius) of Vellore during **a** 1994, **b** 2003, and **c** 2018

urban built-up is due to the larger concentration of industrial areas which results in the highest average temperature.

As mentioned in the literature (Raghavan et al., 2015) the temperature distribution patterns over the urban regions in Vellore district range from 38 ° C to 41 ° C. Figure 11 shows the Mean LST value for the four different LULC classes during 1994–2018. The

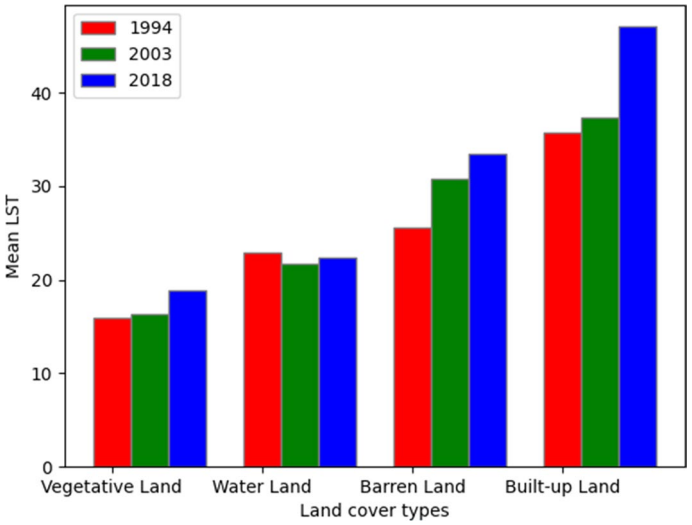



**Fig. 11** Mean LST statistics

**Table 5** Mean LST statistics for the Vellore district during 1994, 2003, and 2018

| Land use types | 1994 | | 2003 | | 2018 | | Change during 2003 – 1994 | |
|---|---|---|---|---|---|---|---|---|
| | Mean LST | Area | Mean LST | Area | Mean LST | Area | LST | % of Area |
| Vegetative land | 15.99 | 1138.11 | 16.40 | 702.56 | 18.95 | 501.2 | 0.41 | – 11% |
| Water body | 22.84 | 541.07 | 21.70 | 206.01 | 22.32 | 104.62 | 1.14 | – 8% |
| Barren land | 25.52 | 4574.74 | 30.8 | 4112.86 | 33.41 | 4304.54 | 5.28 | 65% |
| Built-up land | 35.7 | 303.96 | 37.38 | 995.38 | 47.08 | 1106.45 | 1.68 | 17% |

**Table 6** Mean LST for 2018 – 2003

| Change during 2018 – 2003 | |
|---|---|
| LST | % of area |
| 2.55 | – 33% |
| 0.62 | – 17% |
| 2.61 | 32% |
| 9.7 | 18% |

findings from the overall analysis show that the mean LST for vegetative, barren, and built-up land get expanded from 15 ° C to 18 ° C, 25 ° C to 33 ° C, and 33 ° C to 48 ° C respectively. On the other hand, the mean LST of the water body gets lowered from 23 ° C to 22 ° C during this period.

The mean LST values are computed individual years for the three different years with four different classes are summarized in Tables.5, and 6.

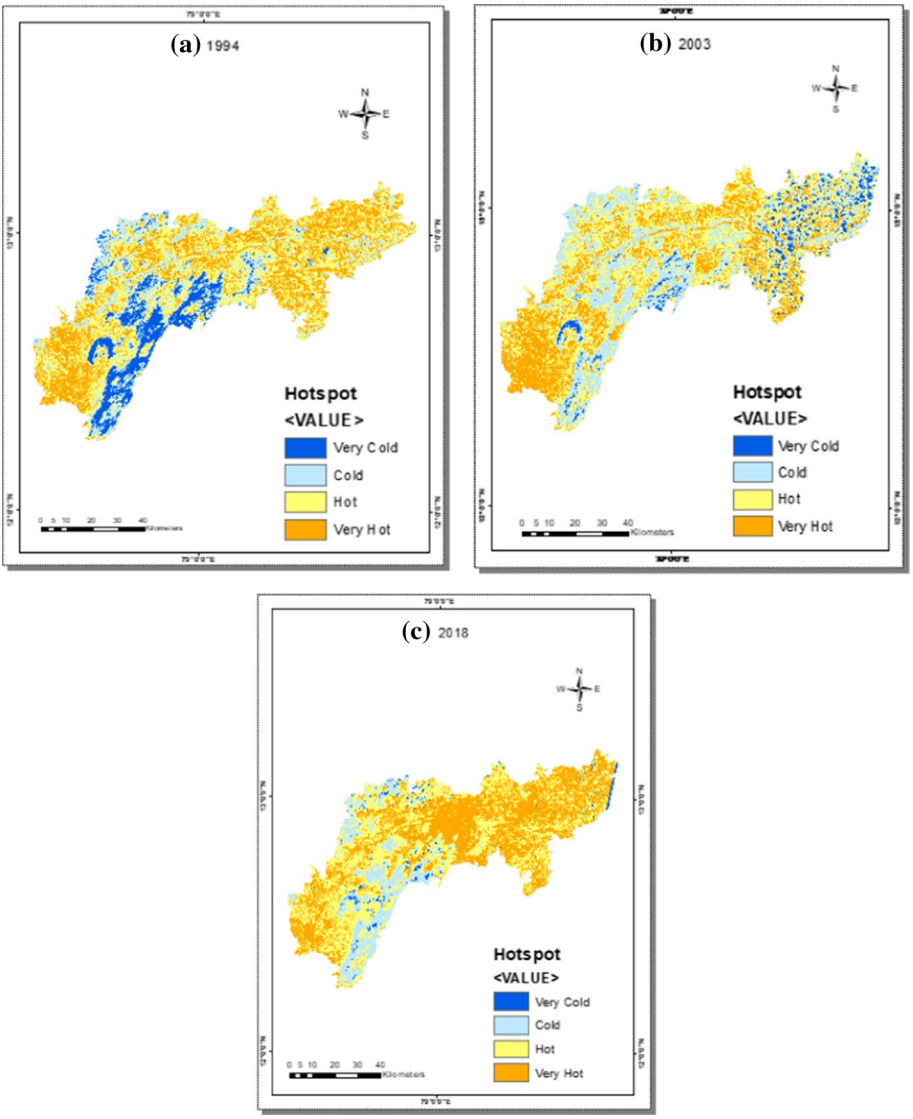



**Fig. 12** Hotspot zones for three different years **a** 1994, **b** 2003, and **c** 2018

### 4.4 Hotspot analysis

From the hotspot analysis of the examined region which is depicted in the Fig. 12, shows that the hotspot zones are increased from minimum to a large number of areas over the district. The total cold spot was decreased by a great number that represents the non-stop urbanization in those specific areas during the period 1994–2018.

### 4.5 Discussion

This section presents the deep analysis of research findings which lies in the study of LULC changes for years (1994–2018), confusion matrices, accuracy assessments, the correlation between LULC and LST, NDVI, and LST. The literature of (Ghosh and, Porchelvan, 2017) reported that the total build-up area get increased rapidly, especially after 1977. Similarly in this study, the built-up land get increased drastically as shown in Fig. 5c. As from the literature (Ghosh & Porchelvan, 2017), the land use changes for the water bodies, agricultural land, forest cover areas get reduced while built-up/urban, cropland get increased. Similarly, the findings from this study achieve the highest rise in barren, and built-up land while the vegetative, and water bodies get reduced, as shown in Table 3. The overall accuracy for LULC during 1994, 2003, and 2018 time period are 92.38%, 89.17%, and 91.34% respectively as illustrated in Table 4. Also, the kappa value greater than 0.75 classified results is termed as compatible. For NDVI analysis, the literature (Prabu & Anouncia, 2018), shows that the agricultural areas are changed for residential purpose. In this study, we have achieved the lower vegetation area due to the arrival of larger industries around the Vellore district over the study period, shown in Fig. 7. Then for NDWI analysis, in literature (Nagamani and Manoharan), the study area experiences a rapid decrease in water bodies and ground water table. The lower value of NDWI value lowers the air humidity and greatly increases the LST values in the study area, shown in Fig. 8. In NDBI analysis, the literature by (Ghosh J, Porchelvan, 2017) stated that the urban/built-up land get increased due to the direct impact of urbanization. Likewise in this study, higher reflectance of urban distribution increased over the study period, shown in Fig. 9. In the earlier literature (Raghavan et al., 2015), the LST calculation recorded for the Vellore district with a minimum temperature of 30 $^{o}$ C and maximum temperature of 46 $^{o}$ C nearby the urban regions of Vellore taluk. Similarly in this study spatial changes in LST of 1994, 2003, and 2018 have been increased in the Vellore district, shown in Fig. 10. The percentage changes of area correlated with LST measure between 2003–1994, 2018–2003 is examined that tends to have a positive impact on barren, and built-up classes while negative value on the water body, and vegetative land, in Table 5. In the literature (Raghavan et al., 2015), the highest temperature is obtained near the urban regions of Vellore district due to the effect of urbanization. For hot spot analysis, the hotspot zones are increased from minimum to a large number of areas during the period 1994–2018.

## 5 Conclusion

In this research, we have investigated the variations of LULC and LST using the sensing images acquired from the USGS website. From this analysis, it is evident that the built-up areas get increased for the time interval from 1994–2018. These changes are directly linked to an increase in LST that greatly impacts a better life quality. ***Effects of vegetation indices-*** The vegetation indices such as NDVI indicate the density of a green area and estimate the crop production whereas the NDBI highlights the urban distribution, NDWI measures extent of water bodies. Hence, it is concluded, NDBI has an inverse relation with NDVI and NDWI where the higher percentage of NDBI directly lowers the NDVI, and NDWI and vice versa. The clear picture of the increase in LST alerts the environmentalist about the future consequences over the entire land cover region. ***Simulation result-*** Various land

cover indices and their relationship with LST are estimated which indicates that the annual change rate surges to a noticeable change in percentage. The simulation results show that the vegetative land has decreased from 19% to -65% in the year 1994–2018. Barren land surges from 24 to 290% show the majority of the agricultural land has been converted toward the built-up area. The examination of hotspot zones shows that they increased particularly in the big cities during the entire study period. These changes are occurred due to the result of population growth and increases in migration towards the urban cities. The result shows the transformation of vegetated land into the built-up area rises LST which is caused due by the higher rate of the urban population. ***Effects of LST-*** From the result, most of the regions are detected with the hot spot areas as depicted in Fig. 10c and negative impact on % of the area mentioned in four classes in Table 5. The factors such as heat waves, population rate increase, and emission of high CO2 cause climatic variability that negatively impacts the crop yield and food shortage in agriculture. Also, climate change leads to the rise of sea level, more droughts, and an increase in infectious diseases. ***Limitations-*** This current study has certain limitations: a) the climatic variations due to the greenhouse emission are ignored, the four LULC classes limit the estimation of the percentage of change area and classification accuracy, the impact of temperature over mountain rich regions are not well explored. ***Future scope—***Based on the achieved result, it is revealed that there is a need to change urban policies related to heavy industrial and commercial sectors that have a great impact on surface temperature. The findings from the study can be used as a reference by the land use planner for the management of the district in the coming future. Future research will work towards highlighting the changes occurs in other different classes like grasslands, open-land, settlement areas, etc.

**Acknowledgements** Not applicable.

**Authors contributions** All the authors have participated in writing the manuscript and have revised the final version. All authors read and approved the final manuscript.

**Funding** There is no funding for this study.

## Declarations

**Conflict of interest** Authors declare that they have no conflict of interest.

**Human and animal participants** This article does not contain any studies with human participants and/or. animals performed by any of the authors.

**Consent to participate** There is no informed consent for this study.

**Consent for publication** Not applicable.

## References

Balogun, I., & Ishola, K. (2017). Projection of future changes in landuse/landcover using cellular automata/markov model over Akure city, Nigeria. *Journal of Remote Sensing Technologies, 5*, 22–31.

Choudhury, D., Das, K., & Das, A. (2019). Assessment of land use land cover changes and its impact on variations of land surface temperature in Asansol-Durgapur Development Region. *The Egyptian Journal of Remote Sensing and Space Science, 22(2),* 203–18 (2019).

Das, S., & Angadi, D. P. (2020). Land use-land cover (LULC) transformation and its relation with land surface temperature changes: A case study of Barrackpore Subdivision, West Bengal, India. *Remote Sensing Applications: Society and Environment, 19*, 100322.

Fonseka, H. P., Zhang, H., Sun, Y., Su, H., Lin, H., & Lin, Y. (2019). Urbanization and its impacts on land surface temperature in Colombo metropolitan area, Sri Lanka, from 1988 to 2016. *Remote Sensing, 11*(8), 957.

Ghosh, J., & Porchelvan, P. (2017). Remote sensing and GIS technique enable to assess and predict landuse changes in Vellore district, Tamil Nadu, India. *IJAER, 12*(12), 3474–3482.

Gohain, K. J., Mohammad, P., & Goswami, A. (2021). Assessing the impact of land use land cover changes on land surface temperature over Pune City, India. *Quaternary International, 575*, 259–269.

Grigoraș, G., & Urițescu, B. (2019). Land use/land cover changes dynamics and their effects on surface urban heat island in Bucharest, Romania. *International Journal of Applied Earth Observation and Geoinformation, 80*, 115–126.

Hart, M. A., & Sailor, D. J. (2009). Quantifying the influence of land-use and surface characteristics on spatial variability in the urban heat island. *Theoretical and Applied Climatology, 95*(3), 397–406.

Huang, Q., & Lu, Y. (2018). Urban heat island research from 1991 to 2015: A bibliometric analysis. *Theoretical and Applied Climatology, 131*(3), 1055–1067.

John, J., Bindu, G., Srimuruganandam B, Wadhwa A, & Rajan P. (2020). Land use/land cover and land surface temperature analysis in Wayanad district, India, using satellite imagery. Annals of GIS, pp. 1–8.

Kafy, A. A., Rahman, M. S., Hasan, M. M., & Islam, M. (2020). Modelling future land use land cover changes and their impacts on land surface temperatures in Rajshahi, Bangladesh. *Remote Sensing Applications: Society and Environment, 18*, 100314.

Lilly Rose, A., & Devadas, M. D. (2009). Analysis of land surface temperature and land use/land cover types using remote sensing imagery-a case in Chennai City, India. In: *Proceedings of the 7th International Conference on Urban Climate (ICUC-7), Yokohama, Japan,* p. *29.*

Madanian, M., Soffianian, A. R., Koupai, S. S., Pourmanafi, S., & Momeni, M. (2018). The study of thermal pattern changes using Landsat-derived land surface temperature in the central part of Isfahan province. *Sustainable Cities and Society., 39*, 650–661.

Maimaitiyiming, M., Ghulam, A., Tiyip, T., Pla, F., Latorre-Carmona, P., Halik, Ü., Sawut, M., & Caetano, M. (2014). Effects of green space spatial pattern on land surface temperature: Implications for sustainable urban planning and climate change adaptation. *ISPRS Journal of Photogrammetry and Remote Sensing, 89*, 59–66.

Nagamani, K., & Manoharan, N. Multi-temporal land use/land cover change detection in semi urban vellore district using landsat tm and ETM+ DATA.

Nayak, S., & Mandal, M. (2019). Impact of land use and land cover changes on temperature trends over India. *Land Use Policy, 89*, 104238.

Pan, X. Z., Uchida, S., Liang, Y., Hirano, A., & Sun, B. (2010). Discriminating different landuse types by using multitemporal NDXI in a rice planting area. *International Journal of Remote Sensing, 31*(3), 585–596.

Piyoosh, A. K., & Ghosh, S. K. (2020). Analysis of land use land cover change using a new and existing spectral indices and its impact on normalized land surface temperature. *Geocarto International,* pp. 1–23.

Piyoosh, A. K., & Ghosh, S. K. (2018). Development of a modified bare soil and urban index for Landsat 8 satellite data. *Geocarto International, 33(4),* 423–42 (2018).

Piyoosh, A. K., & Ghosh, S. K. (2017). Semi-automatic mapping of anthropogenic impervious surfaces in an urban/suburban area using Landsat 8 satellite data. *Giscience and Remote Sensing, 54*(4), 471–494.

Prabu, M., & Anouncia, S. M. (2018). Prediction of Land Cover Changes in Vellore District of Tamil Nadu by Using Satellite Image Processing. *InKnowledge Computing and its Applications Springer, Singapore,* 87–100.

Pramit, V., Aditya, R., Srivastava, P. K., & Raghubanshi, A. S. (2020). Appraisal of kappa-based metrics and disagreement indices of accuracy assessment for parametric and nonparametric techniques used in LULC classification and change detection. *Modeling Earth Systems and Environment., 6*(2), 1045–1059.

Raghavan, K., Mandla, V., & Franco, S. (2015). Influence of urban areas on environment: Special reference to building materials and temperature anomalies using geospatial technology. *Sustainable Cities and Society., 219*, 349–358.

Shi, Y., & Zhang, Y. (2018). Remote sensing retrieval of urban land surface temperature in hot-humid region. *Urban Climate., 24*, 299–310.

Tariq, A., Riaz, I., Ahmad, Z., Yang, B., Amin, M., Kausar, R., Andleeb, S., Farooqi, M. A., & Rafiq, M. (2020). Land surface temperature relation with normalized satellite indices for the estimation of

spatio-temporal trends in temperature among various land use land cover classes of an arid Potohar region using Landsat data. *Environmental Earth Sciences., 79*(1), 1–5.

Thakur, S., Maity, D., Mondal, I., Basumatary, G., Ghosh, P. B., Das, P., & De, T. K. (2021). Assessment of changes in land use, land cover, and land surface temperature in the mangrove forest of Sundarbans, northeast coast of India. *Environment, Development and Sustainability., 23*(2), 1917–2439.

Ullah, S., Tahir, A. A., Akbar, T. A., Hassan, Q. K., Dewan, A., Khan, A. J., & Khan, M. (2019). Remote sensing-based quantification of the relationships between land use land cover changes and surface temperature over the Lower Himalayan Region. *Sustainability., 11*(19), 5492.

Uttara, S., Bhuvandas, N., & Aggarwal, V. (2012). Impacts of urbanization on environment. *International Journal of Research in Engineering and Applied Sciences, 2(2),* 1637–45 (2012).

Vohra, R., & Tiwari, K. C. (2020). Comparative Analysis of SVM and ANN Classifiers using Multilevel Fusion of Multi-Sensor Data in Urban Land Classification. *Sensing and Imaging, 21(1),* pp. 1–21.

Vohra, R., & Tiwari, K. C. (2019). Spatial shape feature descriptors in classification of engineered objects using high spatial resolution remote sensing data. *Evolving Systems*, pp. 1–4.

Wang, R., Cai, M., Ren, C., Bechtel, B., Xu, Y., & Ng, E. (2019). Detecting multi-temporal land cover change and land surface temperature in Pearl River Delta by adopting local climate zone. *Urban Climate, 28*, 100455.

Wang, S., Ma, Q., Ding, H., & Liang, H. (2018). Detection of urban expansion and land surface temperature change using multi-temporal landsat images. *Resources, Conservation and Recycling, 128*, 526–534.

Yuan, F., & Bauer, M. E. (2007). Comparison of impervious surface area and normalized difference vegetation index as indicators of surface urban heat island effects in Landsat imagery. *Remote Sensing of Environment., 106*(3), 375–386.

Zhang, Y., Odeh, I. O., & Han, C. (2009). Bi-temporal characterization of land surface temperature in relation to impervious surface area, NDVI and NDBI, using a sub-pixel image analysis. *International Journal of Applied Earth Observation and Geoinformation, 11*(4), 256–264.

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

**PAPER** • **OPEN ACCESS**

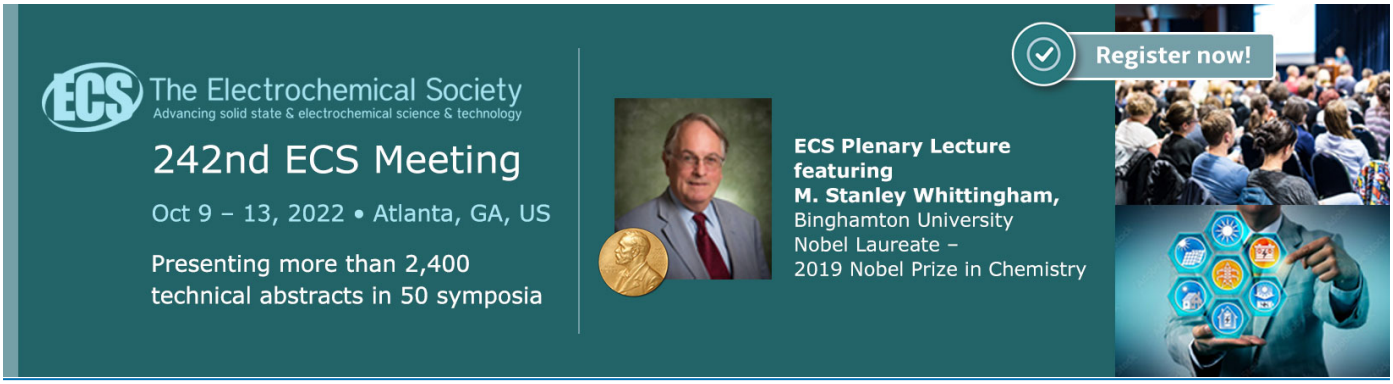
# Analysis of various Pre-processing filtering methods for MRI brain tumour detection in a noisy environment

To cite this article: Priyanka Jain and Priyasha Parmar 2022 *J. Phys.: Conf. Ser.* **2335** 012030

View the article online for updates and enhancements.

You may also like

- Synthesis of pseudo-CT images from pelvic MRI images based on an MD-CycleGAN model for radiotherapy
Hongfei Sun, Qianyi Xi, Rongbo Fan et al.
- Comparative study of microelectrode recording-based STN location and MRI-based STN location in low to ultra-high field (7.0 T) T2-weighted MRI images
Rens Verhagen, P Richard Schuurman, Pepijn van den Munckhof et al.
- Multi-to-binary network (MTBNet) for automated multi-organ segmentation on multi-sequence abdominal MRI images
Xiangming Zhao, Minxin Huang, Laquan Li et al.




This content was downloaded from IP address 14.139.251.98 on 21/09/2022 at 07:42

# Analysis of various Pre-processing filtering methods for MRI brain tumour detection in a noisy environment

**Priyanka Jain, *Priyasha Parmar**

Department of ECE, Delhi Technological University

Priyankajain@dtu.ac.in, priyashaparmar15@gmail.com

**Abstract**. Magnetic Resonance Imaging (MRI) is playing a very important role in the detection of various abnormalities in the human body. Biomedical image processing is helping us in the proper identification of various abnormalities such as a malignant tumour in various parts of the body. In this paper different types of filtering methods applied in the pre-processing stage have been compared. When MRI images have been corrupted by Gaussian noise then various filters are applied to obtain recovered MRI images. After performing filtering operation segmentation has been applied on MRI images to obtain tumour region. The processes have been implemented with the help of MatLab 2018.

## 1. Introduction

Recently several advances have been made in the field of application of image processing in medical applications [1]. Brain tumours are one of the deadly diseases with a low survival rate [2]. According to the statistics provided on cancer.net, it is the 10$^{th}$ leading cause of death in both men and women [2]. According to the World Health Organization (WHO) [3], tumours can be graded from grade 1 to grade 4 and this can be used to classify them into benign or malignant tumours. Benign tumours are from grade 1 and grade 2 and malignant tumours are classified under grade 3 and grade 4 and hence called high-grade tumours [3]. Early detection of low-grade tumours is very crucial because if these tumours are left unchecked then they may turn into high-grade tumours.

Generally, oncologists use MRI to obtain brain scans and detect any abnormal growth in it. However, manually reading MRI Scans is highly prone to errors. Sometimes MRI scans may even be corrupted by noise which again hinders proper identification of brain tumours. Hence, various image pre-processing techniques such as image filtering, segmentation, morphological operations etc. can provide very detailed images of the various tissues inside our body. With these detailed images, doctors can make a timely and efficient proper diagnosis [4].

There are different steps involved in brain tumour detection using image processing techniques. Major steps involved in medical image processing are listed in figure 1.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

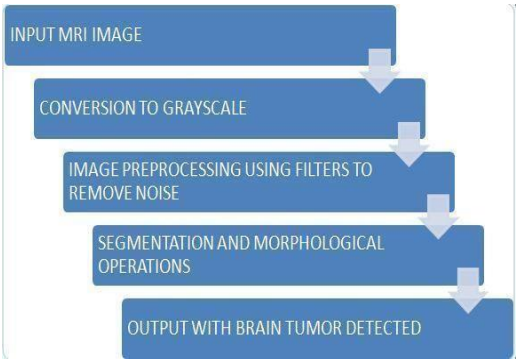



**Figure 1.** Block Diagram for brain tumour detection

First input, MRI image is converted to grayscale for ease of performing further operations Applying image processing operations like segmentation on corrupted images give us inaccurate results. Hence, pre-processing filtering techniques are applied. After removing noise, segmentation operation is applied. Segmentation involves breaking up images into subgroups called segments. This makes performing further operations easier with reduced complexity. Then Morphological operations such as erosion and dilation. Dilation adds pixels on object boundaries while erosion removes pixels on object boundaries are applied. These operations process the digital images based on the shape. The corresponding pixel in the input image is compared with its neighbours and the output image is obtained. Hence at the output brain tumour region is detected inside the MRI brain scan.

## 2. Motivation

Timely and efficient diagnosis of various cancerous tissues growing inside the body is crucial. However, it depends on doctors manually identifying the malignant region inside the brain. This process is highly inefficient. Sometimes it becomes very difficult to take MRI brain scans in a noisy environment. Image processing techniques based on anisotropic diffusion filtering can further help in faster and more efficient detection of various tumours drawing inside the brain. To enhance the quality of MR Image various denoising filters are available in the image processing. In this paper, a comparison of various nonlinear filters has been done on the bases of calculation of Peak Signal to Noise Ratio(PSNR) and Signal to Noise ratio (SNR).

## 3. Literature Review

Anisotropic diffusion filter technique was proposed by Witkins [1] where original images are convolved with Gaussian kernels to obtain the final image. However, accurate locations of different meaningful edges could not be obtained. Perona and Malik *et al.* [1] have used partial differential equations to described anisotropic diffusion and filtering technique. Using anisotropic diffusion filtering techniques to obtain sharp region boundaries in the images [1]. Hence the issue of blurring was solved and sharp edges could be retained. However, the suggested anisotropic diffusion filters technique has not been applied to MRI images.

MRI evaluation used anisotropic diffusion filtering for preserving edges and MRI intensities were suggested by Palma Cappabianco ide Miranda *et al.* [4]. However, no analysis on the performance of the anisotropic diffusion technique has been done if MRI images are corrupted by Gaussian noise.

Bahadure *et al.*[5] worked on image analysis for MRI-based brain tumour detection and feature extraction using biologically inspired Berkely wavelet transform and Support Vector Machine[5]. Berkeley wavelet transformation for the brain tumour segmentation has been performed. Support

vector machine classifiers are also being used so as to extract certain features from the tissues [5]. They have then utilized a certain threshold value for identifying and contouring images. However, the behaviour of MRI images under noisy conditions in pre-processing steps is not observed.
We have studied the various filtering techniques in digital image processing which is being tabularized in Table 1.

**Table 1. Previous Works**

| Sno. | Title | Author | Findings |
|---|---|---|---|
| **1.** | **Using Image Processing Techniques, Automatic Brain Tumor Detection in MRI [8]** | **Mariam Saii, Zaid Kraitem** | a) The pre-processing level, anisotropic diffusion filter is applied to the MRI image.<br>b) Pre-processing stage is not analysed under noisy conditions. |
| **2.** | **Denoising MRI images [7]** | **Hanafy m Ali** | a) salt and pepper and Gaussian noises are being added to the MRI images then a median filter is applied to extract the desired image.<br>b)Other edge detection techniques like guided filters or non-local means have not been explored. |
| **3.** | **MRI-based brain tumour detection and feature extraction using biologically inspired BWT and SVM [5]** | **NB Bahadure AK Ray and HP Thethi** | a) Performed Berkeley wavelet transformation for the brain tumour segmentation.<br>b) Support vector machine classifiers are also being used so as to extract certain features from the tissues. However, the scenario where MRI images are corrupted by Gaussian noise is not considered |

Rashid *et al*. [6] research work on anisotropic techniques and SVM classification have also proposed various morphological operations and anisotropic diffusion filters for segmentation and identification of infected brain tumour tissues. The author has tried to filter MRI images that were corrupted because of salt and pepper noise. However, MRI images corrupted by Gaussian noise is not considered.

Hanafy m Ali *et al*. [7] noise removal can cause various challenges such as various important features of MRI could be blurred due to which visualization and detection of the tumour could become a problem [7]. Here the median filter algorithm is being used and it is modified. First salt and pepper and Gaussian noises are being added to the MRI images then a median filter is applied to extract the desired image[7]. Here denoising of MRI images is performed only using median and adaptive median filters. Other edge detection techniques like guided filters or non-local means have not been explored.
In this paper, anisotropic filtering is applied on an MRI image under noisy conditions and performance comparison has been done with the other existing nonlinear filters.

## 4. Methodology

Filters used in image processing could be linear or nonlinear. Usually, Non-linear filtering techniques perform better in comparison to linear filtering methods [10]. This is because they are able to preserve significant brain tumour tissue important for tumour detection and identification [10]. The comparison of the behaviour of linear vs non-linear filter is given below in figure 2.

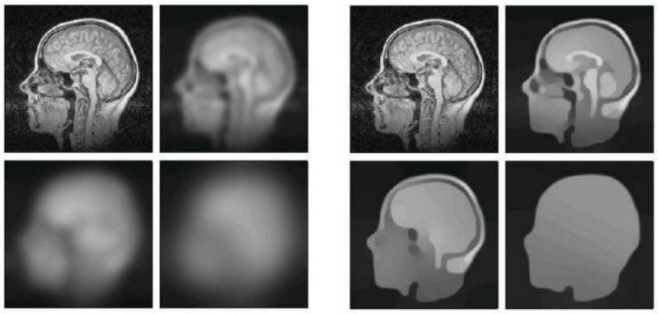


**Figure 2.** Left to right: Linear Filtering performed on MRI image corrupted with noise, Non-Linear Filtering performed on MRI image corrupted with noise, Comparison Between linear and nonlinear filtering[10]

In Figure 2 filtering operation on MRI images is compared between linear and non-linear filter. Linear filtering(Left) while removing noise from the MRI image distorts the edges present in the image. As a result, the image becomes blurred. Non-Linear Filtering(Right) effectively removes the noise from images while preserving their edges.

## 5. Non-Linear Filtering Techniques

In this section various nonlinear filters such as anisotropic diffusion, guided etc. has been explained in brief.

### *5.1 Anisotropic Diffusion Filtering Techniques:*

This Edge preserving diffusion filtering technique in image processing was formulated by Perona and Malik [1]. They are applied in images for mainly smoothing and restoration.

Anisotropic diffusion equation is [1],

$$I=\mathrm{div}(c(x,y,t)\nabla I)=c(x,y,t)\Delta I+ \nabla c.\nabla I \quad (1)$$

I = Image under consideration
c = diffusion coefficient
div = divergence operator
$\nabla$ and $\Delta$ = Gradient and Laplacian operator

These filtering operations are applied to obtain enhanced images whose edges are preserved.

*5.2 Average filtering*

Average filter image processing technique tries to reduce the amount of intensity variation which is present between the neighbourhood pixels. It usually moves through the image pixel by pixel. Further, replace each value with the average value of the neighbouring pixels [7].

*5.3 Guided Filter*

In guided filter, there are two images: input image and guided image. Weights are generated using a guiding image for a particular pixel. Then each of the pixels is multiplied by the input image with that weight [15]. Although guided filter preserves the sharpness and the edges of the MR Image but it has high computational complexity. A variant of the input image could be used as a guidance image.

$$q_i \qquad \text{qi} = \text{Wi,j (I) pj} \tag{2}$$

Where I = guidance image
$q$ = output
$p$ = input
$W$ = weight
$i, j$ = pixel indexes

Significant advantages in guided image filtering over other filters like better quality and efficiency in noise reduction in detail smoothing, no gradient reversal problem [15]. The only disadvantage is guided filtering without a sharp guidance image might not work.

*5.4 Non-local means filter*

Local mean and non – local mean filters algorithm exists in the literature. In the local means algorithm mean value of a group of pixels surrounding a target pixel is taken to smooth the image [12]. However, in non local means filter takes mean of all the pixel in the image weighted by how similar these pixels are to the target pixel [12]. Hence it provides better post-filtering clarity in MRI images. Non local means filter cleans the image containing noise without losing much of the significant details and edges [12].

However, the main disadvantage of non-local means filter is that then the noise increases and its performance deteriorates substantially [12]. That the noised image often suffers from the problem of blurring and there is the subsequent loss of image details [12].

## 6. Experimental Results

Different datasets and MRI images were taken for the implementation of anisotropic diffusion techniques. The coding is performed using MATLAB-R2018b. Following is the implementation of filtering operation MRI brain scans corrupted by noise.

*6.1 Comparative Analysis of Filter in Pre Processing Stage*

Comparative Analysis of filtering techniques applied on noisy MRI images is performed and their signal to noise ratio and peak signal to noise ratio is calculated. Here MRI images are first subjected to Gaussian noise then different filters are applied to check their performance.Gaussian

noise is considered because it is additive in nature and it is easier to model noise which might be introduced in MRI scans.

**Table 2. Shows the comparison between noisy and filtered MRI images**

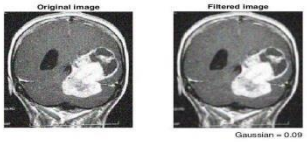
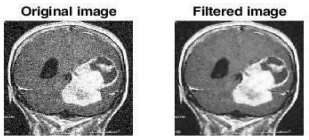
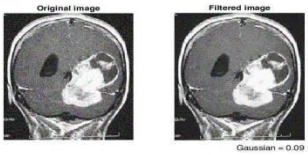
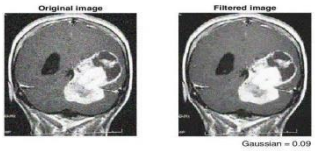
| Sno. | Filter Applied | Results |
|---|---|---|
| 1. | Averaging Filter |   |
| 2. | Anisotropic Diffusion Filter |   |
| 3. | Guided Filter |   |
| 4. | Non-Local Means |   |

**Table 3.** SNR and PSNR were obtained for filtering operations on MRI images

| SNo. | Filter Applied | SNR(Signal to Noise Ratio) | PSNR(Peak Signal to Noise Ratio) |
|---|---|---|---|
| 1. | Averaging Filter | 11.0165 | 17.7974 |
| 2. | Anisotropic Diffusion Filter | 16.4767 | 22.1250 |
| 3. | Guided Filter | 11.0111 | 17.7920 |
| 4. | Non-Local Means Filter | 11.0089 | 17.7899 |

Table 3 shows the SNR and PSNR of different filters applied at the pre-processing stage. SNR and PSNR are calculated using the formulas as described below.
Let AxB be the image. If original AxB image is f(a,b) and filtered image be g(a,b)
SNR is calculated using below formula[7]:

$$SNR = \sum_{i=0}^{A-1}\sum_{j=0}^{B-1} g(a,b)^2 / \sum_{i=0}^{A-1}\sum_{j=0}^{B-1}[g(a,b)-f(a,b)]^2 \quad (3)$$

To calculate the PSNR, Mean squared error is measured using the formula given below[7]:

$$MSE=\sum_{A,B}[I_1(a,b)-I_2(a,b)]^2 / A*B \quad (4)$$

Here a, b are number of rows and columns of input image.
Then PSNR is calculated using below formula[7]:

$$PSNR = 10\log_{10}(R^2/MSE) \quad (5)$$

Here R represents the maximum fluctuation in input image datatype. For eg. for an 8 bit unsigned integer datatype R is 255.On comparing values obtained in Table 3 it can be observed that the anisotropic filter performs better in comparison to other filters.
Then anisotropic diffusion filtering is applied to perform further operations to detect brain tumours. Tumour detection was performed further using anisotropic diffusion filtering.

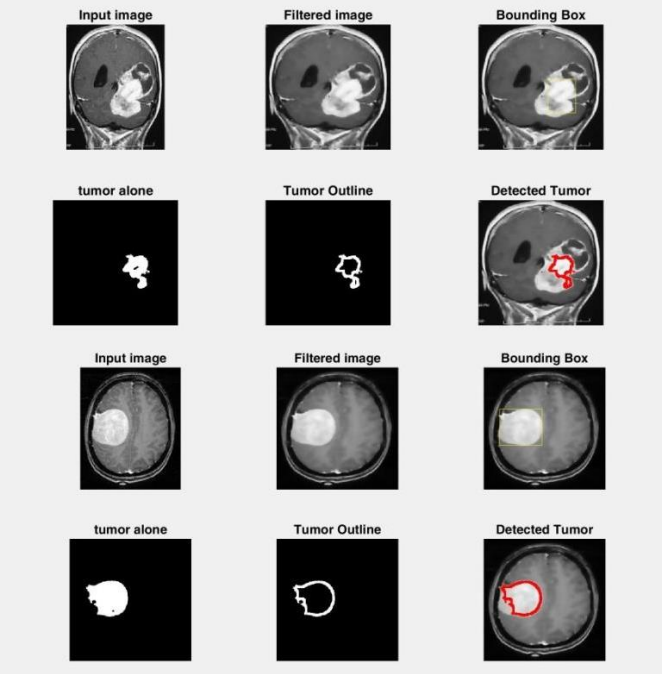



**Figure 3.** Tumor detection performed on different MRI images

Figure 3 shows the tumour detected after applying anisotropic filtering operation. With the help of an anisotropic filtering technique, detection of malignant tumour region in the reconstructed MRI image can be done.

## 7. Conclusion

A comparative analysis of different nonlinear filters has been done by applying these filters in pre-processing stage to reconstruct MRI images corrupted by Gaussian noise. It can be seen that techniques like anisotropic diffusion filtering when applied in pre-processing stage filter MRI images efficiently. These noise-free reconstructed MRI images help in the proper analysis of malignant tumours in brain tumour detection.

## References

[1] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," in *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 7*, pp. 629-639, July 1990, DOI: 10.1109/34.56205.
[2] https://www.cancer.net/cancer-types/brain-tumor/statistics#:~:text=It%20is%20estimated%20that%2018%2C280,and%20CNS%20tumors%20in%202020.
[3] https://academic.oup.com/neuro-oncology/article/23/8/1231/6311214
[4] Caio A. Palma, Fábio A.M. Capobianco, Jaime S. Ide, Paulo A.V. Miranda, Anisotropic Diffusion Filtering Operation and Limitations - *Magnetic Resonance Imaging Evaluation, IFAC Proceedings Volumes, Volume 47*, Issue 3,2014
[5] ResearchArticle "Image Analysis for MRI Based Brain Tumor Detection and Feature Extraction Using Biologically Inspired BWT and SVM" Nilesh Bhaskar Rao Bahadur,1 ArunKumarRay,1 and HarPal Thethi2 *Hindawi International Journal of Biomedical Imaging Volume 2017,* Article ID 9749108
*[6]* "Brain Tumor Detection Using Anisotropic Filtering, SVM Classifier, and Morphological Operation from MR Images," M.H.O Rashid, M.A. Mamun, M.A. Hossain, M. P. Uddin, *Proceedings of the international conference on computer, communication, chemical, material, and electronic engineering, 8-9 February 2018, Bangladesh, IEEE 2018.*
[7] Fundamental Filters for MRI Medical Image Denoising Ali, Hanafy M. at http://dx.doi.org/10.5772/intechopen.72427 towards the end of the chapter.
[8] Mariam Saii, Zaid Kraitem. Automatic Brain Tumor Detection in MRI Using Image Processing Techniques. *Biomedical Statistics and Informatics. Vol. 2, No. 2, 2017,* pp. 73-76. DOI: 10.11648/j.bsi.20170202.16
[9] Platero C., Sanguino J., Velasco O. Nonlinear Diffusion Filters without Parameters for Image Segmentation. In: Jiang X., Petkov N. (eds) Computer Analysis of Images and Patterns. CAIP . Lecture Notes in Computer Science, vol 5702. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-03767-2_63
[10] http://webfiles.portal.chalmers.se/s2/undergraduate/SSY095/PDFdocuments/ForScreen/Notes/NonLinearDiffusion.pdf
[11] Dual Filter Based Images Fusion Algorithm for CT and MRI Medical Images M.N. Narsaiah, S. Vathsal, D. Venkat Reddy. *International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-9, July 2019*
[12] Baozhong Liu Jianbin LIU "Overview of image noise reduction based on non-local mean algorithm " January 2018 *MATECWeb of Conferences 232(10):0302* DOI:10.1051/matecconf/201823203029
[13] T. S. Kumar, K. Rashmi, S. Ramadoss, L. K. Sandhya and T. J. Sangeetha, "Brain tumor detection using SVM classifier," *2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS), 2017,* pp. 318-323, DOI: 10.1109/SSPS.2017.8071613.
[14] W. Zhu and Y. Shen, "A Region Growing Segmentation Approach for MRI Brain Image Processing," 2019 *IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID), 2019,* pp. 188-191, DOI: 10.1109/ICASID.2019.8925066.
[15] J. Wu, C. Wang and Y. Xu, "An Improved Guided Filtering Algorithm for Image Enhancement," *2018 IEEE International Conference on Multimedia and Expo (ICME), 2018,* pp. 1-6, doi: 10.1109/ICME.2018.8486575.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 July 2022, accepted 8 August 2022, date of publication 19 August 2022, date of current version 29 August 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3200386*

RESEARCH ARTICLE

# Chaotic Mapping Based Advanced Aquila Optimizer With Single Stage Evolutionary Algorithm

**MONIKA VERMA[1], (Student Member, IEEE), MINI SREEJETH[1], (Senior Member, IEEE), MADHUSUDAN SINGH[1], (Senior Member, IEEE), THANIKANTI SUDHAKAR BABU[2], (Senior Member, IEEE), AND HASSAN HAES ALHELOU[3], (Senior Member, IEEE)**

[1]Department of Electrical Engineering, Delhi Technological University, New Delhi 110042, India
[2]Department of Electrical and Electronics Engineering, Chaitanya Bharathi Institute of Technology, Hyderabad 500075, India
[3]Department of Electrical Power Engineering, Faculty of Mechanical and Electrical Engineering, Tishreen University, Lattakia 2230, Syria

Corresponding authors: Hassan Haes Alhelou (alhelou@ieee.org) and Thanikanti Sudhakar Babu (sudhakarbabu66@gmail.com)

**ABSTRACT** The intelligent optimization techniques have been introduced by carefully observing the behavior of various hunters like a whale, grey wolf, Aquila, and lizards for estimating global optimum solutions in fair time by forming appropriate mathematical models. However, hunting-based algorithms suffer from slow and pre-requisite convergence and get caught up in local minimum. Aquila optimizer (AO) is one of the recently developed hunting-based methods that encounter a similar type of shortcoming in a few situations. This research introduces the concept of chaotic mapping to the standard AO in order to increase the convergence speed. Also to maintain the balance of exploration performed by AO with its exploitation capability, a single stage evolutionary algorithm is also integrated with it. The performance of standard AO and modified AO are tested for well-defined unimodal and multimodal Benchmark functions. The proposed framework produces one population by standard AO and a new population by single stage genetic algorithm based evolutionary concept in which binary tournament selection, roulette wheel selection, shuffle crossing over and displacement mutation occur to generate a new population. The chaotic mapping criteria are then applied to obtain various variants of the standard AO technique. The general results obtained from the proposed novel chaotic mapping-based advanced AO with single stage evolutionary algorithm shows that it outperforms the standard AO. This advanced technique is thus applied to real-world design engineering problems to study its significance from an industrial point of view.

**INDEX TERMS** Hunting-based algorithm, evolutionary algorithm, Aquila optimizer (AO), chaotic maps.

## I. INTRODUCTION

Several metaheuristic methods are utilized for addressing numerous real-world applications. For example, routing of vehicles [1], robot path tracker, robot gripping [2], electric motor geometry optimization [3], and design of distribution system of water [4].

The development in metaheuristic techniques has gained pace due to its flexibility, simplicity, ability to achieve accuracy while solving linear, non-linear or complex real world optimization problems. These metaheuristic techniques are developed on the basis of different inspirations. The search strategies in different optimization can be Stochastic based, Enumerative based or Calculus based. Based on the search methods, the classification of state-of-the-art optimization methodologies [5], [6], [7] is presented in Figure 1.

The associate editor coordinating the review of this manuscript and approving it for publication was Kan Liu.

### A. MOTIVATION AND PRIOR WORK

Out of so many inspirations, stochastic search based have raised the interest of researcher for obtaining advancement


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

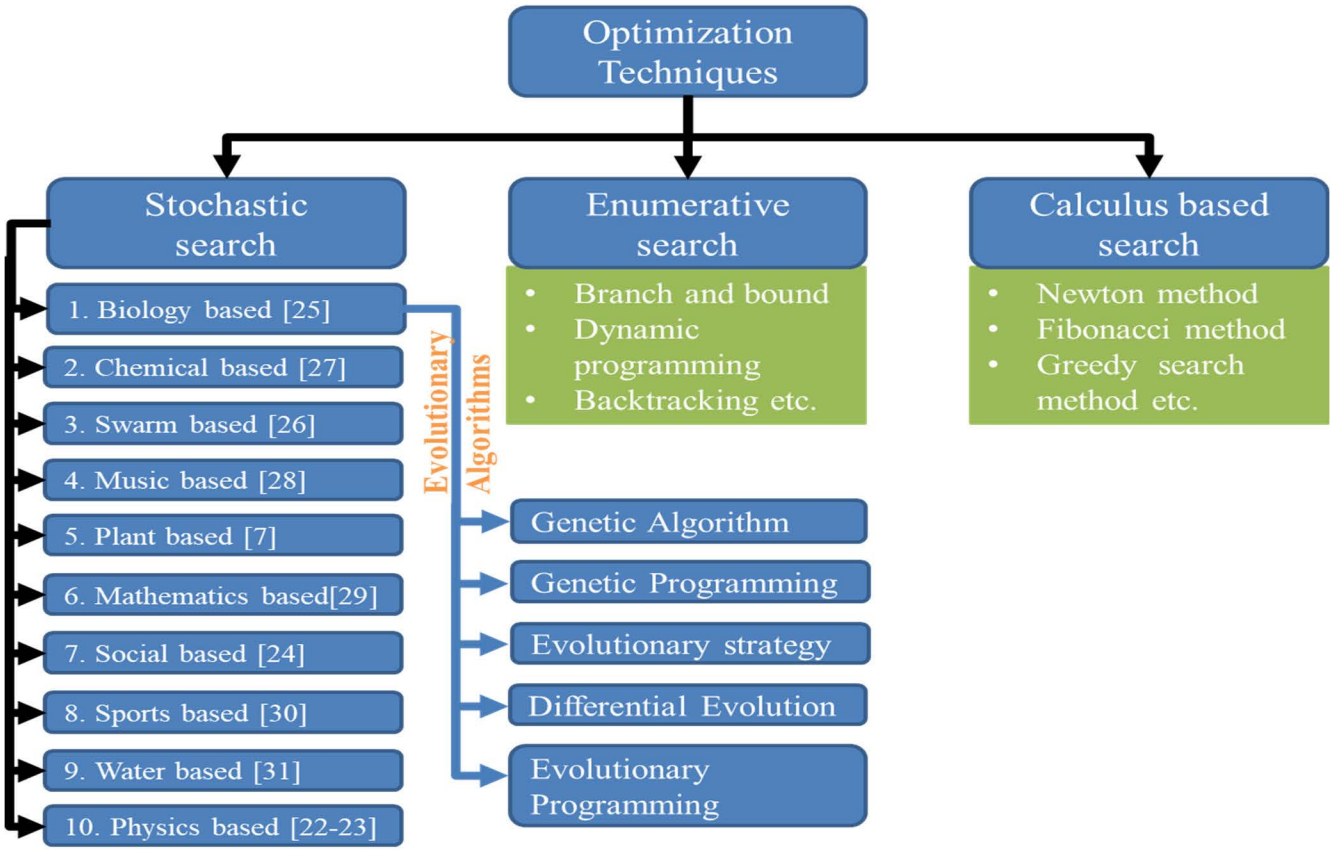



**FIGURE 1.** Taxonomy of state-of-the-art of optimization techniques.

in the performance of existing frameworks. By doing so, the updated versions of existing metaheuristic frameworks become more efficient for solving many complex real world problems [8], [9], [10], [11], [12], [13], [14], [15]. The evolutionary algorithms are based on the laws of natural evolutions. The typical evolutionary algorithms are Genetic Algorithms (GA) [16], Genetic Programming (GP) [17], Evolutionary Strategy (ES) [18], Differential Evolution (DE) [19], and Evolutionary Programming (EP) [20]. For obtaining single string of solution, single stage evolutionary algorithm of GA is the most appropriate among all [21].

Apart from this, stochastic search based algorithms include physics based [22], [23], social based [24], biology based [25], swarm behavior [26], chemical based [27], music based [28], plant based [7], mathematics based [29], sports based [30], water based [31] algorithms.

The Aquila Optimizer (AO) is one of the new technique which is recently designed based on the hunting behavior of Aquila. The genus of 'eagle', i.e., Aquila is suspected to be carrying Zeus's thunderbolt in Roman mythology. Aquila is a proficient hunter. AO algorithm is designed by sub-dividing the exploration as well as exploitation phases on the basis of expanded and narrowed vision. This unique feature guarantees that AO traverse the search region with diverse vision. However, it exhibits few limitations of showing disproportion in the speed of convergence during stochastic search. The other drawbacks are falling for locally optimized region due to discrimination issues and imbalance between exploration-exploitation phases of AO.

The evolutionary algorithm, GA is designed using a strong base of biology inspired mechanism. Originally, GA is capable of generating reprints of procedures of natural selection, the concept of which was introduced by Charles Darwin. Although the concept of biological evolution is obscure, the experimental support exists for few facts related to it; i.e., (1) The concept of evolution is applicable more to genotype (chromosomes) than to phenotype (organisms) themselves. Those genotypes are considered to be the tools responsible for coding the life and the corresponding biological form is deciphered using the information possessed by the chromosomes. (2) The mechanism of natural selection is responsible for the relationship between genotypes and phenotypes. It gives rise to better expectation of life and enables those individuals to reproduce that are the most adapted. (3) The evolution occurs during the phase of reproduction.

Contrarily, the technicalities of biological theory based concepts of natural selection, reproduction, and mutation which are an integral part of the basic genetic algorithm (GA) show sluggish elongated search time if applied solo to optimization problems. However, its high accuracy, robustness,

and efficiency of solving complex problems make it a proficient contender to implementation in the integrated structure [32].

The three main reasons, which rises limitations in the stochastic computations in origins of AO and GA are given as:

- The populations are dynamic in nature. Thus, a variation is observed in the individuals' number, which is in contrary to the fixed population used in both AO and GA.
- The crossover and mutation processes, which occur during the phase of reproduction in GA, are totally chaotic and are independent of any parameter.
- The initial population does not possess completely random characteristics. If there are chaos in reproduction, the chaos will be present in the population as well.

In [33] and [34], the authors introduced the algorithms possessing variable populations which is defined by chaos. The performance of metaheuristic algorithms gets improved using this dynamic configuration.

In Chaotic map theory, the study of dynamical systems is carried out which corresponds to the systems that gets evolved over time. In chaotic mapping, the present state depends upon the characteristics of the previous state. It is in contrast to the random systems, in which the prediction of the upcoming state is not possible due to lack of correlation between the previous and upcoming state.

The use of chaotic maps in the generation of initial population of AO helps to avoid convergence to a locally optimized region. The integration of single stage evolutionary algorithm, GA with AO helps to increase the convergence speed and avoid imbalance between exploration-exploitation phase of AO. As per the literature, there have been few researchers who attempted to improve the performance of AO using different techniques. Table 1 shows the brief overview of modified versions of AO indicating innovations and research gaps.

The GA, on the other hand, was introduced in 1992 [25]. Although, there have been various variants of the basic genetic algorithm since then shown in [35], from a classical point of view, the basic steps are used to update the current solution as per the single-stage evolutionary genetic algorithm due to the following advantages:

- A binary tournament selection strategy, being easier to implement, ensures perseverance of population diversity.
- A simulated binary crossover strategy, ensures better exploration.
- The polynomial mutation strategy, ensures manifoldness.

Hence in order to improve the performance of standard AO, the evolutionary genetic algorithm is integrated with the standard AO to propose seven novel variants of AOs by introducing chaos to it. The proposed novel technique is named as chaotic mapping-based modified Aquila Optimizer with single-stage evolutionary (CMAOE) algorithm.

By single-stage evolutionary algorithm, it is meant that only one stage of the overall genetic algorithm containing selection, crossover, and mutation is being implemented in the proposed CMAOE method in this article.

**TABLE 1. Literature related to prior work in improvement of AO.**

| Algorithm | Authors | Year | Innovations | Gaps |
|---|---|---|---|---|
| IHAOHHO[1] [25] | Wang et. al. | 2021 | Based on hybridization of AO and HHO[2] | Originality of AO is compromised. |
| IAO[3] [26] | Ma et. al. | 2021 | Concept of quasi-opposition learning and wavelet mutation strategies is used | Additional uncertainty in AO is observed. |
| AOAAO[4] [27] | Zhang et. al. | 2022 | Based on hybridization of AOA[5] and AO | High computation time |
| IAO [28] | Ewees et. al. | 2022 | Search strategy of AO is boosted using operators of WOA[6] | Originality of WOA and AO are compromised. |
| AGWO[7] [29] | Ma et. al. | 2022 | Based on hybridization of GWO[8] and AO | Originality of both techniques are not retained. |
| enAO[9] [30] | Ekinci et. al. | 2022 | Enhanced version of AO is developed using modified opposition-based learning and Nelder-Mead simplex search method | Additional complexity in computation is increased. |

[1]Improved Hybrid Aquila Optimizer and Harris Hawks Optimization (IHAOHHO)
[2]Harris Hawk Optimization (HHO)
[3]Improved Aquila Optimizer (IAO)
[4]Arithmetic Optimization Algorithm with Aquila Optimizer (AOAAO)
[5]Arithmetic Optimization Algorithm (AOA)
[6]Whale Optimization Algorithm (WOA)
[7]Aquila Grey Wolf Optimization (AGWO)
[8]Grey Wolf Optimization (GWO)
[9]enhanced Aquila Optimizer (enAO)

### B. AIM

The primary goal of this article is to study the utilization of Chaotic maps for generating the initial population of CMAOE. The performance of the proposed CMAOE algorithm is compared with standard AO for constrained Benchmark functions. The best out of all proposed variants of AO is thus tested on a real world design engineering problem. In addition, the entropy in the initial population is measured and analyzed.

### C. CONTRIBUTIONS

This article contributes the following novel features of the proposed CMAOE technique:

- The dynamical system of the stochastic initial population has consistent irregularity in several solutions. This happens during the expanded and narrowed diversification stage of standard AO and reproduction stage of GA which is entirely chaotic in nature. Hence, the

chaotic mapping systems are introduced to strengthen the prominence of the initial population.
- The originality of standard AO is retained by not altering the parameters of AO.
- The performance evaluation of the CMAOE method is carried out by employing it for 13 benchmark functions consisting of unimodal and multimodal functions.
- The effectiveness of variants of AO is compared with the standard AO technique by comparing their respective convergence profiles.
- The potency of CMAOE is inspected by implementing it for a real-world designed optimization problem which shows the significance of research carried out in this work from an industrial point of view.

### D. PAPER ORGANIZATIONS

The paper organization is as follows: In addition to the introduction prescribed in section I, Section II presents a detailed explanation of the methods and materials exploited in achieving the goals of this work. In section III, modified AO with chaotic mapping and evolutionary genetic algorithm is explained. Section IV presents empirical studies of proposed CMAOE along with the performance analysis results and discussions. Section V shows the application of the proposed algorithm to solve the real-world design engineering of electric motors. The results, obtained by the proposed CMAOE are compared with the analytical results obtained by the means of finite element analysis. The concluding remarks and future scope of this research are listed in section V.

## II. METHODS AND MATERIALS

This section presents the details of the materials and methods utilized in order to achieve the aims of this study.

### A. STANDARD AO

The standard AO begins with the generation of a stochastic population consisting of competent solutions ($x$) belonging to the feasible spectrum of decision variables of the problem (*Equation (1)*).

$$x = \begin{bmatrix} x_{1,1} & \cdots & x_{1,D} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,D} \end{bmatrix} \tag{1}$$

where $x_{n,D}$ represents value of $D^{th}$ variable in the $n^{th}$ solution, where $D$ denotes the number of variables. The present solution $x$ is computed using *Equation (2)*, where $i = 1, 2, \ldots n$ and $j = 1, 2, \ldots D$, $L_j$ is the lower bound of solution $x$ value, $U_j$ is the upper bound of solution $x$ value and $r$ is a random constant.

The standard AO takes into consideration the hunting characteristics of Aquila in mathematical form, expressed in four separate stages to follow.

$$x_{ij} = L_j + r \times (U_j - L_j) \tag{2}$$

- ***Stage-I: Steep soar with upright dip***

At first, the apt area for prey hunting is selected by exploring the search zonal area from steep soar with upright sag. The mathematical form of this action is expressed as *Equation (3)*:

$$x_1(t+1) = x_b(t) \times \left(1 - \frac{t}{T}\right) + (x_m(t) - x_b(t)) \times r \tag{3}$$

where at $t^{th}$ iteration, $x_b(t)$ and $x_m(t)$ are the best and the mean solution respectively, $x_1(t+1)$ is the revised solution due to Stage-I and factor $\left(1 - \frac{t}{T}\right)$ is controlling the D-ph in accordance with the iteration where $T$ represents the upper limit of iterations. *Equation (4)* determines the term $x_m(t)$.

$$x_m(t) = \frac{1}{n}\sum_i^n x_{ij}(t); \quad j = 1, 2, \ldots D \tag{4}$$

- ***Stage-II: Contour navigation with concise drift attack***

Now a circular contour is formed above the target prey thereby narrowing the zonal area for hunting. The action is named 'contour navigation with concise drift attack', represented as *Equation (5)*:

$$x_2(t+1) = x_b(t) \times L(D) + x_r(t) + (y - x) \times r \tag{5}$$

where at $t^{th}$ iteration, $x_r(t) \epsilon [1, n]$ is a randomly selected solution, $L(D)$ is a flight levy function defined by *Equation (6)* and $x_2(t+1)$ is the revised solution due to Stage-II. Factor $(y - x)$ is controlling the spiral contour shape in the course of exploration, where $x$ and $y$ are represented by *Equations (8) and (9)*.

$$L(D) = \frac{s \times u \times \sigma}{|v|^{\frac{1}{\beta}}} \tag{6}$$

$$\sigma = \frac{\Gamma(1+\beta) \times \sin\frac{\pi\beta}{2}}{\Gamma(\frac{1+\beta}{2}) \times \beta \times 2^{\frac{\beta-1}{2}}} \tag{7}$$

$$x = r \times \sin\theta \tag{8}$$

$$y = r \times \cos\theta \tag{9}$$

$$r = r_1 + p \times d_1 \tag{10}$$

$$\theta = -\omega \times d_1 + \theta_1 \tag{11}$$

where $u, v \epsilon [0, 1]$ are random numbers obtained from standard normal distribution, $s$, $\beta$, $r_1$, $p$, $\omega$ and $\theta_1$ are constant values and $d_1 \epsilon [1, D]$ represents an integer vector.

- ***Stage-III: Lowering navigation with gradual plunge attack***

After stating the prey area, the lower navigation is executed with a gradual drop. The area of prey is widely exploited by this action called 'lowering navigation with gradual plunge attack', given by *Equation (12)*.

$$x_3(t+1) = (x_b(t) - x_m(t)) \times a - r + ((U - L) \times r + L) \times \delta \tag{12}$$

where $a$ and $\delta$ are controlling parameters of wide exploitation and $x_3(t+1)$ is the revised solution due to Stage III.

- ***Stage-IV: Stroll and catch prey***

When prey is nearby located, it is strolled and grabbed corresponding to its random movement. This action is represented by *Equation (13)*.

$$x_4(t+1) = Q \times x_b(t) - (g_1 \times x(t) \times r) - (g_2 \times L(D)) + r \times g_1 \quad (13)$$

where at $t^{th}$ iteration, quality function $Q$ is computed using *Equation (14)* and $x_4(t+1)$ is the revised solution due to Stage-IV. Parameters $g_1$ and $g_2$ are controlling the track path of the prey using *Equations (15) and (16)* respectively.

$$Q(t) = t^{\frac{2\times r-1}{(1-T)^2}} \quad (14)$$

$$g_1 = 2 \times r - 1 \quad (15)$$

$$g_2 = 2 \times (1 - \frac{t}{T}) \quad (16)$$

### B. PROPOSED CMAOE ALGORITHM

The CMAOE algorithm consists of replacement of random procedures of generating initial population by Chaotic maps. The chaotic map functions, for example, the Henon map [36], build a numerical series to produce the population by following *Equation (17)*:

$$x_n = 1 - A * x_{n-1}^2 + B * x_{n-1} \quad (17)$$

where $x_n$ represents each individual member of population $x$ (described in *Equation (1)*) whose position depends on the position of the previous individual, $x_{n-1}$.

In CMAOE, the standard AO is upgraded by integrating the single stage GA to further escalate the search precision and accuracy.

Each revised solution vector at $t^{th}$ iteration obtained at *stage IV* of standard AO is treated as the population of candidate solutions for further steps of GA workflow, having the following steps:

- ***Step (i) of GA in CMAOE:***

In this step, binary tournament selection is implemented and thus, the parent solution is obtained. This is a recurring process for the selection of the agents as long as the CMAOE algorithm is administered.

- ***Step (ii) of GA in CMAOE:***

Two-parent agents $P_1 = [P_{1,1}, P_{1,2}, .., P_{1,D}]$ and $P_2 = [P_{2,1}, P_{2,2}, \ldots, P_{2,D}]$ are picked randomly by Simulated Binary Crossover (SBX). The generation of each element of offspring $O_1$ and $O_2$ is done using *Equation (18)*.

$$O_1 = 0.5 \times [(1+\mu) \times P_1 + (1-\mu) \times P_2]$$
$$O_2 = 0.5 \times [(1-\mu) \times P_1 + (1+\mu) \times P_2] \quad (18)$$

$$\mu = \begin{cases} (2r)^{\frac{1}{(\eta_c+1)}}; & if\ r \le 0.5 \\ \left(\frac{1}{2(1-r)}\right)^{\frac{1}{(\eta_c+1)}}; & otherwise \end{cases} \quad (19)$$

where $\mu$ is the crossover SBX controlling factor, computed by *Equation (19)*, where $r$ is a random number, $\eta_c = 20$ is the crossover probability distribution mean.

- ***Step (iii) of GA in CMAOE:***

This step involves the mutation of offspring. Each offspring agent is mutated to produce mutated offspring $O_m$ from offspring $O$ using *Equation (20)*.

$$O_m = O + (U - L) \times \Delta \quad (20)$$

where $\Delta$ is the mutation control factor, calculated using *Equation (21)*.

Where $r$ is a random number, $\eta_m = 20$ is the mutation probability distribution mean. Thereafter, the combined population is obtained and sorted non-dominantly. Thus, the best solution vector of dimension $[1, n]$ is updated.

$$\Delta = \begin{cases} (2r)^{\frac{1}{(\eta_m+1)}} - 1; & if\ r < 0.5 \\ 1 - [2(1-r)]^{\frac{1}{(\eta_m+1)}}; & if\ r \ge 0.5 \end{cases} \quad (21)$$

The other parameters related to CMAOE, such as size of population, number of generation, Levy flight control parameters, crossover probability, mutation probability etc. are kept constant. Table 2 shows the parameter settings of the CMAOE algorithm. The process flow diagram of chaotic mapping-based advanced Aquila Optimizer with single-stage evolutionary algorithm (CMAOE) is presented in Figure 2.

**TABLE 2. Parameter setting for experimental analysis.**

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| Size of population | 20 | $p$ | 0.0265 |
| Number of generations | 1000 | $\omega$ | 0.005 |
| Crossover probability | 0.8 | $\theta_1$ | $3\pi/2$ |
| Mutation probability | 0.2 | $a$ | 0.1 |
| $s$ | 0.01 | $\delta$ | 0.1 |
| $\beta$ | 1.5 | $r_1$ | 10 |

### C. BENCHMARK FUNCTIONS

The Benchmark Functions are set of functions, defined by Siva Sathya and Radhika [37] to assess the performance of algorithms. These functions are classified by the; (i) ability to be non-separable or separable functions, (ii) dimensionality of search region, (iii) multimodality. The separable functions having $m$ variables can be re-written in the form of sum of $m$ functions having only one variable. In case of random distribution of local optima in the search region, complexity of searching for the global optimum point becomes more complex. This complexity is tested using multimodal functions. The dimensionality of search space also affects the performance of algorithms. It signifies the accuracy in computation times of particular algorithm. This work uses the benchmark functions as presented in Figure 3. The unimodal functions used are Sphere, Schwefel 2.22, Rotated hyper ellipsoid, Schwefel 2.21, Rosenbrock, Step and Quartic. The multimodal functions used are Schwefel, Rastrigin, Ackley, Griewank, Penalized1 and Levy N.13.

### D. INITIAL POPULATION

The CMAOE method implements the chaotic processes for the generation of initial populations. The generation of the

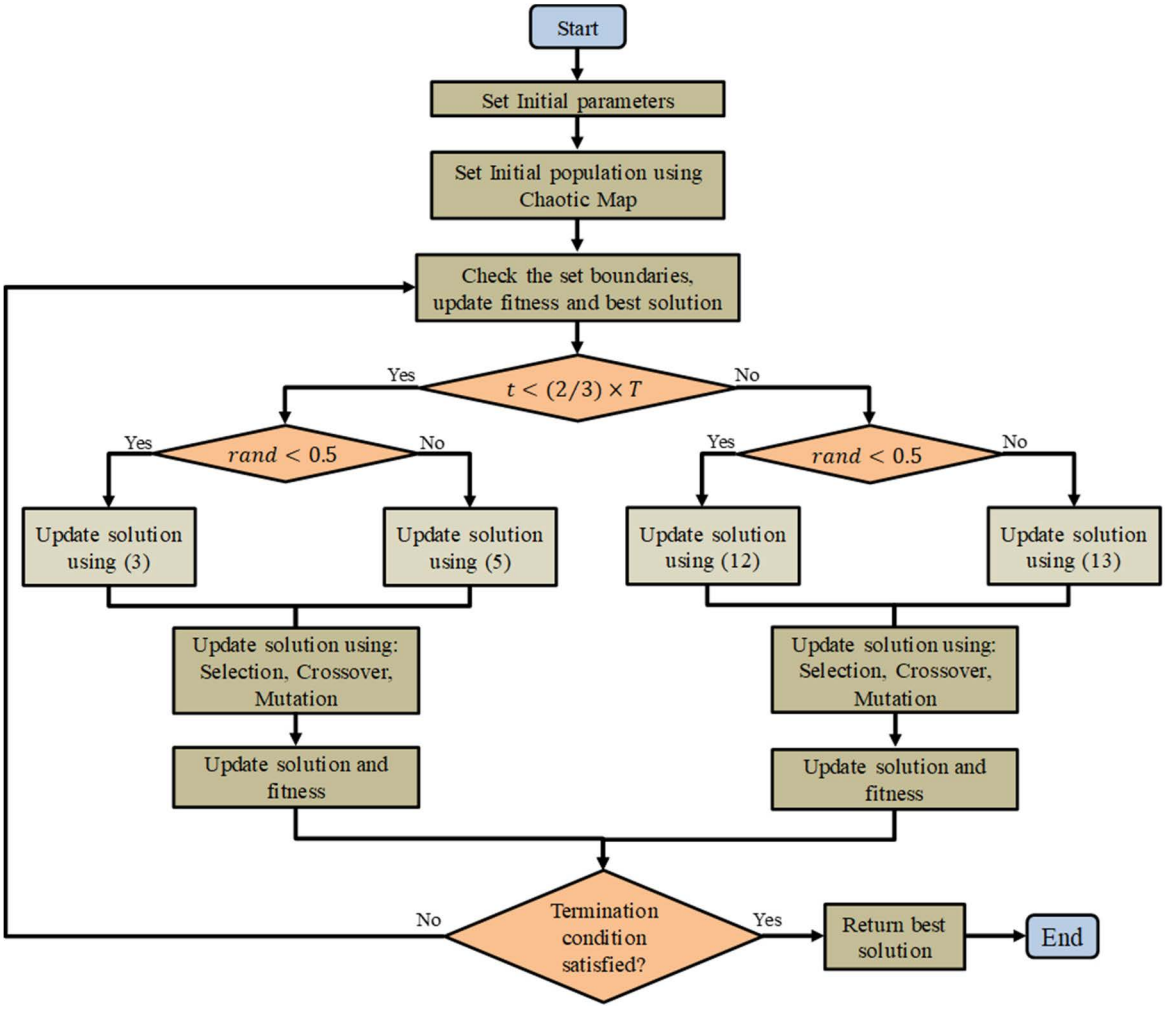



**FIGURE 2.** Diagram of process flow of CMAOE algorithm.

initial population is carried out by eleven chaotic functions, having parameters given in Table 3. By selecting different chaotic maps, the different variants of the proposed algorithm are developed. The corresponding names of different algorithms are also mentioned in Table 3 along with the chaotic map chosen for generation of initial population. For each execution of the algorithm, the initial population is computed using chaos and its corresponding entropy is also computed. For enhancing the reusability of this work, the proposed CMAOE is presented in the form of pseudo code in Figure 4.

### E. COMPLEXITY ANALYSIS OF PROPOSED METHOD

The complexity of an algorithm relies on three main processes: initialization, fitness function computation and solution updating process. The process of CMAOE involves the computation of initial population using a particular chaotic map, calculation of fitness function and updating of solution using hunting behavior of Aquila and the usual steps of evolution, like selection, crossover, mutation. The complexity analysis of initialization process which involves chaotic maps is carried out using Entropy computation in the upcoming section. Apart from initialization process, the complexity of CMAOE is $O(n \times (T \times D + 1)) + O(T \times (n \times D + n \times D + n))$. Here, $n$ is the population size, $T$ is number of generations and $D$ is the dimension size. Therefore, the overall order of computational complexity is $O(T \times n \times D)$. Since, the overall order of complexity is not disturbed. Hence, the proposed method is flexible in producing better optimum solution without the increase in complexity.

## III. EXPERIMENTAL RESULTS ON BENCHMARK FUNCTIONS

The statistical analyses are carried out for two experiments consisting of various benchmark functions. For fair comparison, the start point is same for all the algorithms. The various fitness results obtained by simulating variants of CMAOE are presented in the form of a table. The convergence speeds of various variants of CMAOE are compared in the form of figures as a function of simulation size. The coding of all the programs were carried out in Matlab R2016a and the corresponding executions were performed on a LENOVO Slim 3 with $10^{th}$ Gen Intel® Core i3 CPU @ 1.20 GHz.

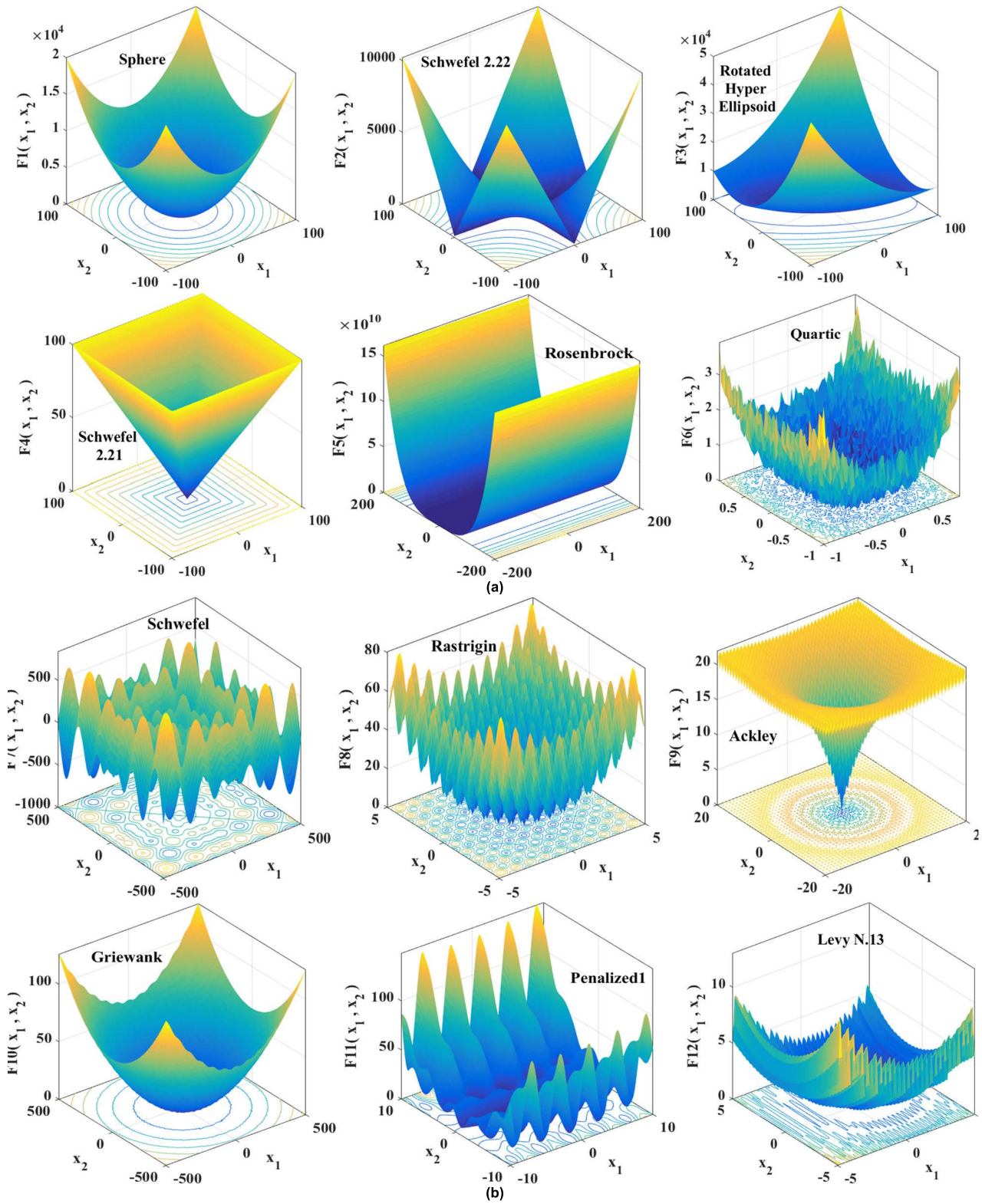



**FIGURE 3.** Benchmark functions used in this work (a) Unimodal functions (b) Multimodal functions.

*Experiment 1:* Seven unimodal functions are selected. These possess a single optimal value. The parameters for simulations are taken as: dimension size = 10, population size = 20, simulation size = 1000. Eleven variants of AO are simulated for solving Sphere function, Schwefel 2.21 function, Schwefel 2.22 function, Rotated hyper ellipsoid function, Rosenbrock function, Step function, and Quartic function.

**TABLE 3. Chaotic maps with prescribed parameters with the corresponding nomenclature of proposed CMAOE algorithms.**

| S. No. | Chaotic map | Parameters | CMAOE variant nomenclature |
|---|---|---|---|
| 1 | Henon | A=1.4, B=0.3 | CMAOE1 |
| 2 | Chebyshev | - | CMAOE2 |
| 3 | Circle | A=0.5, B=0.2 | CMAOE3 |
| 4 | Gauss | - | CMAOE4 |
| 5 | Iterative | A=0.7 | CMAOE5 |
| 6 | Logistic | A=4 | CMAOE6 |
| 7 | Piecewise | P=0.4 | CMAOE7 |
| 8 | Sine | A=4 | CMAOE8 |
| 9 | Singer | Mu=1.07 | CMAOE9 |
| 10 | Sinusoidal | A=2.3 | CMAOE10 |
| 11 | Tent | - | CMAOE11 |

*Discussions:* Table 4 presents the optimal results obtained by simulating standard AO and its eleven AO variant algorithms designed with different chaotic maps for different unimodal benchmark functions.

By observing the performance of CMAOE algorithms with different chaotic maps, following inferences are realized:

- The logistic chaotic map-based CMAOE6 outperforms the standard AO and other variants for most of the functions such as the Sphere function, Rotated hyper ellipsoid function, Rosenbrock function, and step function.
- Other chaotic maps like sine, Chebyshev, and iterative maps generate initial populations which enhances the diversification capability of standard AO for solving Quartic, Sphere, and Step functions and thus secures higher ranks from 2 to 4.
- The maps such as henon, circle, gauss, and tent in CMAOE algorithms produce mediocre optimal results for Schwefel 2.21, 2.22, and Rosenbrock functions as compared to the CMAOE algorithms which secured the higher ranks from range 1 to 4. However, they exhibit better performance than standard AO. Hence, they secure decent ranks from 5 to 8.
- Piecewise, singer and sinusoidal chaotic maps based on CMAOE algorithms display nearly similar behavior in producing the best optimal results for most of the unimodal functions.

Hence, by taking reference to standard AO performance, three categories of chaotic maps embedded in CMAOE algorithms, showing best, better, and good performance are generated. Table 5 presents the categorical distribution of chaotic maps for their selection in CMAOE algorithms for solving unimodal functions.

*Experiment 2:* In the second experiment, six multimodal functions are selected. These possess various local optima and a single global optimal value. The parameters for simulations are taken as: dimension size = 10, population size = 20, simulation size = 1000. Eleven variants of AO are simulated for solving the Rastrigin function, Schwefel function, Ackley function, Griewank function, Penalized1 function, and Levy N.13 function.

## Pseudo Code of CMAOE Algorithm

**Input:** $f_{obj}$, chaotic map for $x$ (initial population), $n$ (size of population),$T$(number of generations)
**Output:** $x_b$ (best solution),$f_b$ (best fitness)

1 Generate $x$ using chaotic map
2 Compute Entropy of population $x$
3 Initialize parameters
4 **While** (termination criteria ← not satisfied)
5 $x_b(t) \leftarrow$ store best solution
6 $f_b \leftarrow$ store best fitness
7 **for** $i = 1,2,\ldots n$
8 update $x_m(t), x, y, g_1, g_2, L(D)$
9 **if** $t \leq (2/3) \times T$
% Stage I
10 update $x \leftarrow$ using Eq. (3)
11 examine and renew fitness
% Stage II
12 update $x \leftarrow$ using Eq. (5)
13 examine and renew fitness
14 **else**
% Stage III
15 update $x \leftarrow$ using Eq. (12)
16 examine and renew fitness
% Stage IV
17 update $x \leftarrow$ using Eq. (13)
18 examine and renew fitness
19 **end if**
20 **end for**
21 compute parent solution $x$
%(using selection over $x$ obtained at line no 19)
22 generate offspring solution $x$
%(using Eq. (18))
23 compute mutated offspring solution $x$
%(using Eq. (20))
24 $x \leftarrow x \cup x_m$
25 sort fitness of new $x$
26 select $n$ solutions from new $x$
27 examine and renew fitness
28 $t = t + 1$
29 **end while**
30 Return $x_b$ and $f_b$

**FIGURE 4. Pseudocode for the proposed CMAOE; (line numbers 1-3 indicate initialization process; line numbers 4-29 indicate the recurring process of execution).**

*Discussions:* The optimal results obtained by simulating the proposed eleven CMAOE algorithms and standard AO for solving different multimodal benchmark functions are presented in Table 6.

Following presumptions are realized by observing the performance of CMAOE algorithms with different chaotic maps for solving multimodal functions:

- While solving the Rastrigin function, Ackley function, and Griewank function, all the variants of CMAOE display the identical global optimum.
- For multimodal functions like Penalized1 function, Schwefel function, and Levy N.13, CMAOE embedded with iterative, Chebyshev, logistic, and singer chaotic maps are found to be outperforming other algorithms.
- The CMAOE with gauss, circle, sine, and sinusoidal mapping exhibits decent performance in locating the global optimum and hence secure better than the AO category.

**TABLE 4. Performance analysis results present the best optimal value for unimodal benchmark functions; Ranks (R) from 1 to 12 are awarded based on their respective performance.**

| Algorithm | Sphere | | Schwefel 2.22 | | Rotated hyper ellipsoid | | Schwefel 2.21 | | Rosenbrock | | Quartic | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Best | R | Best | R | Best | R | Best | R | Best | R | Best | R |
| CMAOE1 | 9.508e-274 | 5 | 4.348e-142 | 3 | 7.903e-277 | 2 | 2.541e-136 | 4 | 7.083e-5 | 5 | 5.380e-5 | 9 |
| CMAOE2 | 3.735e-287 | 3 | 1.618e-140 | 5 | 9.864e-263 | 5 | 1.054e-115 | 8 | 6.578e-5 | 4 | 2.449e-5 | 3 |
| CMAOE3 | 2.510e-257 | 7 | 2.656e-105 | 10 | 2.625e-266 | 4 | 1.398e-140 | 3 | 5.517e-4 | 12 | 3.614e-5 | 7 |
| CMAOE4 | 2.089e-270 | 6 | 1.146e-147 | 1 | 1.932e-237 | 6 | 3.139e-124 | 7 | 4.864e-4 | 11 | 8.855e-6 | 1 |
| CMAOE5 | 3.582e-281 | 4 | 1.839e-134 | 7 | 7.833e-226 | 7 | 5.979e-136 | 5 | 6.525e-4 | 3 | 3.006e-5 | 5 |
| CMAOE6 | 9.266e-296 | 1 | 1.158e-143 | 2 | 2.998e-282 | 1 | 6.569e-135 | 6 | 4.617e-5 | 1 | 2.502e-5 | 4 |
| CMAOE7 | 2.632e-293 | 2 | 2.032e-139 | 6 | 9.617e-268 | 3 | 2.615e-100 | 10 | 5.532e-5 | 3 | 6.090e-5 | 10 |
| CMAOE8 | 1.947e-200 | 9 | 1.913e-97 | 11 | 4.543e-200 | 9 | 1.426e-144 | 1 | 1.653e-4 | 7 | 2.140e-5 | 2 |
| CMAOE9 | 1.892e-194 | 11 | 1.409e-92 | 12 | 4.721e-189 | 11 | 2.215e-98 | 11 | 3.736e-4 | 10 | 2.203e-4 | 12 |
| CMAOE10 | 7.743e-190 | 12 | 8.348e-122 | 8 | 2.438e-187 | 12 | 5.396e-98 | 12 | 2.235e-4 | 8 | 1.939e-4 | 11 |
| CMAOE11 | 4.688e-220 | 8 | 1.075e-140 | 4 | 1.205e-193 | 10 | 6.963e-141 | 2 | 1.537e-4 | 6 | 4.602e-5 | 8 |
| AO | 1.704e-199 | 10 | 1.155e-115 | 9 | 1.394e-222 | 8 | 5.236e-113 | 9 | 3.329e-4 | 9 | 3.518e-5 | 6 |

**TABLE 5. Categorical distribution of chaotic maps based on performance analysis results obtained by implementing the corresponding CMAOE for solving unimodal functions.**

| Category | Chaotic map | Inference |
|---|---|---|
| 1 | Logistic, Sine, Chebyshev, iterative | Best performance |
| 2 | Henon, Gauss, Circle, Tent | Better than standard AO |
| 3 | Piecewise, Singer, Sinusoidal | Good and/or similar to AO |

- The chaotic maps using sinusoidal, henon, tent, and piecewise distribution present almost similar results as compared to the results produced by standard AO.

As a result, the categorical distribution of chaotic maps for best, better and good performance with respect to standard AO performance is concluded. Table 7 presents the corresponding categorical distribution of chaotic maps.

*Experiment 3:* In this experiment, the study of the importance of using single stage evolutionary search in the proposed technique is performed. The Wilcoxon test [38] has been performed to compute the performance factor, *p-value*, to test the null hypothesis that the solution vectors obtained from CMAOE and from standard AO samples from continuous distributions have equal medians.

*Discussions:* The Wilcoxon test result for the sphere benchmark function is obtained as *p-value* = 1.59e-5. As per the 5 % significance level, the rejection of the null hypothesis is obtained [39]. This indicates the two sample solutions had different distributions [40]. Other benchmark functions were also studied using this test and the null hypothesis is found out to be rejected indicating that the solutions obtained are diverse and have dissimilar distributions. The Sphere benchmark function is plotted in Figure 5 along with the evolutionary search plot presenting the offspring solutions and parent solutions with respect to simulation size.

*Experiment 4:* In this experiment, the convergence speeds of designed algorithms are compared and presented in the form of a log |Best| vs simulation size plot. Four unimodal and two multimodal functions are utilized to obtain various plots.

*Discussions:* By observing convergence rate of optimal solutions obtained by simulating CMAOE algorithms, shown in Figure 6, it is concluded that the sluggishness in diversification and intensification stages of AO has been improved to large extent using the advanced CMAOE techniques.

*Experiment 5:* This experiment shows the comparative analysis of the proposed CMAOE with some of the other existing optimization algorithm. For fair comparison, the population size of 20, dimension of 10 and generations of 1000 are kept constant. For unimodal Benchmark functions, CMAOE6 secured Rank 1 for most of the functions.

Hence, it is used for comparison with other prescribed algorithm. Similarly, for multimodal Benchmark functions, CMAOE2 secured Rank 1 for most of the Benchmark functions. So, it is compared with the prescribed algorithms.

*Discussions:* The best solution obtained by the proposed algorithm is compared with that produced by the renowned optimization algorithms, like Grasshopper Optimization Algorithm (GOA), Equilibrium Optimizer (EO), Particle Swarm Optimization (PSO), Dragonfly Algorithm (DA), Ant Lion Optimizer (ALO), Grey Wolf Optimizer (GWO), Marine Predators Algorithm (MPA), Salp Swarm Algorithm (SSA), Sine Cosine Algorithm (SCA), Whale

**TABLE 6.** Performance analysis results present the best optimal value for multimodal Benchmark functions; Ranks (R) from 1 to 12 are awarded based on their respective performance.

| Algorithm | Schwefel | | Rastrigin | | Ackley | | Griewank | | Penalized1 | | Levy N.13 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Best | R | Best | R | Best | R | Best | R | Best | R | Best | R |
| CMAOE1 | -4104.6 | 11 | 0.000 | 1 | 8.882e-16 | 3 | 0.000 | 1 | 5.571e-7 | 5 | 2.181e-7 | 2 |
| CMAOE2 | -4193.3 | 2 | 0.000 | 1 | 8.881e-16 | 1 | 0.000 | 1 | 6.852e-8 | 1 | 2.863e-7 | 3 |
| CMAOE3 | -4122.7 | 10 | 0.000 | 1 | 8.880e-16 | 2 | 0.000 | 1 | 2.273e-7 | 4 | 1.406e-6 | 4 |
| CMAOE4 | -4103.1 | 12 | 0.000 | 1 | 8.881e-16 | 1 | 0.000 | 1 | 2.930e-6 | 8 | 3.407e-6 | 9 |
| CMAOE5 | -4188.4 | 1 | 0.000 | 1 | 8.884e-16 | 4 | 0.000 | 1 | 1.331e-7 | 3 | 2.323e-6 | 8 |
| CMAOE6 | -4147.8 | 7 | 0.000 | 1 | 8.885e-16 | 5 | 0.000 | 1 | 5.185e-6 | 10 | 1.952e-6 | 7 |
| CMAOE7 | -4130.3 | 9 | 0.000 | 1 | 8.881e-16 | 1 | 0.000 | 1 | 7.087e-6 | 12 | 5.919e-6 | 11 |
| CMAOE8 | -4160.5 | 6 | 0.000 | 1 | 8.883e-16 | 4 | 0.000 | 1 | 1.162e-7 | 2 | 1.756e-6 | 6 |
| CMAOE9 | -4197.4 | 3 | 0.000 | 1 | 8.883e-16 | 4 | 0.000 | 1 | 1.937e-6 | 7 | 1.371e-5 | 12 |
| CMAOE10 | -4172.4 | 4 | 0.000 | 1 | 8.886e-16 | 6 | 0.000 | 1 | 3.368e-6 | 9 | 5.862e-6 | 10 |
| CMAOE11 | -4143.5 | 8 | 0.000 | 1 | 8.881e-16 | 1 | 0.000 | 1 | 9.241e-7 | 6 | 1.419e-6 | 5 |
| AO | -4169.9 | 5 | 0.000 | 1 | 8.886e-16 | 6 | 0.000 | 1 | 6.913e-6 | 11 | 2.111e-7 | 1 |

**TABLE 7.** Categorical distribution of chaotic maps based on performance analysis results obtained by implementing the corresponding CMAOE for solving multimodal functions.

| Category | Chaotic map | Inference |
|---|---|---|
| 1 | Logistic, Iterative, Chebyshev, Singer | Best performance |
| 2 | Circle, Gauss, Sine | Better than standard AO |
| 3 | Piecewise, Henon, Sinusoidal | Good and/or similar to AO |

Optimization Algorithm (WOA) and Slime Mold Algorithm (SMA). It is observed from Table 8 variant of CMAOE is producing better performance for unimodal functions as well as for multimodal functions as compared to other algorithms.

*Experiment 6:* In this experiment, the computational time of the proposed algorithm is computed. The overall computational time of CMAOE is compared with that of other hybrid versions of standard AO. (say, running time of one instruction is $t_i$)

*Discussions:* The computational time of algorithms is computed for 1000 iterations, population size of 20 and dimensional size of 10. For fair comparison, the time complexity of other four hybrid versions of AO are computed and compared in Table 9. It is found that CMAOE has less burden on CPU computational time complexity for reaching the global optimum.

## IV. APPLICATION OF CMAOE TO AN INDUSTRIAL REAL-WORLD PROBLEM

In this section, an optimization problem of obtaining minimum supplementary cogging torque in a new surface inset type of permanent magnet synchronous motor (si-PMSM) is formulated.

In this motor, the drastic variation in cogging torque is obtained by varying inter-polar length in the rotor. The reasons for applying the proposed CMAOE algorithms to solve this problem are listed below [41], [42], [43], [44], [45], [46]:

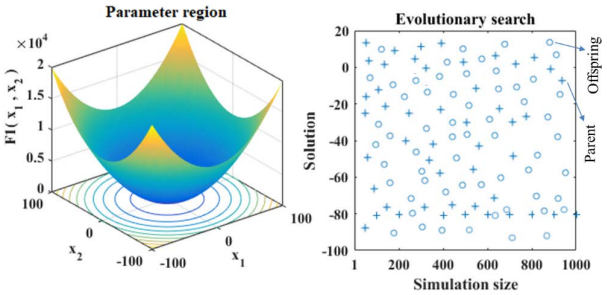



**FIGURE 5.** Solution dispersion plot obtained by standard AO (parent) and CMAOE (offspring) for sphere benchmark function.

- For the higher electromagnetic performance of si-PMSM, the necessity to suppress the cogging supplement in its torque potential is worth addressing since it procreates noisy behavior of the motor.
- The inter-polar length $l_i$ or the gap between rotor PMs in si-PMSMs is of original rotor material and induces the enhanced distortion in torque behavior if not optimized at the design stage. Hence, the dependency of cogging torque over $l_i$ is derived by the assistance of surrogacy.

Recently, the design and optimization methods done on the basis of approximation have gained thorough recognition. This approach, generally, approximates the comprehensive operations to a suitable and simpler form of models analytically. Such models are called 'Meta-models' or 'surrogate models' and the concept of constructing meta-models is

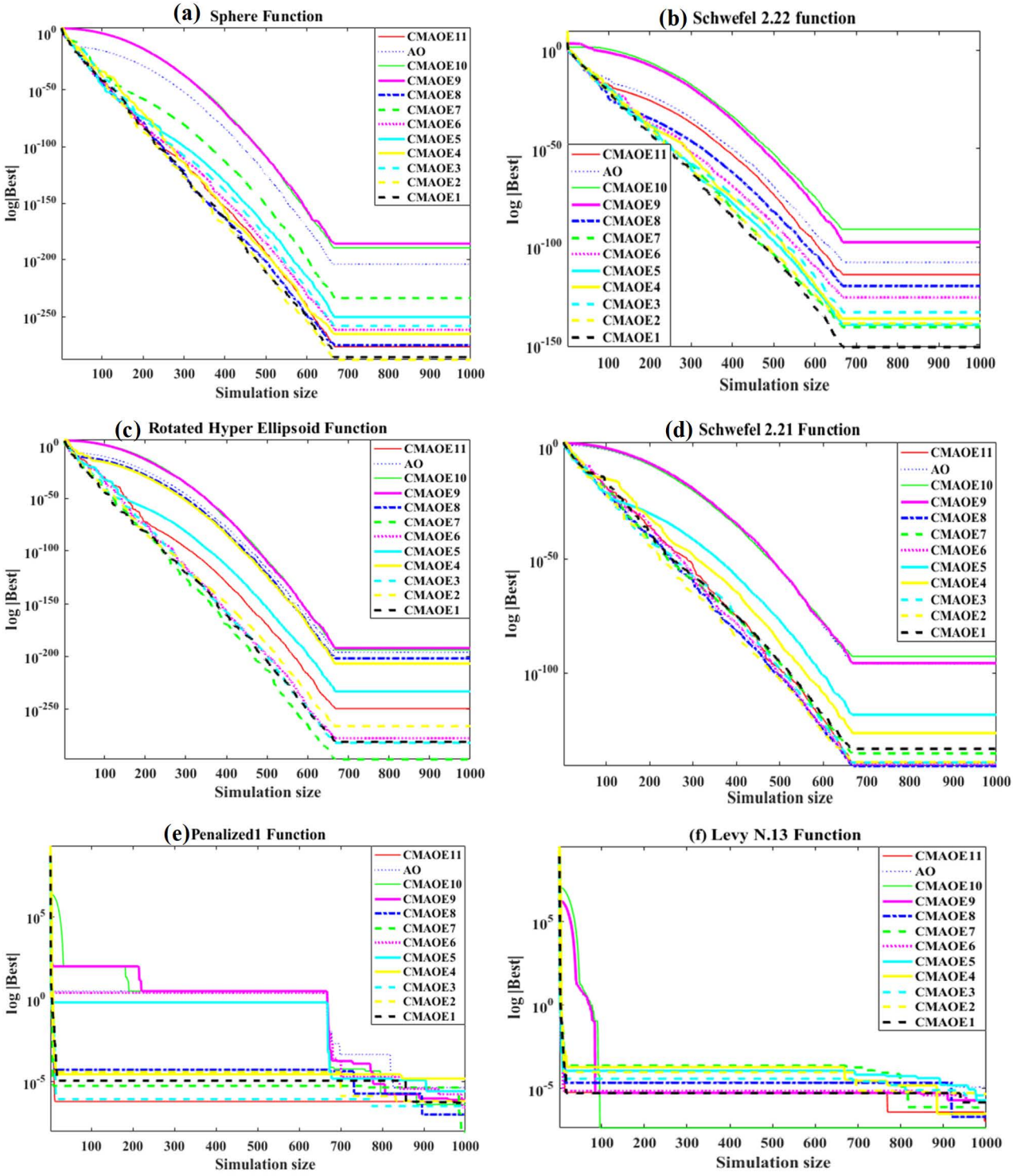



**FIGURE 6.** Convergence plots of standard AO and eleven variants of CMAOE for (a)–(d) unimodal functions and (e)–(f) multimodal functions.

called 'meta-modelling' or 'surrogate modelling'. Using the obtained meta-models, optimization procedure is applied in order to search the optimum solution. Barthelemy and Haftka [47] and Haftka *et al.* [48] has discussed about the relationship between experiments and optimization.

An extensive consolidated review is elucidated by Simpson et. al regarding meta-models by exploring different sampling methods, meta-modelling procedures, approximating models, practical systems for experiments and their applications [49].

**TABLE 8. Comparative analysis results for the best optimal solution of classical benchmark functions obtained by different algorithms.**

| Algorithms | Unimodal Functions | | | | | |
|---|---|---|---|---|---|---|
| | Sphere | Schwefel 2.22 | Rotated hyper ellipsoid | Schwefel 2.21 | Rosenbrock | Quartic |
| CMAOE6 | 9.266E-296 | 1.158E-143 | 2.998E-282 | 6.569E-135 | 4.617E-5 | 2.502E-5 |
| GOA | 4.5311E-06 | 4.2831E-02 | 7.6675E-01 | 8.7472E-02 | 3.5924E+00 | 5.4523E-02 |
| EO | 4.3340E-58 | 4.4565E-34 | 3.2828E-30 | 7.2504E-22 | 5.3001E+00 | 7.1835E-04 |
| PSO | 2.8877E-21 | 1.8509E-10 | 3.0206E-07 | 2.0935E-05 | 1.0189E-01 | 3.9634E-03 |
| DA | 9.4062E-01 | 1.1170E+00 | 4.9375E+01 | 1.5899E+00 | 1.3482E+02 | 7.3884E-03 |
| ALO | 1.3931E-08 | 2.0566E-05 | 7.7055E-01 | 1.0461E-03 | 5.9959E+00 | 1.6443E-02 |
| GWO | 3.3951E-51 | 5.2330E-29 | 5.8410E-22 | 2.3617E-16 | 6.2379E+00 | 3.9526E-04 |
| MPA | 4.8287E-31 | 1.5592E-17 | 4.3286E-15 | 2.3066E-13 | 8.7560E+00 | 3.9490E-04 |
| SSA | 4.5257E-10 | 7.8584E-06 | 5.5017E-06 | 1.6171E-05 | 7.8930E+00 | 1.3986E-02 |
| SCA | 2.2968E-15 | 2.2667E-11 | 8.7897E-06 | 4.3787E-06 | 7.2782E+00 | 1.4779E-03 |
| WOA | 1.2964E-74 | 2.3267E-54 | 1.8044E+01 | 2.1035E-03 | 6.8245E+00 | 1.9935E-03 |
| SMA | 0.0000E+00 | 4.0840E-185 | 0.0000E+00 | 4.1535E-184 | 1.3987E-01 | 6.3985E-05 |
| Algorithms | Multimodal Functions | | | | | |
| | Schwefel | Rastrigin | Ackley | Griewank | Penalized1 | Levy N.13 |
| CMAOE2 | -4193.3 | 0.000 | 8.881e-16 | 0.000 | 6.852e-8 | 2.863e-7 |
| GOA | -3.3212E+03 | 2.8854E+01 | 1.0370E-03 | 1.9581E-01 | 8.3499E-04 | 8.2498E-04 |
| EO | -3.5946E+03 | 0.0000E+00 | 4.4409E-15 | 0.0000E+00 | 5.9790E-17 | 4.0278E-16 |
| PSO | -2.5933E+03 | 1.9899E+00 | 3.2066E-10 | 1.0083E-01 | 2.0418E-21 | 8.6127E-22 |
| DA | -2.9985E+03 | 1.0558E+01 | 1.4270E+00 | 6.5156E-02 | 7.5827E-02 | 2.8877E-01 |
| ALO | -3.2226E+03 | 1.5919E+01 | 8.1753E-05 | 1.2060E-01 | 2.6276E-01 | 2.2534E-07 |
| GWO | -3.1415E+03 | 0.0000E+00 | 7.9936E-15 | 0.0000E+00 | 2.7411E-06 | 6.5501E-06 |
| MPA | -3.9514E+03 | 0.0000E+00 | 4.4409E-15 | 0.0000E+00 | 1.1688E-11 | 3.5877E-11 |
| SSA | -3.2590E+03 | 5.9697E+00 | 9.2279E-06 | 1.3262E-01 | 5.1467E-11 | 5.7806E-10 |
| SCA | -2.1001E+03 | 2.7143E-12 | 5.6642E-08 | 6.6321E-12 | 6.6012E-02 | 3.3532E-01 |
| WOA | -3.8213E+03 | 0.0000E+00 | 8.8818E-16 | 0.0000E+00 | 1.4821E-03 | 2.4993E-03 |
| SMA | -4.1898E+03 | 0.0000E+00 | 8.8818E-16 | 0.0000E+00 | 8.9676E-05 | 6.6998E-03 |

The state-of-the-art categorization of different approximation methods based on selection, sampling, and curve fitting is presented in Table 10.

A model of si-PMSM having 9 slots and 6 poles is developed having a concentrated winding configuration. The dimensions of the initial design are chosen to comply with sufficient electric vehicle powertrain conditions. Table 11 displays the model specifications.

The problem, in this work, is formulated using approaches based on finite element (FE) and regression analysis of supervised learning via surrogate modelling (SMO) concepts. The FE technique helps to collect the cogging torque response of the motor at different inter-polar lengths. The least square method is the simplest and efficient approximation methods to derive the desired relationship of input design parameters with the output response, i.e., cogging torque of the motor. Figure 7 depicts the decision variable and its variation performed to obtain ten experimental results for cogging torque.

The SMO method assists to derive the relationship between the design parameter ($l_i$) and the response of interest (i.e., supplementary cogging torque). The various FE responses for different values of $l_i$ are treated as the design of experiments (DOEs) of SMO.

In design industries, by the administration of FE analysis to complex system architecture, a steep level of accuracy is ensured while studying the influence of critical factors such as structural parameters and their respective materialistic properties on the system performance. FE modeling safeguards the user-friendly environment to encounter susceptibility and versatility of the design. For different inter-polar lengths, the corresponding FE analysis results for ten experiments are registered in Table 12.

**TABLE 9. Computational time comparison.**

| Algorithms | Total computational time = (initialization time + time to update solution + fitness calculation time) |
|---|---|
| CMAOE | $10t_i + 1000 \times 51t_i + 1000 \times 20t_i$ |
| IHAOHHO | $10t_i + 1000 \times 61t_i + 1000 \times 20t_i$ |
| AOAAO | $10t_i + 1000 \times 62t_i + 1000 \times 20t_i$ |
| AGWO | $10t_i + 20t_i + 1000 \times 41t_i + 1000 \times 20t_i$ |

**TABLE 10. Categorization of related work to meta-models.**

| Meta-model selection | Sampling | Curve fitting |
|---|---|---|
| Polynomial<br>• Linear<br>• Quadratic<br>• Higher order<br>• Least interpolating polynomial | Classic techniques<br>• Factorial/fractional<br>• Box – Behnken<br>• Plackett – Burman<br>• Central composite<br>• Alphabetical optimum | Least square regression (weighted) |
| Splines<br>• Linear<br>• Cubic<br>• Rational B splines (NURBS)<br>• Multi-variate adaptive regression splines (MARS) | Space-filling<br>• Simple grids<br>• Orthogonal arrays<br>• Latin hypercube<br>• Hammersley sequence<br>• Minimax and maximin<br>• Uniform designs | Best linear predictor |
| Gaussian process | Hybrid techniques | Best linear unbiased predictor (BLUP) |
| Kriging model | Human or random selection | Multipoint approximation (MPA) |
| Radial basis function (RBF) | Based on importance | Log likelihood |
| Hybrid model | Directional simulation | Back propagation (in ANN) |
| Support vector machine (SVM) | Adaptive methods | Entropy (for inductive learning on decision tree |
| Knowledge based or decision tree | Based on discrimination | |

Using the regression analysis through the least square approach, the coefficients of the meta-model of cogging torque is estimated using the MATLAB 2016a tool [50], [51]. Thus, the polynomial response surface (RS) model is derived. The appropriateness of the RS model is inspected via the root mean square error (RMSE) test. The lower the value of RMSE value, the better the accuracy of the RS model. It is observed that the obtained RMSE value $\approx 0.9577$ which is less than 1 since the coefficient of determination ($R^2 = 0.9991$) and adjusted coefficient of determination ($R^2_{adj} = 0.9990$) are nearly equal with the error of $\sim$0.0001. It ensures that no unwanted terms are included in the designed RS model. The real world objective problem is thus formulated as a cogging

**TABLE 11. Specifications of the si-PMSM model (HP = Horse power, RPM = Revolutions per minute, EMF = Electromotive force).**

| Specification | Value | Unit |
|---|---|---|
| Rated Output Power | 3 | HP |
| Rated Speed | 3000 | Rpm |
| Operation Frequency | 150 | Hz |
| Input Voltage | 200 | V |
| Stator Inner Diameter | 67.6 | Mm |
| Stator Outer Diameter | 110 | Mm |
| Rotor Outer Diameter | 66.6 | Mm |
| Stack Length | 50 | Mm |
| Air gap Length | 2.5 | Mm |
| Coil Turns | 30 | - |
| Inter-polar length | 15.34 | Mm |
| Electrical Steel | M400-50A | - |
| Permanent Magnet | NdFe35 | - |

segment of torque ($f_1$) denoted by *Equation (22)*. Here, $T_{out} \geq 11.18\ Nm$ is the output torque produced at the shaft by the motor which is used as one of the inequality constraints due to the fact that the optimized model should be able to produce at least 11.18 Nm of torque, produced by the initial model.

$$\begin{aligned} &\min f_1\ (l_i) = -5.664 + 8.4925 l_i - 0.4661 l_i^2 \\ &\text{subjected to : } 1.5\ mm \leq l_i \leq 15.94\ \text{mm} \\ &\qquad T_{out} \geq 11.18\ Nm \end{aligned} \tag{22}$$

## V. VALIDATION TEST OF CMAOE ALGORITHMS FOR DESIGNED PROBLEM

In this section, first the effect of entropy is evaluated for various variants of CMAOE by implementing it on *Equation (22)*. Then, a validation test is employed to compare the change in performance of the initial model and the optimized model.

Entropy measures randomness in a population statistically. It is used to characterize the texture of the input information. By normalizing each solution in $x$ (see equation (1)), the entropy of the initial population is computed in the CMAOE algorithm (see Figure 4). Following inferences were observed from this study (see Figure 8).

- The bar (l) in Figure 8 corresponds to the entropy computed by the random function used in standard AO to generate the initial population. The Henon (a), Piecewise (g), and Tent (k) maps show similar behavior as that of the bar (l) and have the entropy concentration of a lower range, i.e., $< 3$.
- The populations generated using maps Circle (c), Gauss (d), Sine (h), and Sinusoidal (j) exhibit entropy concentrations of medium-range, i.e., 3 - 5.
- The map Logistic (f) performs the best performance with respect to the highest entropy concentration of 6.8955. Maps Chebyshev (b), Iterative (e), and singer (i) also present the entropy concentration range of high values, i.e., $> 5$.

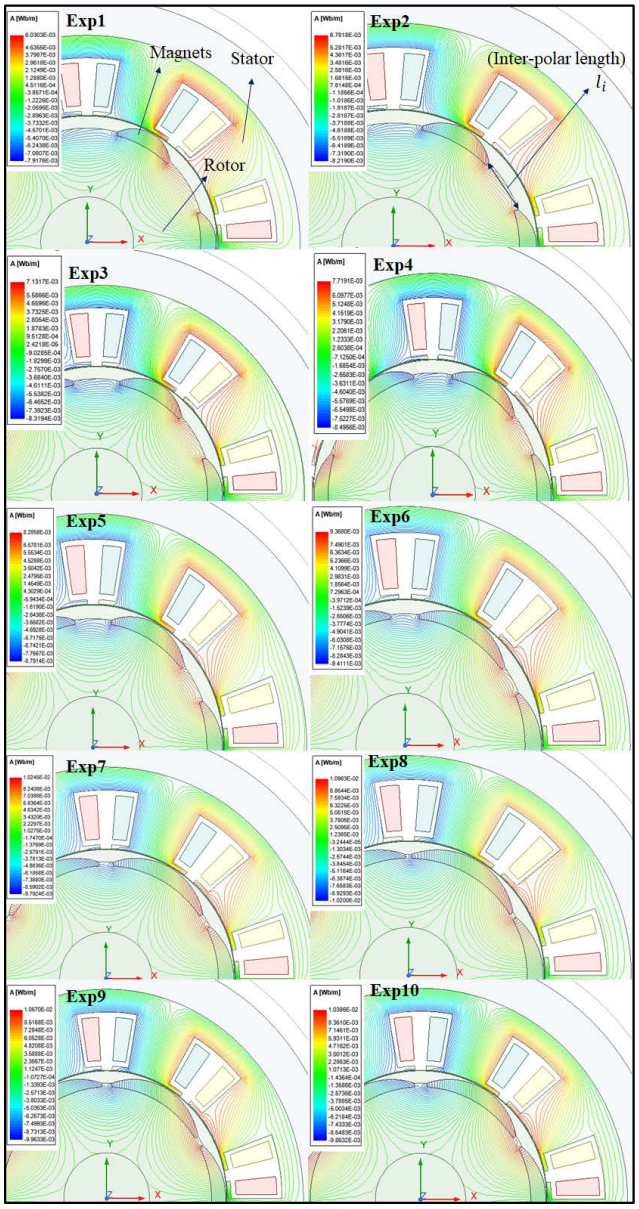



**FIGURE 7.** Depiction of design (decision) variable in response surface model.

**TABLE 12.** Training data of cogging torque, $f_1$, via FE investigations.

| Experiment (Exp) | Inter-polar length $l_i$ (mm) | $f_1$ (mNm) |
|---|---|---|
| 1 | 15.3414 | 12.6 |
| 2 | 13.1936 | 19.6 |
| 3 | 12.2731 | 28.8 |
| 4 | 10.7390 | 32.8 |
| 5 | 9.2048 | 20.7 |
| 6 | 7.7502 | 13.1 |
| 7 | 5.5847 | 9.9 |
| 8 | 4.5341 | 7.5 |
| 9 | 3.7614 | 6.2 |
| 10 | 2.2217 | 8.3 |

The optimized model shows a noticeable improvement in its electromagnetic performance with respect to the initial model, shown in Figure 9. The validation test analysis results

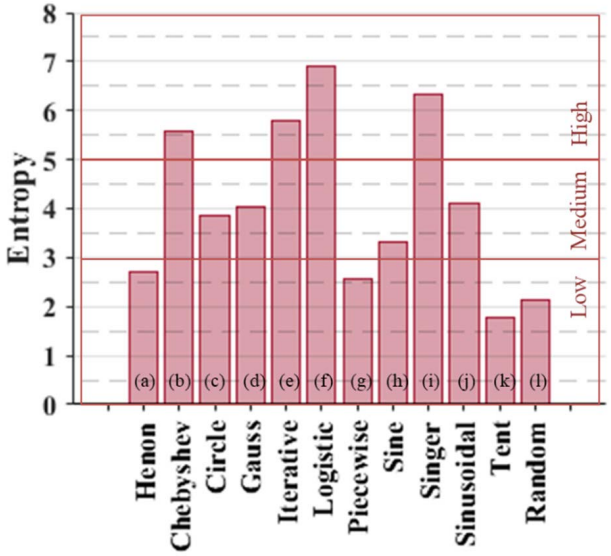



**FIGURE 8.** Bar plots of various chaotic maps as a function of entropy in the population.

**TABLE 13.** Validation test analysis results to compare results of the motor optimization problem.

| Parameter | Initial model | Optimized model | |
|---|---|---|---|
| | | AO | CMAOE |
| $l_i\ (mm)$ | 15.3 | **3.21** | 3.18 |
| $f_1(mNm)$ | 12.6 | 6.10 | 5.91 |
| $T_{out}(Nm)$ | 11.18 | 15.77 | 16.04 |

are reported in Table 13. It is observed that the cogging torque of the optimized model obtained by AO is reduced by 51.5 % while that of the model obtained by CMAOE is reduced by 53 %. It is also found that the optimized model is able to produce enhanced output torque at the shaft ($T_{out}$) by 41 % using AO while that 43.4 % using CMAOE.

## VI. DISCUSSIONS ON PERFORMANCE OF CMAOE

On the basis of source of inspiration, all the hybrid versions of Aquila Optimizers have been collected for the first time to encourage the future research. The chaotic maps are integrated to many metaheuristic algorithms like Bird Swarm algorithms with chaotic mapping [52], Chaotic Bee Colony Algorithms for global numerical optimization [53], Chaotic based Optics inspired optimization algorithms as global solution search approach [54], Chaotic League Championship algorithm [55], and Chaotic Harmony Search Algorithms [56], etc., in order to generate random numbers involved in the computations of converging solutions. However, this paper integrates the chaotic mappings to generate initial population to hybrid version of AO for the first time.

Thus, the proposed CMAOE presents the following features highlighting certain pros and cons:

- It has higher diversity in the population search region using the hybridization of AO with GA.

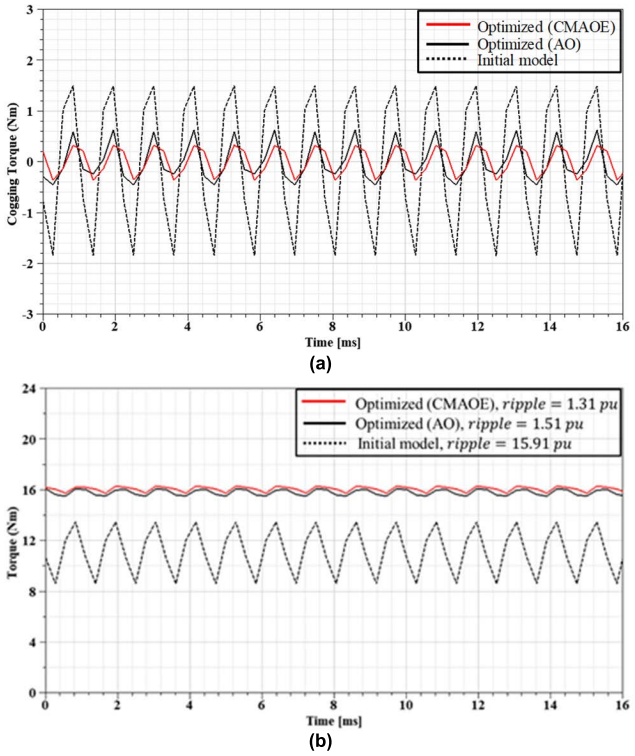



**FIGURE 9.** Electromagnetic performance comparison for initial and optimized model obtained by AO and CMAOE. (a) Cogging torque vs. time plot. (b) Output torque vs. time plot.

- Through comparison of entropy of initial population, generated via integration of chaotic maps, with that generated via stochastic method, it is found that the performance of the algorithm has been improved.
- The order of computational complexity has not been affected as compared to the standard AO.
- As compared to the other hybrid versions of AO, the CMAOE shows inferior burden on CPU computational time. However, as compared to the standard AO, the computational time of proposed CMAOE has been increased to solve certain complex multimodal functions like Rastrigin and Penalized1 etc.
- The hybridization of algorithms includes the computation of extra variables which require extra CPU memory.

The proposed CMAOE has following practical advantages:

- The high accuracy of CMAOE helps the design engineers to apply the proposed hybrid version of AO to design optimization of electric motors.
- The improvement in convergence speed of CMAOE ensures that the proposed algorithm can be applied to high constrained problems like power system optimal power flow.

The limitations of the research work are listed below:

- To use a particular chaotic map for generation of initial population in CMAOE, the complexity of the problem has to be considered. It effects the overall computational time of the algorithm [57].
- To formulate the real world problem using methods other than least square method, the feasibility ranges of the design variables have to be thoroughly analyzed.

## VII. CONCLUSION

This paper successfully illustrates the following accomplishments:

- A novel advanced version of the standard AO algorithm, obtained by incorporating chaotic mapping and single-stage evolutionary algorithm concept, is developed and successfully implemented to optimize problems efficiently.
- The proposed CMAOE technique is executed for retrieving the optimal value of various benchmark functions.
- On the basis of its performance for unimodal and multimodal functions, the chaotic maps are categorized in 'Best', 'Better', and 'Good' levels.
- For unimodal functions, CMAOE with Logistic, sine, Chebyshev and Iterative chaotic maps are found to be giving the results, hence are categorized as ''Best'' performance category.
- For multimodal functions, CMAOE with Logistic, Iterative, Chebyshev and Singer chaotic maps are giving the results falling in the ''Best'' category.
- The convergence profiles of CMAOE algorithms are found to be more efficient than the standard AO framework.
- The solution samples obtained from standard AO and CMAOE show dissimilarity in distribution which is verified using Wilcoxon null hypothesis test.
- A real-world industrial effective motor optimization problem is formulated using FE and SMO-based realizations. The analyzed results are thus obtained by both AO and CMAOE algorithms.
- It is found that the CMAOE technique produces 53% reduction in cogging torque and 43.4% increment in torque output of the motor. Whereas, standard AO produces 51.5% reduction in cogging torque and 41% increment in torque output of the motor.

Thus, this paper illustrates that the proposed CMAOE algorithms can efficiently solve variety of problems. For future works, the authors plan to examine the evolution in the average entropy. Also, the effects of multifractals can be monitored during the formation of initial populations.

## REFERENCES

[1] Y. N. Guo, J. Cheng, S. Luo, and Y. Xue, "Robust dynamic multi-objective vehicle routing optimization method," *IEEE Trans. Comput. Biol. Bioinf.*, vol. 15, pp. 1891–1901, Nov. 2018.
[2] R. Saravanan, S. Ramabalan, N. G. R. Ebenezer, and C. Dharmaraja, "Evolutionary multi criteria design optimization of robot grippers," *Appl. Soft Comput.*, vol. 9, no. 1, pp. 159–172, Feb. 2009.
[3] G. Linzzi, S. Lucidi, F. Parasiliti, and M. Villani, "Multi-objective optimization techniques for the design of induction motors," *IEEE Trans. Magn.*, vol. 39, no. 3, pp. 1261–1264, Mar. 2003.
[4] M. Helena, S. Nargiz, and S. Dragan, "Lost in optimization of water distribution systems," *Environ. Model. Softw.*, vol. 93, pp. 209–254, Jul. 2017.

[5] B. Alatas and H. Bingol, "A physics based novel approach for travelling tournament problem: Optics inspired optimization," *Inf. Technol. Control*, vol. 48, no. 3, pp. 373–388, Sep. 2019.
[6] B. Alatas and H. Bingol, "Comparative assessment of light-based intelligent search and optimization algorithms," *Light Eng.*, vol. 28, no. 6, pp. 51–59, 2020.
[7] S. Akyol and B. Alatas, "Plant intelligent based metahueristic optimization algorithms," *Artif. Intell. Rev.*, vol. 47, pp. 417–462, May 2016.
[8] K. C. Tan, T. H. Lee, and E. F. Khor, "Evolutionary algorithms with dynamic population size and local exploration for multiobjective optimization," *IEEE Trans. Evol. Comput.*, vol. 5, no. 6, pp. 565–588, Jan. 2001.
[9] A. Pratap, T. Meyarivan, S. Agarwal, and K. A. Deb, "Fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Feb. 2002.
[10] G. G. Yen and H. Lu, "Dynamic multiobjective evolutionary algorithm: Adaptive cell-based rank and density estimation," *IEEE Trans. Evol. Comput.*, vol. 7, no. 3, pp. 253–274, Dec. 2003.
[11] C. A. C. Coello, "Evolutionary multi-objective optimization: A historical view of the field," *IEEE Comput. Intell. Mag.*, vol. 1, no. 1, pp. 28–36, Feb. 2006.
[12] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms: An Introduction*. Hoboken, NJ, USA: Wiley, 2001.
[13] C. A. C. Coello, G. B. Lamont, and D. A. Van Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems* (Genetic and Evolutionary Computational Series). Norwell, MA, USA: Kluwer, 2002.
[14] M. Verma, M. Sreejeth, and M. Singh, "Application of hybrid Metaheuristic technique to study influence of core material and core trench on performance of surface inset PMSM," *Arabian J. Sci. Eng.*, vol. 47, no. 3, pp. 3037–3053, Mar. 2022, doi: 10.1007/s13369-021-06017-4.
[15] M. Verma, M. Singh, and M. Sreejeth, "Integrated Taguchi method-assisted polynomial metamodelling & genetic algorithm based optimisation of a surface inset permanent synchronous motor for performance improvement," *IET Electr. Syst. Transp.*, vol. 12, no. 1, pp. 26–35, Mar. 2022, doi: 10.1049/els2.12035.
[16] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis With Applications to Biology, Control, and Artificial Intelligence*. Cambridge, MA, USA: MIT Press, 1992.
[17] P. J. Angeline, "Genetic programming: On the programming of computers by means of natural selection," *Biosystems*, vol. 33, no. 1, pp. 69–73, Jan. 1994.
[18] I. Rechenberg, "Evolutionstrategie: Optimierungtechnischersystemenachprinzipien der biologischen evolution," Frommann-HolzboogVerlag, Stuttgart, Germany, Tech. Rep., 1973, pp. 157–160.
[19] R. Storn and K. Price, "Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces," *J. Global Optim.*, vol. 11, no. 4, pp. 341–359, 1997.
[20] L. J. Fogel, A. J. Owens, and M. J. Walsh, *Artificial Intelligence Thorough Simulated Evolution*. New York, NY, USA: Wiley, 1966.
[21] L. Giraud-Moreau and P. Lafon, "A comparison of evolutionary algorithms for mechanical design components," *Eng. Optim.*, vol. 34, no. 3, pp. 307–322, Jan. 2002.
[22] E. Rashedi, H. Nezamabadi-Pour, and S. Saryazdi, "GSA: A gravitational search algorithm," *J. Inf. Sci.*, vol. 179, no. 13, pp. 2232–2248, Sep. 2009.
[23] S. Mirjalili, S. M. Mirjalili, and A. Hatamlou, "Multi-verse optimizer: A nature-inspired algorithm for global optimization," *Neural Comput. Appl.*, vol. 27, no. 2, pp. 495–513, Aug. 2015.
[24] M. Kumar, A. J. Kulkarni, and S. C. Satapathy, "Socio evolution & learning optimization algorithm: A socio-inspired optimization methodology," *Future Gener. Comput. Syst.*, vol. 81, pp. 252–272, Dec. 2018.
[25] H. John, "Genetic algorithm," *Sci. American*, vol. 267, pp. 66–73, Mar. 1992.
[26] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Proc. 6th Int. Symp. Micro Mach. Hum. Sci. (MHS)*, Oct. 1995, pp. 39–43.
[27] L. Astudillo, P. Melin, and O. Castillo, "Introduction to an optimization algorithm based on the chemical reactions," *Inf. Sci.*, vol. 291, pp. 85–95, Aug. 2014.
[28] E. V. Altay and B. Alatas, "Music based Metaheuristic methods for constrained optimization," in *Proc. 6th Int. Symp. Digit. Forensic Secur. (ISDFS)*, Mar. 2018, pp. 1–6.
[29] L. Abualigah, A. Diabat, S. Mirjalili, M. A. Elaziz, and H. A. Gandomi, "The arithmetic optimization algorithm," *Comput. Methods Appl. Mech. Eng.*, vol. 376, no. 4, pp. 1–38, Apr. 2021.
[30] B. Alatas, "Sports inspired computational intelligence algorithms for global optimization," *Artif. Intell. Rev.*, vol. 52, no. 3, pp. 1579–1627, Oct. 2017.
[31] M. Ghasemi, I. F. Davoudkhani, E. Akbari, A. Rahimnejad, S. Ghavidel, and L. Li, "A novel and effective optimization algorithm for global optimization and its engineering applications: Turbulent flow of water-based optimization (TFWO)," *Eng. Appl. Artif. Intell.*, vol. 92, Jun. 2020, Art. no. 103666.
[32] K. Chen, E. Fun, and T. Heng, "Multi-objective evolutionary algorithms and applications," in *Advanced Information and Knowledge Proceedings*. Springer, 2006.
[33] M. Angelova and T. Pencheva, "Tuning genetic algorithm parameters to improve convergence time," *Int. J. Chem. Eng.*, vol. 2011, pp. 1–7, Sep. 2011.
[34] G. Karafotias, M. Hoogendoorn, and A. E. Eiben, "Parameter control in evolutionary algorithms: Trends and challenges," *IEEE Trans. Evol. Comput.*, vol. 19, no. 2, pp. 167–187, Feb. 2015.
[35] R. R. Sharapov, "Genetic algorithms: Basic ideas, variants and analysis," in *Vision Systems: Segmentation and Pattern Recognition*, G. Obinata and A. Dutta, Eds. IntechOpen, 2007.
[36] A. Leontitsis, "Chaotic systems toolbox: Analysis of chaotic systems," MATLAB, Mathworks, Natick, MA, USA, Tech. Rep., 2004.
[37] S. S. Sathya and M. V. Radhika, "Convergence of nomadic genetic algorithm on benchmark mathematical functions," *Appl. Soft Comput.*, vol. 13, no. 5, pp. 2759–2766, May 2013.
[38] R. Craig, "An empirical study on Wilcoxon signed rank test," *J. Negros Oriental State Univ.*, 2018.
[39] E. Slud and L. J. Wei, "Two-sample repeated significance tests based on the modified Wilcoxon statistic," *J. Amer. Stat. Assoc.*, vol. 77, no. 380, pp. 862–868, Mar. 1982.
[40] N. Feltovich, "Nonparametric tests of difference in medians: Comparison of Wilcoxon–Mann–Whitney and robust rank order tests," *Exp. Econ.*, vol. 6, pp. 273–297, Feb. 2003.
[41] Z. Zhang, C. Xia, H. Wang, and T. Shi, "Analytical field calculation and analysis of surface inset permanent magnet machines with high saliency ratio," *IEEE Trans. Magn.*, vol. 52, no. 12, pp. 1–12, Dec. 2016, doi: 10.1109/TMAG.2016.2599147.
[42] K. V. Vidvanandan, "Overview of electric and hybrid vehicles," NTPC, Tech. Rep., Mar. 2018.
[43] K.-C. Kim, "A novel method for minimization of cogging torque and torque ripple for interior permanent magnet synchronous motor," *IEEE Trans. Magn.*, vol. 50, no. 2, pp. 793–796, Feb. 2014, doi: 10.1109/TMAG.2013.2285234.
[44] D. Wang, X. Wang, and S.-Y. Jung, "Cogging torque minimization and torque ripple suppression in surface-mounted permanent magnet synchronous machines using different magnet widths," *IEEE Trans. Magn.*, vol. 49, no. 5, pp. 2295–2298, Jan. 2013.
[45] G. Feng, C. Lai, and N. C. Kar, "An analytical solution to optimal stator current design for PMSM torque ripple minimization with minimal machine losses," *IEEE Trans. Ind. Electron.*, vol. 64, no. 10, pp. 7655–7665, Aug. 2017.
[46] W. Zhao, X. Wang, C. Gerada, H. Zhang, C. Liu, and Y. Wang, "Multi-physics and multi-objective optimization of a high speed PMSM for high performance applications," *IEEE Trans. Magn.*, vol. 54, no. 11, pp. 1–5, Nov. 2018, doi: 10.1109/TMAG.2018.2835504.
[47] J. F. M. Barthelemy and R. Haftka, "Approximation concepts for optimal structural design: A review," *Struct. Optim.*, vol. 5, pp. 129–144, Sep. 1993.
[48] R. T. Haftka, E. P. Scott, and J. R. Cruz, "Optimization and experiments: A survey," *Appl. Mech. Rev.*, vol. 51, no. 7, pp. 435–448, Jul. 1998.
[49] T. W. Simpson, J. Peplinski, P. N. Koch, and J. K. Allen, "Metamodels for computer-based engineering design: Survey and recommendations," *Eng. Comput.*, vol. 17, no. 2, pp. 129–150, Jul. 2001.
[50] J. Gong, F. Gillon, J. T. Canh, and Y. Xu, "Proposal of a Kriging output space mapping technique for electromagnetic design optimization," *IEEE Trans. Magn.*, vol. 53, no. 6, pp. 1–4, Jun. 2017, doi: 10.1109/TMAG.2017.2666263.
[51] T. Kutsuta, S. Wakao, Q. Yamashita, and Y. Fujishima, "Design optimization of permanent magnet synchronous motor by response surface methodology," *J. Appl. Phys.*, vol. 91, pp. 8305–8307, Sep. 2002.
[52] E. V. Altay and B. Alatas, "Bird swarm algorithms with chaotic mapping," *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 1373–1414, Apr. 2019.
[53] B. Alatas, "Chaotic bee colony algorithms for global numerical optimization," *Expert Syst. Appl.*, vol. 37, no. 8, pp. 5682–5687, Aug. 2010.

[54] H. Bingol and B. Alatas, "Chaos based optics inspired optimization algorithm as global solution search approach," *Chaos, Solitons Fractals*, vol. 141, pp. 1–12, Dec. 2020.
[55] H. Bingol and B. Alatas, "Chaotic league championship algorithms," *Arabian J. Sci. Eng.*, vol. 41, no. 12, pp. 5123–5147, Dec. 2016, doi: 10.1007/s13369-016-2200-9.
[56] B. Alatas, "Chaotic harmony search algorithms," *Appl. Math. Comput.*, vol. 216, no. 3, pp. 2687–2699, 2010.
[57] T. S. L. V. Ayyarao, N. S. S. Ramakrishna, R. M. Elavarasan, N. Polumahanthi, M. Rambabu, G. Saini, B. Khan, and B. Alatas, "War strategy optimization algorithm: A new effective Metaheuristic algorithm for global optimization," *IEEE Access*, vol. 10, pp. 25073–25105, 2022.

**MONIKA VERMA** (Student Member, IEEE) received the B.Tech. degree in electrical engineering from the G. B. Pant University of Agriculture and Technology, Uttarakhand, India, in 2013, and the M.Tech. degree in power electronics and drives from the Vellore Institute of Technology, India, in 2016. She is currently pursuing the Ph.D. degree in electrical engineering with Delhi Technological University, India. She is an Associative Member of the Institution of Engineering and Technology (IET).

**MINI SREEJETH** (Senior Member, IEEE) was born in Kerala, India. She received the B.Tech. degree from Mahatma Gandhi University, Kerala, the M.Tech. degree from Calicut University, Kerala, and the Ph.D. degree from the University of Delhi, New Delhi, India. She joined the Delhi College of Engineering (now DTU), in 2007, as a Senior Lecturer, where she is currently working as a Professor with the Department of Electrical Engineering. Her research interests include power electronics, modeling, control, and the operation of drives. She is a Life Member of the Indian Society for Technical Education, New Delhi, India, and a Senior Member of WIE.

**MADHUSUDAN SINGH** (Senior Member, IEEE) received the Ph.D. degree from the University of Delhi, New Delhi, India. He is currently a Professor with the Department of Electrical Engineering, Delhi Technological University. His research interests include the areas of modeling and analysis of electrical machines, voltage control aspects of self-excited induction generators, and power electronics and drives. He is a fellow of the Institution of Engineers (IE), India, and the Institution of Electronics and Telecommunication Engineers, New Delhi.

**THANIKANTI SUDHAKAR BABU** (Senior Member, IEEE) received the B.Tech. degree from Jawaharlal Nehru Technological University, Ananthapur, India, in 2009, the M.Tech. degree in power electronics and industrial drives from Anna University, Chennai, India, in 2011, and the Ph.D. degree from VIT University, Vellore, India, in 2017. He is currently working as a Postdoctoral Researcher with the Department of Electrical Power Engineering, Institute of Power Engineering, Universiti Tenaga Nasional (UNITEN), Malaysia. He has published more than 60 research articles in various renowned international journals. His research interests include design and implementation of solar PV systems, renewable energy resources, power management for hybrid energy systems, storage systems, fuel cell technologies, electric vehicle, and smart grid. He has been acting as an Editorial Board Member and a Reviewer for various reputed journals, such as the IEEE and IEEE Access, IET, Elsevier, and Taylor & Francis.

**HASSAN HAES ALHELOU** (Senior Member, IEEE) received the B.Sc. degree (Hons.) from Tishreen University, in 2011, and the M.Sc. and Ph.D. degrees (Hons.) from the Isfahan University of Technology (IUT), Iran. He was with the School of Electrical and Electronic Engineering, University College Dublin (UCD), Dublin, Ireland, from 2020 to 2021, and with IUT. He is currently a Professor and a Faculty Member with Tishreen University, Syria, and a Consultant with Sultan Qaboos University (SQU), Oman. He is also with the Department of Electrical and Computer Systems Engineering, Monash University, Clayton, VIC, Australia. He has published more than 200 research papers in high-quality peer-reviewed journals and international conferences. His research papers received more than 3000 citations with H-index of 29 and i-index of 67. He authored/edited 15 books published in reputed publishers, such as Springer, IET, Wiley, Elsevier, and Taylor & Francis. His major research interests include renewable energy systems, power systems, power system security, power system dynamics, power system cybersecurity, power system operation, control, dynamic state estimation, frequency control, smart grids, micro-grids, demand response, and load shedding. He has participated in more than 15 international industrial projects over the globe. He was included in the 2018 and 2019 Publons and Web of Science (WoS) list of the top 1% best reviewer and researchers in the field of engineering and cross-fields over the world. He was a recipient of the Outstanding Reviewer Award from many journals, such as *Energy Conversion and Management* (ECM), *ISA Transactions*, and *Applied Energy*. He was a recipient of the Best Young Researcher in the Arab Student Forum Creative among 61 researchers from 16 countries at Alexandria University, Egypt, in 2011. He also received the Excellent Paper Award 2021/2022 from IEEE *CSEE Journal of Power and Energy Systems* (SCI IF: 3.938; Q1). He serves as an Editor for a number of prestigious journals, such as IEEE Systems Journal, *Computers and Electrical Engineering* (CAEE-Elsevier), *The Journal of Engineering* (IET), and *Smart Cities*. He has also performed more than 800 reviews for high prestigious journals, including IEEE Transactions on Power Systems, IEEE Transactions on Smart Grid, IEEE Transactions on Industrial Informatics, IEEE Transactions on Industrial Electronics, *Energy Conversion and Management*, *Applied Energy*, and the *International Journal of Electrical Power & Energy Systems*.

# Circuit Complexity in an interacting quenched Quantum Field Theory

Sayantan Choudhury[1],* Rakshit Mandish Gharat[2],† Saptarshi Mandal[3],‡ and Nilesh Pandey[4]§
[1] *Centre For Cosmology and Science Popularization (CCSP), SGT University, Gurugram, Delhi- NCR, Haryana- 122505, India,*
[2] *Department of Physics, National Institute of Technology Karnataka, Surathkal, Karnataka-575025, India,*
[3] *Department of Physics, Indian Institute of Technology Kharagpur, Kharagpur-721302, India, and*
[4] *Department of Applied Physics, Delhi Technological University, Delhi-110042, India.*

In this work, we explore the effects of a quantum quench on the circuit complexity for a quenched quantum field theory having weakly coupled quartic interaction. We use the invariant operator method, under a perturbative framework, for computing the ground state of this system. We give the analytical expressions for specific reference and target states using the ground state of the system. Using a particular cost functional, we show the analytical computation of circuit complexity for the quenched and interacting field theory. Further, we give a numerical estimate of circuit complexity with respect to the quench rate, $\delta t$ for two coupled oscillators. The parametric variation of the unambiguous contribution of the circuit complexity for an arbitrary number of oscillators has been studied with respect to the dimensionless parameter $(t/\delta t)$. We comment on the variation of circuit complexity for different values of coupling strength, different number of oscillators, and even in different dimensions.

arXiv:2209.03372v1 [hep-th] 7 Sep 2022

## I. Introduction

Recently, many of the works in theoretical physics are concerned with the confluence of various ideas in quantum many-body physics, quantum information theory and high energy physics [1–7]. Some of these works involved studying entanglement entropy in quantum systems, generally having a dynamical background [8–19]. Subsequently, it was discovered that entanglement entropy, although playing an important role in black-hole physics, is not sufficient to probe the interior of black holes. Susskind and collaborators, introduced a new measure in relation to the thermalisation properties of black holes, which was termed Quantum Computational Complexity (QCC) [20–27]. Lately, QCC has been studied widely in high energy physics. [28–83]. This stirred up exploring the QCC in the form of circuit complexity, in the context of free-field theories[84–86]. On the other hand, circuit complexity was computed for a weakly interacting field theory in [87].

Also, most of the recent works in many-body physics deal with dynamical systems where a time-dependent parameter is varied suddenly or very slowly. This is commonly refereed to as a Quantum Quench. The quench protocol seemingly drives any out of equilibrium system and thermalises it[88–90]. Entanglement , in the context of quenched systems has been discussed in [18, 19, 91–98]. Recently, we have studied entanglement in a quenched two-body coupled oscillator system with a quartic coupling in [99]. Furthermore, circuit complexity for quenched systems was computed in [100, 101]. The effects of quenches in quantum systems have been explored experimentally for cold atoms in [102–111].

For dynamical systems, the time-dependent Schrodinger equation can be solved using Lewis-Resenfield invariant-operator method[112]. Lewis Resenfield method was used to determine the time-dependent eigenstates in [113–117]. Further, one can consider adiabatic evolution of time-dependent parameters[118, 119] and use invariant-operator method to compute time-independent perturbative corrections to the eigenstates [120]. The exact form of time-dependent parameters in such eigenstates can be evaluated by solving the Ermakov-Milne-Pinney [121–123] equation in Mathematica.

Motivated by all these ideas, in this work, we:

- start with discretising the Hamiltonian for an interacting Quantum Field Theory (QFT) with quartic coupling, on a finite-sized lattice, having a quenched frequency by using the most common protocol in the literature.
- use invariant operator method, in a perturbative framework, to fix the reference and target states which are chosen as the ground state for an interacting field theory (with quartic coupling).
- compute the circuit complexity for quenched-coupled oscillators by using Nielsen's geometric approach [124–128] and choosing a specific cost function by modifying the results of [87].
- numerically evaluate the circuit complexity and comment on its parametric variation, for different cases

The organisation of the paper is as follows:

- Discretising, a Quantum Field Theory (QFT) with quartic interaction, on a lattice, we decouple the

* sayantan_ccsp@sgtuniversity.org, sayanphysicsisi@gmail.com
† rakshitmandishgharat.196ph018@nitk.edu.in
‡ saptarshijhikra@gmail.com
§ nilesh911999@gmail.com

Hamiltonian using Fourier modes in section II. Evidently, the decoupled Hamiltonian refers to that of $N$ coupled oscillators having a quartic perturbative coupling. The frequency of these oscillators is quenched by choosing a particular protocol.

- In section III, we use invariant operator method to compute the time-dependent ground states and also the first-order perturbative corrections to the ground state, for the quenched Hamiltonian.
- Using the ground, we fix a specific reference and target state in section IV. By using a particular cost function we compute the circuit complexity for the chosen reference and target states. The continuum limit of circuit complexity is also evaluated.
- In section V, we numerically evaluate circuit complexity for different sets of parameters and comment on the dynamical behavior of the circuit complexity in three different regimes.
- Section VI encapsulates the conclusions we draw from the results obtained in this work.

## II. The Setup and the Quench protocol

As one of the first steps towards understanding circuit complexity in the context of quenched interacting field theory, in this section we begin with a scalar field theory with $\hat{\lambda}\phi^4$ interaction term. We regulate this QFT by placing it on a lattice. Once discretized the Hamiltonian evidently represents a family of $N$ coupled anharmonic oscillators. We transform the original coordinates to normal modes to decouple the Hamiltonian. This helps us to compute the eigenstates for the system in a much simpler way, which is discussed in the upcoming section. We follow the notations used in [87, 129] to facilitate an easy comparison. Further we mention the time-dependent quench profile chosen to be the frequency of these oscillators.

The Hamiltonian for a scalar field theory with a $\hat{\lambda}\phi^4$ interaction is given by [129],

$$\mathcal{H} = \frac{1}{2}\int d^{d-1}x\Bigg[\pi(x)^2 + (\nabla\phi(x))^2 + m^2\phi(x)^2 + \frac{\hat{\lambda}}{12}\phi(x)^4\Bigg] . \tag{1}$$

Here $d$ is the space-time dimensions. We assume that the coupling $\hat{\lambda} << 1$, so that we can work in a perturbative framework. This theory can be discretized on a $d-1$ dimensional lattice, which is characterised by lattice spacing, $\delta$. It can be shown that, the discretized Hamiltonian becomes,

$$\mathcal{H} = \frac{1}{2}\sum_{\vec{n}}\Bigg\{\frac{\pi(\vec{n})^2}{\delta^{d-1}} + \delta^{d-1}\Bigg[\frac{1}{\delta^2}\sum_i\left(\phi(\vec{n}) - \phi\left(\vec{n}-\hat{x}_i\right)\right)^2 + m^2\phi(\vec{n})^2 + \frac{\hat{\lambda}}{12}\phi(\vec{n})^4\Bigg]\Bigg\}. \tag{2}$$

Here $\vec{n}$ denotes the spatial location of the points on lattice and $\hat{x}_i$ represent the unit vectors along the lattice. Further, we introduce the following substitutions to simplify the form of the Hamiltonian:

$$\begin{aligned} \hat{X}(\vec{n}) &= \delta^{d/2}\phi(\vec{n}), & \hat{P}(\vec{n}) &= \pi(\vec{n})/\delta^{d/2}, \\ M &= \frac{1}{\delta}, & \omega &= m, \\ \eta &= \frac{1}{\delta}, & \lambda &= \frac{\hat{\lambda}}{24}\delta^{-d}, \end{aligned} \tag{3}$$

where $\omega$ represents the frequency of individual oscillators and $\eta$ denotes inter-mass coupling. After these substitutions we get,

$$\mathcal{H} = \sum_{\vec{n}}\Bigg\{\frac{\hat{P}(\vec{n})^2}{2M} + \frac{1}{2}M\Bigg[\omega^2\hat{X}(\vec{n})^2 + \eta^2\sum_i\left(\hat{X}(\vec{n}) - \hat{X}\left(\vec{n}-\hat{x}_i\right)\right)^2 + 2\lambda\hat{X}(\vec{n})^4\Bigg]\Bigg\}. \tag{4}$$

The above Hamiltonian, in Eq.(4) represents a family of infinite coupled anharmonic oscillators. Setting $M = 1$, for simplicity the above Hamiltonian can be reformulated to that of $N$ coupled oscillators with quartic perturbation:

$$H = \frac{1}{2}\sum_{a=0}^{N-1}\left[p_a^2 + \omega^2 x_a^2 + \eta^2(x_a - x_{a+1})^2 + 2\lambda x_a^4\right]. \tag{5}$$

We use normal mode coordinates as dicrete Fourier transform of the original coordinates, given by:

$$x_a = \frac{1}{\sqrt{N}}\sum_{k=0}^{N-1}\exp\left[i\frac{2\pi a}{N}k\right]\tilde{x}_k \tag{6}$$

$$p_a = \frac{1}{\sqrt{N}}\sum_{k=0}^{N-1}\exp\left[i\frac{2\pi a}{N}k\right]\tilde{p}_k \tag{7}$$

The Hamiltonian of Eq.(5) can be rewritten in normal modes as:

$$H = \frac{1}{2}\sum_{a=0}^{N-1}\left[p_a^2 + \omega^2 x_a^2 + \eta^2(x_a - x_{a+1})^2 + 2\lambda x_a^4\right] \quad . \tag{8}$$

We split the above Hamiltonian into two parts,

$$H = H_k + H'_{\phi^4} \quad , \tag{9}$$

where,

$$H_k = \frac{1}{2}\sum_{k=0}^{N-1}\left[|\tilde{p}_k|^2 + \omega_k^2|\tilde{x}_k|^2\right], \tag{10}$$

denotes the unperturbed (free) Hamiltonian which can be decoupled for each of the $N$ oscillators. Here,

$$\omega_k^2 = \omega^2 + 4\eta^2\sin^2\left(\frac{\pi k}{N}\right), \tag{11}$$

denotes the freqency for each of the $N$ oscillators. The exact form of the eigenstates for the unperturbed Hamultonian of Eq.(10) has been computed in the subsection III A. On the other hand the $\lambda\phi^4$ perturbation term in the Hamiltonian of Eq.(9) can be dealt with by transforming the form of perturbations in normal modes:

$$H'_{\phi^4} = \frac{\lambda}{N}\sum_{k_1,k_2,k_3=0}^{N-1}\tilde{x}_\alpha\tilde{x}_{k_1}\tilde{x}_{k_2}\tilde{x}_{k_3}\ ; \qquad \alpha = N - k_1 - k_2 - k_3 \bmod N \tag{12}$$

The contribution of the above Hamiltonian in Eq.(12) is evaluated by approximating the first order correction to the eigenstates of unperturbed Hamiltonian by employing the use of time-independent perturbation theory in the subsection III B.

We now consider the frequency $\omega$ in, Eq.(11) as a time-dependent quench profile. One of the most common quench profiles used in literature [130, 131] is given by:

$$\omega^2(t/\delta t) = \omega_0^2\left[\tanh^2\left(\frac{t}{\delta t}\right)\right]. \tag{13}$$

Here $\omega_0$ can be considered as a free parameter and $\delta t$ measures the quench rate. We choose this particular quench profile chosen since it admits an exact solution for the mode functions given in [130]. Note that this profile attains a constant value at very early and late time. Also, for this chosen form of quench profile, the dynamical changes in the system occur in the time window $[-\delta t, \delta t]$. We will set $t/\delta t = T$ and $\omega_0 = 1$. The respective frequencies in the normal mode basis take the following form,

$$\omega_k = \sqrt{\omega(T)^2 + 4\eta^2\sin^2\left(\frac{\pi k}{N}\right)}, \tag{14}$$

where $\omega(T)$ is the quench profile in Eq.(13) and $k$ runs from 0 to $N-1$.

As the frequency of each oscillator now depends on time, the unperturbed Hamiltonian is evidently time-dependent. We employ the use of invariant operator method to compute the exact form of the unperturbed Hamiltonian. We emphasise that the perturbed Hamiltonian is not time-dependent and hence can be used as a time-independent perturbation applied to $N$ coupled oscillators. Using the ground state of total Hamiltonian of Eq.(8) we construct the reference as well as target states which are further used to evaluate the circuit complexity of this interacting quench model.

## III. Constructing Wave function for a $\phi^4$ quench model

In this section our prime objective is to derive an analytical expression for the eigenstates of Hamiltonian in Eq.(8), by using Lewis-Resenfield invariant opertor method and approximate it to the first order perturbative correction. The expression for eigenstates of decoupled and unperturbed Hamiltonian in Eq.(10) is derived using invariant operator method, in the subsection III A. The first order perurbative correction to the ground state of the decoupled Hamiltonian, is derived in the subsection III B.

### A. Eigenstates and Eigenvalues for unperturbed Hamiltonian

As shown earlier, in the normal mode basis, the unperturbed Hamiltonian, of Eq.(10), for $N$ oscillators decouples. The wavefunction for $N$ oscillators is then a product of eigenstates for each decoupled Hamiltonian, of Eq.(10).

$$\psi_{0,0,\cdots 0}\left(\tilde{x}_0, \cdots \tilde{x}_{N-1}, T\right) = \psi_{n_1}(\tilde{x}_1, T)\psi_{n_2}(\tilde{x}_2, T). \tag{15}$$

Here $T$ denotes the time-dependence of eigenstates emerging due to the quenched frequency from Eq.(13). We deal with these time-dependent eigenstates by employing the use of invariant operator method, closely following the prescription of [132]. Below, we have briefly mentioned the steps one can follow to compute analytical expression for eigenstates of a quenched Hamiltonian of Eq.(10). Note that we have suppressed the time-dependence of all the parameters to have simplicity in the notations.

First, we define the creation ($a_k^\dagger$) and annihilation ($a_k$) operators given by,

$$a_k = \frac{1}{\sqrt{2\dot{\gamma}_k}}\left[\dot{\gamma}_k\left(1 - i\frac{\dot{\rho}_k}{\rho_k\dot{\gamma}_k}\right)\tilde{x}_k + i\tilde{p}_k\right]$$
$$a_k^\dagger = \frac{1}{\sqrt{2\dot{\gamma}_k}}\left[\dot{\gamma}_k\left(1 + i\frac{\dot{\rho}_k}{\rho_k\dot{\gamma}_k}\right)\tilde{x}_k - i\tilde{p}_k\right]. \tag{16}$$

Here, $k = 0, 1, \cdots, N-1$. Further, $\gamma_k$ and $\rho_k$ are time-dependent factors while $\dot{\gamma}_k = \partial_T\gamma_k$, $\dot{\rho}_k = \partial_T\rho_k$ and $\ddot{\rho}_k = \partial_T^2\rho_k$. One can show that the operators in Eq.(16) satisfy the commutation relation $[a_i, a_j^\dagger] = \delta_j^i$. We fix the time-dependent factor $\rho_k$ as the solution to the Ermakov-Milne-Pinney equation for each oscillator,

$$\ddot{\rho_k} + \omega_k^2\rho_k = 0, \tag{17}$$

where $\omega_k$ denotes frequency for each coupled oscillator Eq.(14). If we define,

$$\alpha_k = \omega_0^2 + \frac{4\eta^2\left(\sin^2\left(\frac{\pi i}{N}\right)\right)}{\tanh^2(T)}, \tag{18}$$

for, $k = 0, 1, \cdots, N-1$, then using Eq.(13) one can rewrite Eq.(17) as:

$$\ddot{\rho_k} + \alpha_k \tanh^2(T)\rho_k = 0. \tag{19}$$

Here $k = 0, 1, \cdots, N-1$. We assume that the solution to Eq.(19) is of the form:

$$\rho_k(t, \delta t) = c_1\epsilon_k^1(t, \delta t) + c_2\epsilon_k^2(t, \delta t). \tag{20}$$

Here $c_1$ and $c_2$ represent numerical constants. On the other hand $\epsilon_k^1$ and $\epsilon_k^2$ are two complex valued coefficients for each $k = 0, \cdots, N-1$. However we can consider only the term with, $\epsilon_k^1$ as one of the solutions to Eq.(19), by setting $c_2 = 0$. Using Mathematica one can compute the exact form of $\epsilon_k^1$, following the steps shown in [99], which is mentioned below:

$$\epsilon_k^1 = \left(e^{\frac{2t}{\delta t}}\right)^{-\frac{1}{2}i\delta t\alpha_k}\left(e^{\frac{2t}{\delta t}}+1\right)^{\frac{1}{2}\left(\sqrt{1-4\delta t^2\alpha_k^2}+1\right)}\ {}_2F_1\left(\frac{1}{2}\left(\sqrt{1-4\delta t^2\alpha_k^2}+1\right), \frac{1}{2}\left(\sqrt{1-4\delta t^2\alpha_k^2}-2i\delta t\alpha_k^2+1\right); 1-i\delta t\alpha_k^2; -e^{\frac{2t}{\delta t}}\right), \tag{21}$$

where, ${}_2F_1$ denotes the hyper-geometric function. These complex valued solutions can be rewritten in the form $\epsilon_k^1 = \varepsilon_k + i\zeta_k$, where $\varepsilon_k$ and $\zeta_k$ can be treated as real and linearly independent equations; for each $k = 0, \cdots, N-1$. As evaluated in [133], the solution to Eq.(19) is guaranteed to be of the form:

$$\rho_k(t, \delta t) = \sqrt{A\varepsilon_k^2(t, \delta t)t + 2B\varepsilon(t, \delta t)\zeta_k(t, \delta t) + C\zeta_k^2(t, \delta t)}\,. \tag{22}$$

Also, $\Omega_k = \rho_k^2\dot{\gamma_k}$, is an invariant quantity with respect to time. Hence, $\gamma_k$ can be computed as follows,

$$\gamma_k(t, \delta t) = \int_0^t \frac{\Omega_k}{\rho_i^2(t, \delta t)}dt \quad . \tag{23}$$

Note that all the quantities are now a function of $t$ and $\delta t$, this time-dependence is however suppressed along the most part of this article; until necessary. The creation and annihilation operators in Eq.(16) can now be used to construct invariant operator, for each of the N decoupled Hamiltonians, of Eq.(10):

$$I_k = \Omega_k\left(a_k^\dagger a_k + \frac{1}{2}\right). \tag{24}$$

Here, $k = 0, 1 \cdots, N-1$. Assuming that each invariant operator $I_k$ is one of the complete set of commuting observables for the respective Hamiltonian $H_k$ assures that each $I_k$ has its own eigenstates. The ground state for this spectrum of each invariant operator can be computed using annihilation operator of Eq.(16) by solving, $a_k u_{0_k} = 0$. The expression for the ground state of for the spectrum of each invariant operator $I_k$ of Eq.(24), for each $k = 0, 1 \cdots, N-1$ is given by,

$$u_{0_k} = \left(\frac{\dot{\gamma_k}}{\pi}\right)^{1/4}\exp\left[-\frac{\dot{\gamma_k}}{2}(1 - i\dot{\rho_k}\rho_k\dot{\gamma_k})\tilde{x}_k^2\right]. \tag{25}$$

The creation operator in Eq.(16) can then be used to evaluate the $n^{th}$ eigenstate of the invariant-operator $I_k$ which is given by,

$$u_{n_k} = \frac{1}{\sqrt{n!}}(a_k^\dagger)^n u_{0_k} = \left(\frac{1}{2^{n_k}n_k!}\right)\left(\frac{\dot{\gamma_k}}{\pi}\right)^{1/4}\exp\left[\dot{\gamma_k}\left(1 - \frac{i\dot{\rho_k}}{\rho_k\dot{\gamma_k}}\right)\tilde{x}k^2\right]\mathbf{H}_{n_k}\left[\sqrt{\dot{\gamma_k}}\tilde{x}_k\right]. \tag{26}$$

Here $\mathbf{H}_{n_k}$ denotes the Hermite polynomial of order $n_k$; for each $k = 0, \cdots, N-1$. The eigenstates of the invariant operator shown in Eq.(26) are can now be used to compute the wavefunction for each decoupled Hamiltonian of Eq.(10) such that, $\psi_{n_k} = e^{i\beta_{n_k}}u_{n_k}$, where $\beta_{n_k} = -(1/2 + n_k)$; for $k = 0, \cdots, N-1$. The eigenstates for total unperturbed Hamiltonian of $N$ coupled oscillators is then product for $N$ copies of this wavefunction,

$$\psi^{(0)}_{n_1\cdots n_{N-1}} = \prod_{k=0}^{N-1}\psi_{n_k}. \tag{27}$$

Using Eq.(26)) one can then show that,

$$\psi^{(0)}_{n_1\cdots n_{N-1}} = \left(\frac{1}{2^{n_0+n_1+\cdots+n_{N-1}}n_0!\cdots n_{N-1}!}\right)\left(\frac{g_0g_1\cdots g_{N-1}}{\pi^N}\right)^{1/4}\exp\left[-\frac{i}{2}\sum_{k=0}^{N-1}(2n_k+1)\gamma_k\right]\exp\left[-\frac{1}{2}\sum_{k=0}^{N-1}\tilde{\nu}_k\tilde{x}_k^2\right]$$
$$\times\mathbf{H}_{n_0}\left[\sqrt{\dot{\gamma}_0}\tilde{x}_0\right]\cdots\mathbf{H}_{n_{N-1}}\left[\sqrt{\dot{\gamma}_{N-1}}\tilde{x}_{N-1}\right]. \tag{28}$$

Here $g_k = \dot{\gamma}_k$ and,

$$\tilde{\nu}_k = \dot{\gamma}_k \left(1 - \frac{i\dot{\rho}_k}{\rho_k \dot{\gamma}_k}\right), \tag{29}$$

for $k = 0, 1 \cdots, N-1$. We now focus on the ground state of the wavefunction shown in Eq.(28), which can be written as:

$$\psi^{(0)}_{0\cdots 0} = \left(\frac{g_0 g_1 \cdots g_{N-1}}{\pi^N}\right)^{1/4} \exp\left[-\frac{1}{2}\sum_{k=0}^{N-1}(i\gamma_k + \tilde{\nu}_k \tilde{x}_k^2)\right]. \tag{30}$$

Using the steps shown in [120] the energy eigenvalues for each of the $N$ decoupled oscillators can be evaluated as,

$$\langle \psi_{n_k}| H_i |\psi_{n_k}\rangle = W_k(T)\left[n_k + \frac{1}{2}\right]. \tag{31}$$

Here $W_k(T)$ for $k = 0, \cdots, N-1$ is a time-dependent factor for each oscillator given by,

$$W_k(T) = \frac{\dot{\gamma}_k}{2}\left(\frac{\dot{\rho}_k + \rho_i^2\rho_i^2 + \rho_i^2\dot{\gamma}_i}{\rho_i^2 {\dot{\gamma}_i}^2}\right). \tag{32}$$

Using the above form of time-dependent eigenvalues, one can compute the energy eigenvalues for the decoupled Hamiltonian of Eq.(10), which is given by:

$$\langle \psi^{(0)}_{n_1,\cdots,n_{N-1}}| H |\psi^{(0)}_{n_1,\cdots,n_{N-1}}\rangle = \sum_{k=0}^{N-1} W_k\left(n_k + \frac{1}{2}\right). \tag{33}$$

The above expression for the eigenvalues of unperturbed Hamiltonian can now be used to approximate the first order time-independent perturbative correction to the ground state of Eq.(30).

### B. Wavefunction for $\lambda\phi^4$ perturbation applied to the ground state of $N$ quenched-coupled oscillators

In this section our prime objective is to evaluate analytical expression for the wavefunction of ground state of $N$ coupled oscillators in a perturbative framework, approximated to first order. We consider $\psi^{(1)}$ to be the first order correction arising due to Hamiltonian in Eq.(12). Hence, using Eq.(30), the expression for ground state of total Hamiltonian Eq.(8) corrected to first order in $\lambda$ can be written as,

$$\psi_{0,0,\cdots 0}(\bar{x}_0, \cdots \tilde{x}_{N-1}) = \left(\frac{g_0 g_1 \cdots g_{N-1}}{\pi^N}\right)^{1/4} \times \exp\left[-\frac{1}{2}\sum_{k=0}^{N-1}(i\gamma_k + \tilde{\nu}_k\tilde{x}_k^2 + \lambda\psi^1)\right]. \tag{34}$$

We take note of the fact that for $N$ coupled oscillators the $\lambda\phi^4$ perturbation can give rise to a combination of five terms viz., $x_a^4, x_b^2x_c^2, x_d x_e^3, x_f x_g^2 x_h$ and $x_i x_j x_k x_l$. Hence, we have expressed the form of first order correction by closely following the notations used in [87],

$$\psi^1_4 = \sum_{\substack{a=0 \\ 4a \bmod N \equiv 0}}^{N-1} B_1(a) + \sum_{\substack{b,c=0 \\ (2b+2c) \bmod N \equiv 0 \\ b\neq c}}^{N-1} \frac{B_2(b,c)}{2} + \sum_{\substack{d,e=0 \\ (3e+d) \bmod N \equiv 0 \\ d\neq e}}^{N-1} B_3(d,e) + \sum_{\substack{f,m,h=0 \\ (f+2m+h) \bmod N \equiv 0 \\ f\neq m\neq h}}^{N-1} \frac{B_4(f,m,h)}{2} + \sum_{\substack{i,j,k,l=0 \\ (i+j+k+l) \bmod N \equiv 0 \\ i\neq j\neq k\neq l}}^{N-1} \frac{B_5(i,j,k,l)}{24}. \tag{35}$$

One can compute the exact form of coefficients for each of the five different perturbative terms by first choosing appropriate number of oscillators and then generalising the result for $N$ oscillators. Further one can compute the perturbative correction by setting $V$ as each perturbative term mentioned above, and then using the formula given below,

$$\psi^{(1)}_{0,\cdots 0} = \sum_{(n_0\cdots n_{N-1})\neq(0,\cdots 0)} \frac{\langle \psi^{(0)}_{n_0,\cdots n_{N-1}}| V |\psi^{(0)}_{0,\cdots 0}\rangle \times \psi^{(0)}_{n_0,\cdots n_{N-1}}}{\langle \psi^{(0)}_{0,\cdots 0}| H |\psi^{(0)}_{0,\cdots 0}\rangle - \langle \psi^{(0)}_{n_0,\cdots n_{N-1}}| H |\psi^{(0)}_{n_0,\cdots n_{N-1}}\rangle}. \tag{36}$$

For example, if one wants to get the form of $B_3(d,e)$ in Eq.(35), set number of oscillators to $N = 2$ and put

$V = x_0 x_1^3$ in Eq.(36). The form of the perturbative expansion thus obtained can be generalised for arbitrary number of, $N$ oscillators. We then repeat these steps to fix all the coefficients of perturbative expansion. The exact form of all these coefficients, using Eq.(36) is tabulated in Appendix C.

## IV. Analytical calculation for Circuit Complexity of $\phi^4$ quench model

The ground state of the total Hamiltonian, calculated in the previous section, given by Eq.(34) is used to construct the reference and target states in the subsection IV A. Choosing a specific cost functional, we have derived the analytical expression for circuit complexity, by modifying the results of [87], in the subsection IV B.

### A. Constructing Target/Reference states

In the wavefunction for $N$ oscillators with quartic perturbation shown in Eq.(34) following the prescription given in [87] one can write the exponent in the form of a matrix conjugated by a basis vector, $\vec{v}$. The wavefunction then takes the below given form:

$$\psi^s_{0,0,\cdots 0}\left(\tilde{x}_0, \cdots, \tilde{x}_{N-1}\right) \approx \mathcal{N}^s \exp\left[-\frac{1}{2} v_a A^s_{ab} v_b\right]. \quad (37)$$

Here $\mathcal{N}^s$ denotes the normalisation factor and $A^s$ denotes a block diagonal matrix, for the respective state. Further, the space of circuits is parameterised by setting value of the running parameter $s$. At $s = 1$ the above form of wavefunction coincides with the wavefunction in Eq.(34) such that $\mathcal{N}_{s=1}$ becomes the normalising factor of Eq.(34), by an appropriate choice of the basis $\vec{v}$. Then $\psi^{s=1}_{0,0,\cdots 0}$ is referred to as the target state. There are many possible choices for choosing bases so as to obtain the terms in perturbative expansion in Eq.(35). However as a minimal choice we choose the below mentioned basis,

$$\vec{v} = \{\tilde{x}_0, \cdots \tilde{x}_{N-1}, \tilde{x}_0^2, \cdots, \tilde{x}^2_{N-1}, \cdots, \tilde{x}_a \tilde{x}_b, \cdots\}. \quad (38)$$

In this basis one can show that the matrix $A$, in Eq.(37) has a block diagonal form:

$$A^{s=1}_{ab} = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}. \quad (39)$$

$A_1$ contains coefficients of terms like $x_a^2$ and $x_a x_b$ in Eq.(34) multiplied by $-2$. All the elements of $A_1$ can be fixed to obtain a specific form of target state. This block is often referred to as the *unambiguous* block.

The elements of $A_2$ block consist of coefficients which are basis dependent i.e. there is not any unique choice of basis vector for defining the elements of $A_2$ block because unlike $A_1$ block which only consist of coefficients of quadratic terms and they can be defined uniquely without any ambiguity, the $A_2$ block consists of elements which are coefficients of terms like $\tilde{x}_a^2 \tilde{x}_b^2$, $\tilde{x}_a^2 \tilde{x}_b^2 \tilde{x}_c^2$, $\tilde{x}_a \tilde{x}_b \tilde{x}_c \tilde{x}_d$ which can be defined in several ways. Due to this arbitrariness, the complexity for the ambiguous block will be different for different choices of basis. One cannot therefore fix elements of $A_2$ such that the contribution of $A_2$ to total complexity of the system is independent of choice of basis. Due to these ambiguities, the $A_2$ block is often referred to as *ambiguous* block. We construct the reference state by choosing the value of all the frequencies for each of the $N$ oscillators as $\tilde{\omega}_{ref}$. Since all the oscillators have same frequency, the reference state evidently would be time-independent. Then wavefunction in Eq.(35) can be modified to that of reference state, mentioned below,

$$\psi^{s=0}(x_1, x_2, ...., x_n) = \mathcal{N}^{s=0} \exp\Big[-\sum_{i=0}^{N-1} \frac{\tilde{\omega}_{ref}}{2}\left(x_i^2 + \lambda^0 x_i^4\right)\Big]. \quad (40)$$

Note that, $\lambda^0$ is a parameter denoting the non-Gaussian nature of the reference state and is not to be confused with the perturbative coupling $\lambda$ used in target state. In normal modes the above expression can be recast as shown below,

$$\psi^{s=0}(\tilde{x}_1, \tilde{x}_2, ...., \tilde{x}_n) = \mathcal{N}^{s=0} \exp\Big[-\frac{1}{2}\Big(v_a A^{s=0}_{ab} v_b\Big)\Big], \quad (41)$$

where the matrix $A^{s=0}_{ab}$ can be fixed as:

$$A^{s=0}_{ab} = \begin{pmatrix} \tilde{\omega}_{ref} \mathbb{I}_{N\times N} & 0 \\ 0 & A_2^{s=0} \end{pmatrix}. \quad (42)$$

Again there will be ambiguities in fixing the elements of $A_2^{s=0}$ for the reasons already mentioned above.
Equipped with the target and and reference states for the quenched and interacting oscillators we can proceed to give an analytical expression for circuit complexity by getting around the ambiguities in the upcoming subsection.

### B. Analytical calculation of the complexity functional

In this subsection we outline analytical steps to compute the expression for circuit complexity for the previously mentioned target state Eq.(34) starting with a reference state Eq.(41) using the results of [87].
As shown in [100] the complexity functional depends on the chosen cost function. In this article we work with the following cost function,

$$\mathcal{F}_\kappa(s) = \sum_I p_I |Y^I|^\kappa. \quad (43)$$

As shown in [80, 87] the circuit complexity with this particular cost function becomes,

$$\mathcal{C}_\kappa = \int_{s=0}^{1} \mathcal{F}_\kappa \, ds. \tag{44}$$

Next, we write the complexity as sum of two terms,

$$\mathcal{C}_\kappa = \mathcal{C}_\kappa^{(1)} + \mathcal{C}_\kappa^{(2)}. \tag{45}$$

Here $\mathcal{C}_\kappa^{(1)}$ refers to the contribution to the circuit complexity from $A_1$ block while $\mathcal{C}_\kappa^{(2)}$ refers to the contribution from $A_2$ block. $\mathcal{C}_\kappa^{(1)}$ and $\mathcal{C}_\kappa^{(2)}$ can be given as ratio of eigenvalues of the respective blocks for chosen target (34) and reference states Eq.(41) as prescribed in [87, 100],

$$\mathcal{C}_\kappa = \frac{1}{2^\kappa}\sum_{i=0}^{N-1}\left|log\Big(\frac{\Lambda_i^{(1)}}{\tilde{\omega}_{ref}}\Big)\right|^\kappa + \mathcal{A}\sum_j\left|log\Big(\frac{\Lambda_j^{(2)}}{h_i\tilde{\omega}_{ref}\lambda_0}\Big)\right|^\kappa. \tag{46}$$

Here $\mathcal{A}$ denotes the penalty factor. Due to the ambiguities arising while fixing the form of $A_2$ block, numerically one cannot fix the form of $\Lambda_i^{(2)}$. However, as discussed in the Appendix B we can choose a minimal basis such that elements of $A_2$ are fixed for $N = 2$ oscillators. In the Appendix B, we have computed the total circuit complexity for $N = 2$ coupled oscillators having a quenched Hamiltonian with a quartic coupling.

In this work, have neglected the contribution of the ambiguous block i.e. $\mathcal{C}_\kappa^{(2)}$ as we cannot find any basis to get the numerically exact contribution of $A_2$ block for arbitrary number of oscillators. Nonetheless, we attempt to give an analytical form of $\mathcal{C}_{\kappa=1}^{(2)}$ in terms of renormalized parameters in the Appendix A again following the steps shown in [87].

### C. The Continuum Limit for $\mathcal{C}_1$

Now, we will compute the exact form of eigenvalues of $A_1$ block, taking the continuum limit. To find the form of $\Lambda_i^{(1)}$, we now reinstate the factor of $M$ previously set to $M = 1$ in the section II. In light of this the Hamiltonian, given in Eq.(8) will change, although retaining the previous form of eigenvalues and eigenfunctions. The new Hamiltonian with a factor of $M$ becomes,

$$H = \frac{1}{M}\sum_{\vec{n}}\Big\{\frac{P(\vec{n})^2}{2} + \frac{1}{2}M^2\Big[\omega^2 X(\vec{n})^2 + \Omega^2\sum_i (X(\vec{n}) - X(\vec{n}-\hat{x}_i))^2 + 2\big\{\lambda_4 X(\vec{n})^4\big\}\Big]\Big\}. \tag{47}$$

Considering the reinstated factor of $M$, we rescale some of the parameters as shown below,

$$\omega \to \frac{\omega}{\delta}; \quad \eta \to \frac{\eta}{\delta}; \quad \lambda \to \frac{\lambda}{\delta^2}; \quad \tilde{\omega}_{ref} \to \frac{\tilde{\omega}_{ref}}{\delta}; \quad \lambda^0 \to \frac{\lambda^0}{\delta}.$$

Using these rescaled parameters one can generalise the form of eigenvalues $\Lambda_i^{(1)}$ using Mathematica by considering trial cases for different values of $N$ and eliminating all the factors aside for $\rho_k$ and $\tilde{\omega}_k$ by using appropriate formulae mentioned in previous sections. Below we show the generalised formula for eigenvalues of $A_1^{(s=1)}$ block depending on whether the chosen number of oscillators, $N$ is even or odd,

$$\begin{aligned}\Lambda_i &= \frac{3\lambda\rho_i^2}{2N}\left(\frac{g_\alpha + 2g_i}{g_\alpha\left(\rho_i^2\left(g_i+\omega_i^2\right)+\dot{\rho}_i\right)} - \frac{2g_i\rho_\alpha^2}{\rho_i^2\left(\rho_\alpha^2\left(\omega_i^2 g_\alpha + g_i\left(2g_\alpha+\omega_\alpha^2\right)\right)+g_i\dot{\rho}_\alpha\right)+\dot{\rho}_i g_\alpha\rho_\alpha^2}\right) + \frac{\tilde{\nu}_i}{2}, \quad N\text{: Even} \\ &= \frac{3\lambda g_i^2\rho_i^4\left(\left(g_i+\omega_i^2\right)\rho_i^2+\dot{\rho}_i\right)}{N g_i\left(\rho_i^2\left(g_i+\omega_i^2\right)+\dot{\rho}_i\right)\left(\rho_i^2\left(\rho_i^2\left(g_i\omega_i^2+g_i\left(2g_i+\omega_i^2\right)\right)+g_i\dot{\rho}_i\right)+g_i\dot{\rho}_i\rho_i^2\right)} + \frac{\tilde{\nu}_i}{2}, \quad N\text{: Odd}\end{aligned} \tag{48}$$

where the index, $\alpha = |N/2 - i|$. One can insert this expression for eigenvalues in Eq.(46) to get the desired value of circuit complexity. We emphasize that the form of these eigenvalues make the circuit complexity a time-dependent quantity due to the chosen quench profile. This time-dependence of complexity is explored in numerical plots by choosing an appropriate scale of time.
One can check the behaviour of circuit complexity, $C_1$ at the continuum limit: $N \to \infty$ while $\delta \to 0$ such that $L = (N\delta)$ is finite.
Now, in arbitrary dimensions, the equation for $\mathcal{C}_{\kappa=1}^{(1)}$, can be rewritten as:

$$\mathcal{C}_{\kappa=1}^{(1)} = \frac{1}{2}\sum_{k=1}^{d-1}\sum_{i=0}^{N-1}\left|log\Big(\frac{\Lambda_i^{(1)}}{\tilde{\omega}_{ref}}\Big)\right|. \tag{49}$$

For simplicity, if one chooses to set same frequency (of the respective oscillator) in each dimension, $d$, then one can write:

$$\omega_i = \sqrt{\frac{1}{d-1}\left[4(d-1)\eta^2\sin^2\left(\frac{\pi i}{N}\right) + \omega_0^2\tanh^2\left(\frac{t}{\delta t}\right)\right]}. \tag{50}$$

Hence, in arbitrary dimensions $d$, the circuit complexity can be still be computed using Eq.(46), thus getting rid

of the lattice sums in Eq.(49) such that frequencies are set to Eq.(50).

## V. Numerical Results

In this section we numerically evaluate the value of circuit complexity for coupled oscillators using the expressions computed for total complexity of two coupled oscillators, denoted by $\mathcal{C}$, shown in the Appendix B and the unambiguous contribution of complexity for arbitrary number of oscillators using the results of the subsection IV C, henceforth denoted by $\mathcal{C}_1$. Note that each time-dependent coefficients in these expressions explicitly depend on $\rho_k(t, \delta t)$ given by Eq.(22). To obtain the values of $A, B$ and $C$ which can be inserted in Eq.(22), we set some straightforward initial conditions. The invariant quantities, $\Omega_k$ in Eq.(23) are taken to be $\Omega_k = 1$. Using the values of $\rho_k(t, \delta t)$ for $t \to 0$, we get the desired values of $A, B$ and $C$ by setting $\rho_k(0, \delta t) = 1$ while $\gamma_k(0, \delta t) = 0$ by using $AC - B^2 = \Omega^2$. The coupling between the oscillators is set to be $\eta = 0.25$. The free parameter in the quench profile, Eq.(13) is set to $\omega_0 = 1$. The frequency of reference state is set to $\omega_{ref} = 0.001$.

Using the exact values of circuit complexity we parameterise four different plots in this section. The first plot features the behaviour of total complexity $\mathcal{C}$ computed for two coupled oscillators discretised on a lattice of size $L = 20$, varying with the quench rate $\delta t$. This plot is divided into two regions, $\delta t < 1$ is the sudden quench limit (blue) region while $\delta t > 1$ is the slow quench limit (yellow) region. The next three plots characterise the behaviour of unambiguous contribution of circuit complexity $C_1$ for more than two coupled oscillators discretised on a lattice of size, $L = 100$, varying with dimensionless parameter $(t/\delta t)$. In these plots we mark $(t/\delta t) = 1$ as Quench Point, by a dotted vertical red line. The red region for $(t/\delta t) < 1.2$ features the early time behavior of circuit complexity, the yellow region $0.8 < (t/\delta t) < 1.2$ shows the behavior of complexity near the quench point while the blue region, $(t/\delta t) > 1.2$ characterises the late time behavior of circuit complexity.

In Fig.1, we have plotted the numerical values of total complexity for $N = 2$, coupled oscillator system with quartic interaction , using Eq.(B7), with respect to the quench rate $\delta t$. The blue coloured region shows the behaviour of circuit complexity in the fast regime, $\delta << 1$. In slow regime, initially, the circuit complexity monotonically increases till $\delta t = 0.01$. Beyond, $\delta t = 0.01$ the complexity shows a linear scaling with slope, $\log \mathcal{C}/\log \delta t = 0.0825$, upto $t = 0.1$. This linear scaling is then followed by a monnotonous increase in circuit complexity till $t = 0.2$. Beyond $t = 0.2$, circuit complexity saturates, making transition into slow regime, $\delta t > 1$, marked by a yellow background. Furthermore, it is evident that, for each quartic coupling, $\lambda$, the circuit complexity shows same behaviour. However, it is clear that as one increases quartic coupling, the complexity decreases.

Although one cannot get numerical results for total circuit complexity of more than two oscillators, we have plotted the numerical values of complexity pertaining to the unambiguous block i.e., $\mathcal{C}_1$ with respect to the dimensionless parameter $(t/\delta t)$ for $N = 10$ coupled oscillators, in Fig.2. The figure is divided into three regions, first is the early time behaviour for the dimensionless parameter between $0 < (t/\delta t) < 0.8$ marked by a red background. In this region for all the three considered quartic couplings, circuit complexity decreases linearly upto $(t/\delta t) = 0.7$. Beyond this, for $(t/\delta t) > 0.7$, circuit complexity $C_1$, diverges for each particular quartic coupling $\lambda$, such that higher the value of $\lambda$, higher is the complexity $C_1$. This linear scaling is then followed by a monotonous decrease in the value of $C_1$, as we move near the quench point $(t/\delta t) = 1$, marked by a yellow background. The late time behaviour, is characterised by saturation of $C_1$, for $(t/\delta t) > 1.2$, which is marked by a blue background. It is evident that for different quartic couplings, $\lambda$, the circuit complexity $C_1$ scales similarly in each particular region. However as we increase the quartic coupling, $\lambda$, circuit complexity $C_1$ increases at any particular time near and beyond the quench point.

In Fig.3 we have plotted the unambiguous contribution of circuit complexity $C_1$ for a discretised field theory with respect to the dimensionless parameter $(t/\delta t)$ for different number of oscillators, viz., $N = 10, 11, 12$ (keeping quartic coupling fixed at $\lambda = 0.01$). The early time behavior for a particular $N$ is characterised by a steep linear decrease in value of circuit complexity, $C_1$ upto $(t/\delta t) = 0.4$. Beyond this point the complexity $C_1$ monotonically decreases and finally saturates at $(t/\delta t) \approx 0.5$. After this point, the unambiguous contribution of circuit complexity, $C_1$ remains saturated. The early time behavior of $C_1$ featuring a decrease and saturation in value of complexity is marked by red background. The behavior of $C_1$ near to the quench point as well as at late time is characterised by a constant saturated value at all times, marked by yellow and blue backgrounds respectively. It is evident that unambiguous contribution of circuit complexity, $C_1$ scales similarly for any $N$. However larger the number of oscillators $N$, larger is the value of $C_1$. In the continuum limit when we have large number of coupled oscillators, one can expect that the unambiguous contribution of circuit complexity, $C_1$ will behave similar to that shown in this figure.

In Fig.4, we have plotted the unambiguous contribution of circuit complexity $C_1$ for a discretised field theory with respect to the dimensionless parameter $(t/\delta t)$ for different number of dimensions, viz., $d = 2, 3, 4$ (keeping quartic coupling fixed at $\lambda = 0.01$ while number of oscillators $N = 10$.). For a particular $d$, the early time behavior of unambiguous contribution of circuit complexity is characterised by a monotonous decrease in the value of $C_1$, marked by a red background. The values of $C_1$ for different dimensions, $d$ begin to converge near to the

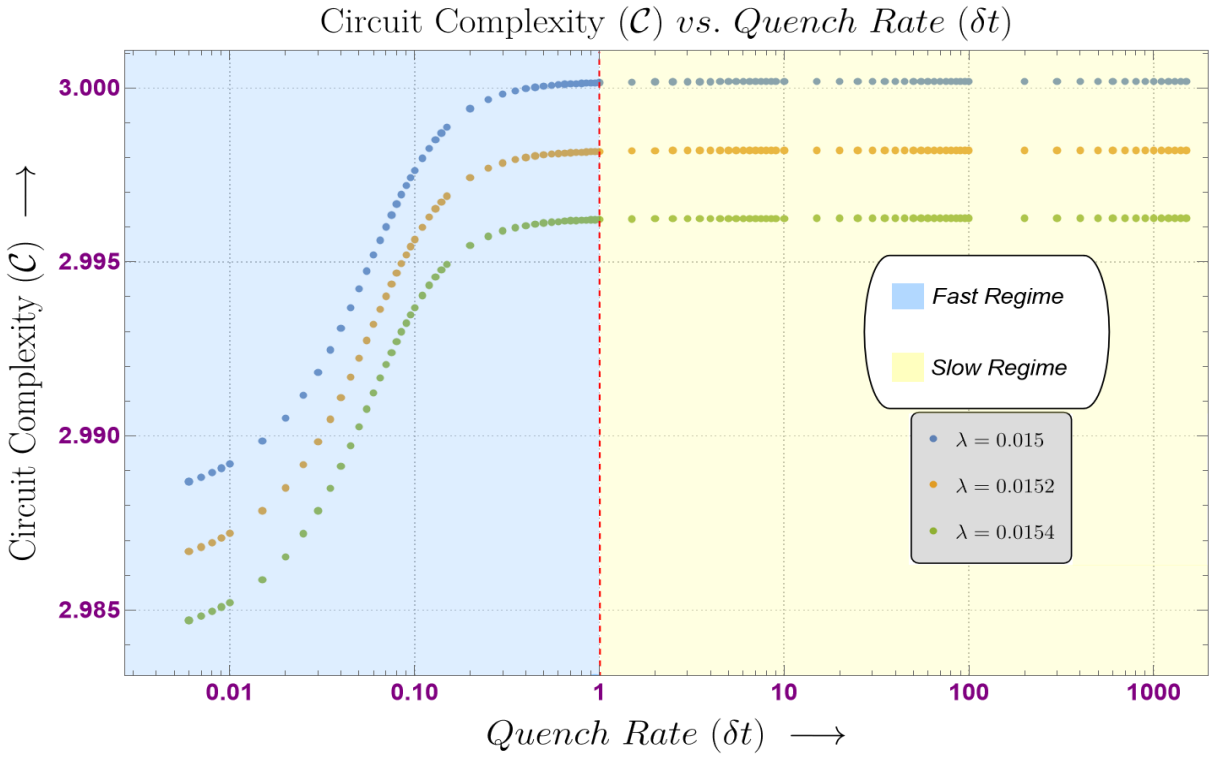



FIG. 1. Log-Log variation of the total circuit complexity ($C$) with respect to the quench rate ($\delta t$) for different orders of the coupling constant $\lambda$ for two coupled oscillators with quartic perturbation.

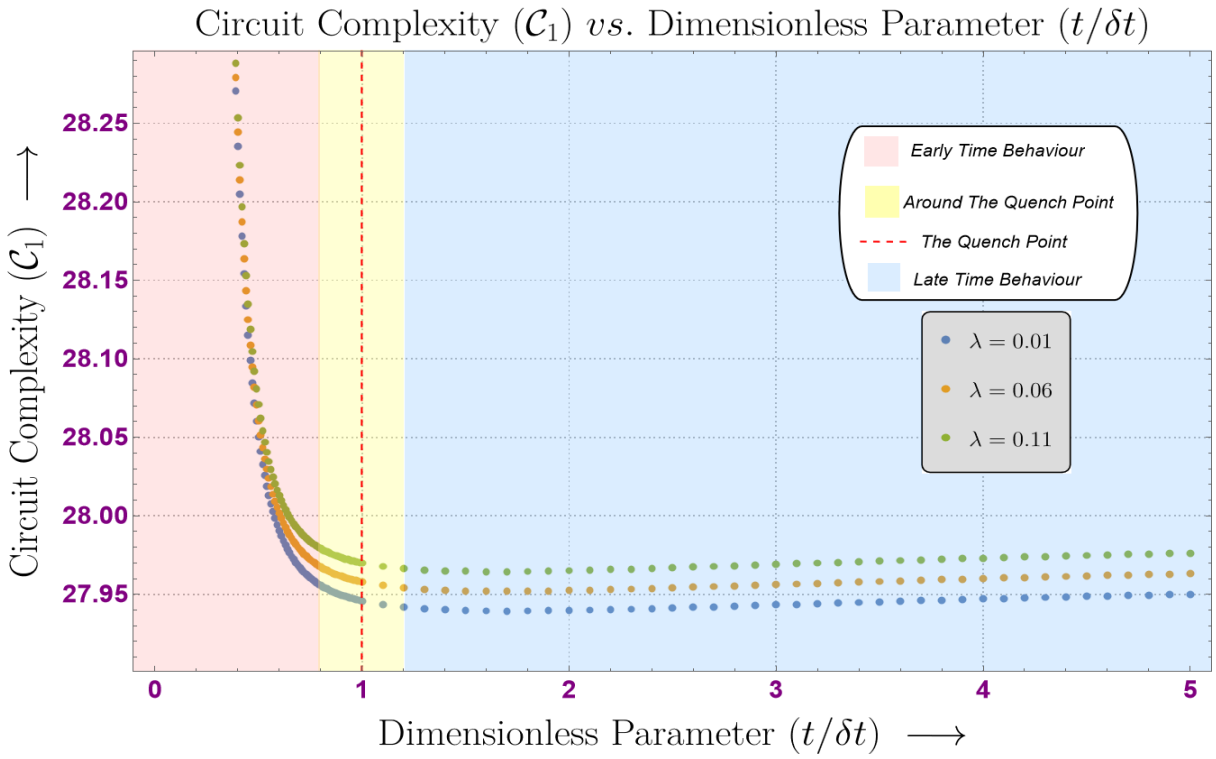



FIG. 2. Variation of the circuit complexity for $A_1$ block ($C_1$) for $N = 10$, with respect to the dimensionless parameter ($t/\delta t$) for different orders of the coupling constant $\lambda$.

quench point ($t/\delta t = 1$) and finally saturate to a constant value, this is marked by yellow background. At

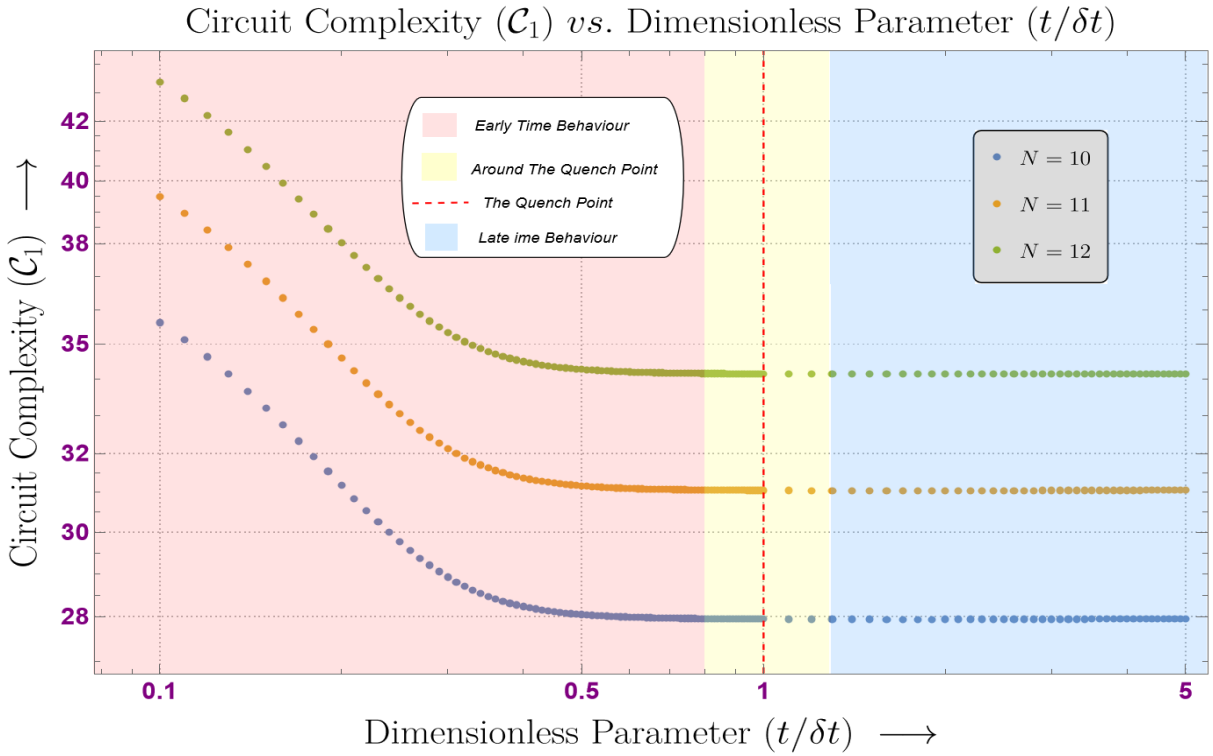



FIG. 3. Semi-Log variation of the circuit complexity for $A_1$ block ($C_1$) at $\lambda = 0.01$, with respect to the dimensionless parameter ($t/\delta t$) for different values of $N$.

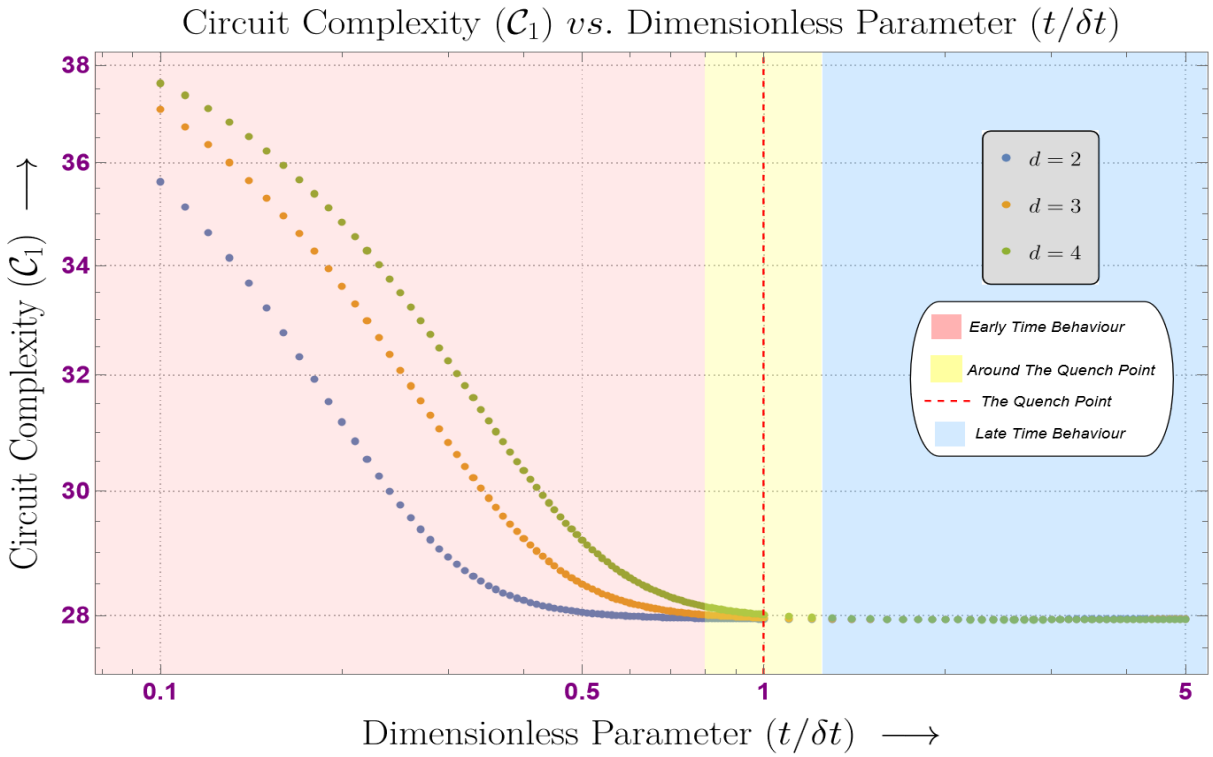



FIG. 4. Semi-Log variation of the circuit complexity for $A_1$ block ($C_1$) at $\lambda = 0.01, N = 10$ with respect to the dimensionless parameter ($t/\delta t$) for different dimensions $d$.

late times, the unambiguous contribution of circuit complexity $C_1$ remains saturated to a constant value, this

is marked by a blue background. It is evident for any dimension $d$, the circuit complexity $C_1$ scales similarly. However higher the dimension $d$, larger is the value of $C_1$ at any particular time in the red region.

## VI. Conclusions

The concluding remarks of this work are appended below point-wise:

- Focusing on a quenched quantum field theory, in this article we have derived the analytical expression for time-dependent quantum computational complexity using Nielsen's method viz., circuit complexity. Further, we have also discussed on various parametric variation of circuit complexity.

- Discretising the field theory on a lattice, we evaluated the time-dependent wavefunction for $N$-coupled oscillators having quartic perturbation and a quenched frequency. The wavefunction for the ground state of this system was derived using invariant operator method in a perturbative framework.

- The analytical form of the reference state and target states was chosen using the ground state wavefunction. Using a specific cost function and a minimal choice of basis, we could compute the exact form for the total complexity for two coupled oscillators. Further, we have computed the exact form of the unambiguous contribution of circuit complexity, for arbitrary, $N$ number of oscillators.

- For two coupled oscillators, we observed that, in most part of the sudden quench, the total circuit complexity monotonously increases at very small values of quench rate, then scales linearly and shows a trend of thermalisation near $\delta t = 1$. In the slow quench limit, the complexity remains saturated irrespective of the quench rate. When parameterised for different values of quartic coupling, $\lambda$, it is evident that as coupling increases, even circuit complexity increases.

- The exact analytical form for unambiguous contribution of the circuit complexity for $N$ coupled oscillators was derived using the results of [87]. The parametric variation of this circuit complexity was then plotted with respect to dimensionless parameter $(t/\delta t)$.

- It is evident from these plots that unambiguous contribution of the circuit complexity decreases with respect to the dimensionless parameter. When parametrised for different quartic couplings, we find that initially the complexity decreases linearly following same line, irrespective of the quartic coupling. Near to the quench point, complexity for each quartic coupling diverges and saturates at late times. After, quench point, the complexity after divergence is clearly in direct proportion to the increasing value of quartic coupling.

- We observed that, the unambiguous contribution of the circuit complexity behaves similarly, irrespective of the number of oscillators, $N$. However, as $N$ increases the respective value of complexity increases at any particular time. Using this, we commented on the conitnuum limit where the results would still be the same.

- Furthermore, it is clear that the unambiguous contribution of circuit complexity for the chosen set of parameters is proportional to the increasing number of dimension at early times. However at the quench point, unambiguous contribution of circuit complexity attains a constant value irrespective of the number of dimensions and thermalises at late times.

**Future Prospects:**

- In this work we explored the effects of quantum quench on QFT with quartic coupling, in the future it would be interesting to explore the effects of quantum quench on Krylov Complexity [134–138] for QFT with quartic coupling.

- Some of the works focusing on understanding the connection between complexity and quantum entanglement are [131, 139–142]. The connection between complexity and quantum entanglement in the case of quenched theories might turn out to be fruitful.

**Acknowledgement:** SC would like to thank the work friendly environment of CCSP, SGT University for providing tremendous support in research and offer the Assistant Professor (Senior Grade) position. SC also thanks all the members of our newly formed virtual international non-profit consortium Quantum Aspects of the Space-Time & Matter (QASTM) for elaborative discussions. RMG, SM and NP would like to thank the members of the QASTM Forum for useful discussions. Last but not least, we would like to acknowledge our debt to the people belonging to the various parts of the world for their generous and steady support for research in natural sciences.

## A. $\mathcal{C}^{(2)}_{\kappa=1}$ in terms of renormalized parameters

As discussed in [87] one can attempt to find the form of $\mathcal{C}^{(2)}_{\kappa=1}$ by using renormalisation. In this subsection we give an outline of the modified expression for the same in case of a quenched interacting field theory discretised on a $d-$ dimensional lattice containing $N$ oscillators.
The renormalized matrix elements for $A_2$ block are given as [87],

$$A_2[m,n] = \frac{a_{mn}\lambda_R\delta^{-d}}{V^{\frac{1}{d-1}} f(\tilde{\omega}_i)}. \tag{A1}$$

Note that all these elements are time-dependent due to the form of frequency, $\tilde{\omega}_i$ chosen as a quench profile. The eigenvalues then take the general form [87],

$$\Lambda^{(2)}_i = \frac{b_j\lambda_R\delta^{-d}}{V^{\frac{1}{d-1}} g(\tilde{\omega}_i)}, \tag{A2}$$

where $j \in \{0,1,\cdots(DimA_2-1\}$ while $i \in \{0,1,\cdots N-1\}$. As shown in [87] the renormalized penalty factor is,

$$\mathcal{A} = (\lambda_R\delta^{4-d})^{\mu}\delta^{-\upsilon}V^{\frac{\upsilon}{d-1}}. \tag{A3}$$

Here $\upsilon$ and $\mu$ are arbitrary integers which can be fixed by using physical arguments as discussed in [87]. Using Eq.(46), the renormalized-complexity contribution for $A_2$ block in $d-$ dimensions can then be written as,

$$\mathcal{C}^{(2)}_{\kappa=1} = \frac{(\lambda_R\delta^{4-d})^{\mu}\delta^{-\upsilon}V^{\frac{\upsilon}{d-1}}}{2}\sum_{k=0}^{d-1}\sum_{i_k=0}^{(\mathrm{Dim}A_2)-1}\Big|\log\Big(\frac{\Lambda^{(2)}_{i_k}\delta^2}{h_i\tilde{\omega}_{ref}\lambda_0}\Big)\Big|. \tag{A4}$$

Note that now each $i_k \in \{1,2,\cdots(\mathrm{Dim}A_2-1)\}$ for $k \in \{1,2,\cdots d-1\}$. Using the renormalized form of eigenvalues from Eq.(A2) we finally obtain,

$$\mathcal{C}^{(2)}_{\kappa=1} = \frac{(\lambda_R\delta^{4-d})^{\mu}\delta^{-\upsilon}V^{\frac{\upsilon}{d-1}}}{2}\sum_{k=0}^{d-1}\sum_{i_k=0}^{(\mathrm{Dim}A_2)-1}\Big|\log\Big(\frac{b_{i_k}\lambda_R\delta^{2-d}}{V^{\frac{1}{d-1}} g(\tilde{\omega}_i)h_{i_k}\tilde{\omega}_{ref}\lambda_0}\Big)\Big| \tag{A5}$$

. Note that, we have not used this expression for getting numerically exact results.

## B. Circuit Complexity for two oscillators

In the subsection IV A, we commented on the ambiguities in fixing the coefficients in the ambiguous $A_2$ block for arbitrary number of oscillators, $N$. Due to these ambiguities one cannot obtain numerical results for contribution from $A_2$ block in circuit complexity for $N$ oscillators. However, in this appendix, we choose a minimal basis for the case of two oscillators and hence get rid of these ambiguities to get the total contribution of $A_1$ as well as $A_2$ block in circuit complexity. We begin by specialising the wavefunction (in normal modes) in Eq.(34) to that of two oscillators by inserting $N=2$, written as:

$$\psi_{0,0}(\bar{x}_0,\bar{x}_1) = \frac{(g_0g_1)^{1/4}}{\sqrt{\pi}}\exp\left[-\iota(\gamma_0+\gamma_1)\right]\exp\left[C_0\right]\exp\Big[-\frac{1}{2}\Big(C_1\bar{x}_0^2+C_2\bar{x}_1^2+C_3\bar{x}_0^2\bar{x}_1^2+C_4\bar{x}_0^4+C_5\bar{x}_1^4\Big)\Big]. \tag{B1}$$

The exact form of $C_i$ for $i=0$ to $i=5$ can be inferred from the table in the Appendix C. Next, we write the above wavefunction in the following form:

$$\psi^s(\bar{x}_0,\bar{x}_1) = \mathcal{N}^s\exp\Big[-\frac{1}{2}\Big(v_aA(s)_{ab}\ v_b\Big)\Big]. \tag{B2}$$

Here $\mathcal{N}^s$ is the normalisation factor. Similar to that of generalised wavefunction inserting $s=1$ in the above equation will correspond to the target state, while inserting $s=0$ corresponds to the reference state. We choose an unentangled and non-Gaussian reference state given by,

$$\psi^{s=0}(\bar{x}_0,\bar{x}_1) = \mathcal{N}^{s=0}\exp\Big[-\frac{\tilde{\omega}_{ref}}{2}(\bar{x}_0^2+\bar{x}_1^2+\frac{\lambda_0}{2}(\bar{x}_0^4+\bar{x}_1^4+6\bar{x}_0^2\bar{x}_1^2))\Big].$$

Here $\lambda_0$ parameterizes the non-Gaussianity of the reference state. The exponential in the above equation can be written in form of a matrix $A(s=0)$ by choosing a basis,

$$\vec{v} = \{\bar{x}_0, \bar{x}_1, \bar{x}_0\bar{x}_1, \bar{x}_0^2, \bar{x}_1^2\}. \tag{B3}$$

The matrix then takes the following form:

$$A(s=0) = \begin{pmatrix} A_1^0 & 0 \\ 0 & A_2^0 \end{pmatrix}, \tag{B4}$$

where,

$$A_1^0 = \begin{pmatrix} \tilde{\omega}_{ref} & 0 \\ 0 & \tilde{\omega}_{ref} \end{pmatrix} \quad ; \quad A_2^0 = \lambda_0\tilde{\omega}_{ref}\begin{pmatrix} b & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2}(3-b) \\ 0 & \frac{1}{2}(3-b) & \frac{1}{2} \end{pmatrix}.$$

One can choose the values of $b$ such that matrix $A_2^0$ is non-singular. To diagonalize $A_2^0$ we set $b=3$. On the other hand to choose a non-Guassian reference state we set $\lambda_0 = 1.5$.
Similarly one can get the target state in form of matrix $A(s=1)$ given by,

$$\psi^s(\bar{x}_0, \bar{x}_1) = \mathcal{N}^{s=1}\exp\Big[-\frac{1}{2}\Big(v_a A(s=1)_{ab}\, v_b\Big)\Big]. \tag{B5}$$

Choosing the same basis as that in Eq.(B3) we can write the matrix for target state as,

$$A(s=1) = \begin{pmatrix} A_1^1 & 0 \\ 0 & A_2^1 \end{pmatrix}, \tag{B6}$$

where,

$$A_1^1 = \begin{pmatrix} C_1 & 0 \\ 0 & C_2 \end{pmatrix} \quad ; \quad A_2^1 = \begin{pmatrix} \tilde{b}C_5 & 0 & 0 \\ 0 & C_3 & \frac{1}{2}(1-\tilde{b})C_5 \\ 0 & \frac{1}{2}(1-\tilde{b})C_5 & C_4 \end{pmatrix}.$$

The parameter $\tilde{b}$ can be chosen such that $A_2^1$ is non-singular. Further to diagonalise $A_2^1$ we set $\tilde{b}=1$. It is clear that with a minimal choice of basis given in Eq.(B3) one can fix all the elements of both $A_1$ and $A_2$ blocks for the case of two oscillators. Aimed with the expression for circuit complexity for $N$ oscillators given in Eq.(46), and setting the penatly factor $\mathcal{A}=\infty$, we have computed the circuit complexity for the two quenched oscillators with quartic coupling, by using:

$$\mathcal{C}_{\kappa=1} = \frac{1}{2}\Big(\log\Big|\frac{\det A_1^1}{\det A_1^0}\Big| + \log\Big|\frac{\det A_2^1}{\det A_2^0}\Big|\Big). \tag{B7}$$

Here *det* refers to the determinant of the respective block of the matrix. Note that the quench profile chosen as the frequency scale of the system imposes time-dependence on each element of the target state matrix block viz., $A_1^1$ and $A_2^1$. The complexity therefore becomes time-dependent, this is clearly shown in the numerical results discussed in the section V.

### C. Tabulated Values of Coefficients

In this appendix, the values of various coefficients we have used in some steps to compute the analytical expression of circuit complexity, are tabulated in respective tables.

- We begin by listing the values of $B_i$ for $i = 1, 2...5$ in equation Eq.(35) of section III B in the table given below.

| $B_i$ | Coefficient of $B_i$ |
|---|---|
| $B_1(a)$ | $-\frac{3x_a^2}{4Ng_aW_a}+\frac{9}{16Ng_a^2W_a}-\frac{x_a^4}{4NW_a}$ |
| $B_2(b,c)$ | $-\frac{3x_b^2W_c}{2NW_bg_c(W_b+W_c)}-\frac{3W_bx_c^2}{2Ng_bW_c(W_b+W_c)}+\frac{3}{4Ng_bg_c(W_b+W_c)}+\frac{3W_c}{4Ng_bW_bg_c(W_b+W_c)}+\frac{3W_b}{4Ng_bg_cW_c(W_b+W_c)}-\frac{3x_b^2x_c^2}{N(W_b+W_c)}$ |
| $B_3(d,e)$ | $-\frac{12x_dW_ex_e}{Ng_e(W_d+W_e)(W_d+3W_e)}-\frac{4x_dx_e^3}{N(W_d+3W_e)}$ |
| $B_4(f,m,h)$ | $-\frac{12x_fx_hW_m}{Ng_m(W_f+W_h)(W_f+W_h+2W_m)}-\frac{12x_fx_hx_m^2}{N(W_f+W_h+2W_m)}$ |
| $B_5(i,j,k,l)$ | $-\frac{24x_ix_jx_kx_l}{N(W_i+W_j+W_k+W_l)}$ |

Note that $g_k$ for $k = 0, \cdots N-1$ is defined in the subsection III A, while $W_k$ in Eq.(32). Also all the indices in the tabulated expressions run from 0 to $N-1$.

- Next, we tabulate the values of coefficients $C_i$ for $i = 1, 2...5$ in equation Eq.(B1) of the Apendix B.

| $C_i$ | Coefficient of $C_i$ |
|---|---|
| $C_0$ | $\frac{9\lambda}{32g_0^2W_0}+\frac{9\lambda}{32g_1^2W_1}+\frac{3\lambda}{8g_0g_1(W_0+W_1)}+\frac{3\lambda W_1}{8g_0g_1W_0(W_0+W_1)}+\frac{3\lambda W_0}{8g_0g_1W_1(W_0+W_1)}$ |
| $C_1$ | $-\frac{3\lambda}{8g_0W_0}-\frac{3\lambda W_1}{4g_1W_0(W_0+W_1)}-\frac{\nu_0}{2}$ |
| $C_2$ | $-\frac{3\lambda}{8g_1W_1}-\frac{3\lambda W_0}{4g_0W_1(W_0+W_1)}-\frac{\nu_1}{2}$ |
| $C_3$ | $-\frac{\lambda}{8W_0}$ |
| $C_4$ | $-\frac{\lambda}{8W_1}$ |
| $C_5$ | $-\frac{3\lambda}{2(W_0+W_1)}$ |

Note that $g_k = \dot{\gamma}_k$ where $\gamma_k$ for $k = 0, 1$ can be computed by solving the EMP equation as shown in the section II, while $W_k$ can be computed using Eq.(32).

[1] P. Calabrese and J. Cardy, "Entanglement entropy and quantum field theory," *Journal of Statistical Mechanics: Theory and Experiment* **2004** no. 6, (June,

2004) 06002, arXiv:hep-th/0405152 [hep-th].
[2] E. Witten, "Notes on Some Entanglement Properties of Quantum Field Theory," *arXiv e-prints* (Mar., 2018) arXiv:1803.04993, arXiv:1803.04993 [hep-th].
[3] T. Nishioka, "Entanglement entropy: Holography and renormalization group," *Reviews of Modern Physics* **90** no. 3, (July, 2018) 035007, arXiv:1801.10352 [hep-th].
[4] D. Blanco, "Quantum information measures and their applications in quantum field theory," *arXiv e-prints* (Feb., 2017) arXiv:1702.07384, arXiv:1702.07384 [hep-th].
[5] M. Headrick, "Lectures on entanglement entropy in field theory and holography," *arXiv e-prints* (July, 2019) arXiv:1907.08126, arXiv:1907.08126 [hep-th].
[6] L. Amico, R. Fazio, A. Osterloh, and V. Vedral, "Entanglement in many-body systems," *Reviews of Modern Physics* **80** no. 2, (Apr., 2008) 517–576, arXiv:quant-ph/0703044 [quant-ph].
[7] J. I. Cirac, "Entanglement in many-body quantum systems," *arXiv e-prints* (May, 2012) arXiv:1205.3742, arXiv:1205.3742 [quant-ph].
[8] A. Lakshminarayan and V. Subrahmanyam, "Multipartite entanglement in a one-dimensional time-dependent ising model," *Phys. Rev. A* **71** (Jun, 2005) 062334.
[9] H. F. Song, S. Rachel, C. Flindt, I. Klich, N. Laflorencie, and K. Le Hur, "Bipartite fluctuations as a probe of many-body entanglement," *Phys. Rev. B* **85** (Jan, 2012) 035409.
[10] C. J. Bardeen, "Time dependent correlations of entangled states with nondegenerate branches and possible experimental realization using singlet fission," *The Journal of Chemical Physics* **151** no. 12, (2019) 124503.
[11] A. Sivaramakrishnan, "Entanglement entropy with a time-dependent hamiltonian," *Phys. Rev. D* **97** (Mar, 2018) 066003.
[12] P. Caputa, G. Mandal, and R. Sinha, "Dynamical entanglement entropy with angular momentum and U(1) charge," *Journal of High Energy Physics* **2013** (Nov., 2013) 52, arXiv:1306.4974 [hep-th].
[13] E. Canovi, E. Ercolessi, P. Naldesi, L. Taddia, and D. Vodola, "Dynamics of entanglement entropy and entanglement spectrum crossing a quantum phase transition," *Phys. Rev. B* **89** no. 10, (Mar., 2014) 104303, arXiv:1311.3612 [cond-mat.stat-mech].
[14] P. Jacquod and C. Petitjean, "Decoherence, entanglement and irreversibility in quantum dynamical systems with few degrees of freedom," *Advances in Physics* **58** no. 2, (2009) 67–196.
[15] S. Akhtar, S. Choudhury, S. Chowdhury, D. Goswami, S. Panda, and A. Swain, "Open Quantum Entanglement: A study of two atomic system in static patch of de Sitter space," *Eur. Phys. J. C* **80** no. 8, (2020) 748, arXiv:1908.09929 [hep-th].
[16] S. Choudhury and S. Panda, "Quantum entanglement in de Sitter space from stringy axion: An analysis using $\alpha$ vacua," *Nucl. Phys. B* **943** (2019) 114606, arXiv:1712.08299 [hep-th].
[17] S. Choudhury and S. Panda, "Entangled de Sitter from stringy axionic Bell pair I: an analysis using Bunch–Davies vacuum," *Eur. Phys. J. C* **78** no. 1, (2018) 52, arXiv:1708.02265 [hep-th].
[18] S. Ghosh, K. S. Gupta, and S. C. L. Srivastava, "Entanglement dynamics following a sudden quench: An exact solution," *EPL* **120** no. 5, (2017) 50005, arXiv:1709.02202 [quant-ph].
[19] S. Ghosh, K. S. Gupta, and S. C. L. Srivastava, "Exact relaxation dynamics and quantum information scrambling in multiply quenched harmonic chains," *Phys. Rev. E* **100** no. 1, (2019) 012215, arXiv:1905.06743 [quant-ph].
[20] L. Susskind, "Three Lectures on Complexity and Black Holes," 10, 2018. arXiv:1810.11563 [hep-th].
[21] A. R. Brown, L. Susskind, and Y. Zhao, "Quantum Complexity and Negative Curvature," *Phys. Rev. D* **95** no. 4, (2017) 045010, arXiv:1608.02612 [hep-th].
[22] A. R. Brown, D. A. Roberts, L. Susskind, B. Swingle, and Y. Zhao, "Holographic Complexity Equals Bulk Action?," *Phys. Rev. Lett.* **116** no. 19, (2016) 191301, arXiv:1509.07876 [hep-th].
[23] A. R. Brown, D. A. Roberts, L. Susskind, B. Swingle, and Y. Zhao, "Complexity, action, and black holes," *Phys. Rev. D* **93** no. 8, (2016) 086006, arXiv:1512.04993 [hep-th].
[24] L. Susskind, "Entanglement is not enough," *Fortsch. Phys.* **64** (2016) 49–71, arXiv:1411.0690 [hep-th].
[25] L. Susskind and Y. Zhao, "Switchbacks and the Bridge to Nowhere," arXiv:1408.2823 [hep-th].
[26] D. Stanford and L. Susskind, "Complexity and Shock Wave Geometries," *Phys. Rev. D* **90** no. 12, (2014) 126007, arXiv:1406.2678 [hep-th].
[27] L. Susskind, "Computational Complexity and Black Hole Horizons," *Fortsch. Phys.* **64** (2016) 24–43, arXiv:1403.5695 [hep-th]. [Addendum: Fortsch.Phys. 64, 44–48 (2016)].
[28] J. L. F. Barbon and E. Rabinovici, "Holographic complexity and spacetime singularities," *JHEP* **01** (2016) 084, arXiv:1509.09291 [hep-th].
[29] M. Alishahiha, "Holographic Complexity," *Phys. Rev. D* **92** no. 12, (2015) 126009, arXiv:1509.06614 [hep-th].
[30] R.-Q. Yang, "Strong energy condition and complexity growth bound in holography," *Phys. Rev. D* **95** no. 8, (2017) 086017, arXiv:1610.05090 [gr-qc].
[31] S. Chapman, H. Marrochio, and R. C. Myers, "Complexity of Formation in Holography," *JHEP* **01** (2017) 062, arXiv:1610.08063 [hep-th].
[32] Y. Zhao, "Complexity and Boost Symmetry," *Phys. Rev. D* **98** no. 8, (2018) 086011, arXiv:1702.03957 [hep-th].
[33] M. Flory, "A complexity/fidelity susceptibility $g$-theorem for $AdS_3/BCFT_2$," *JHEP* **06** (2017) 131, arXiv:1702.06386 [hep-th].
[34] M. Alishahiha and A. Faraji Astaneh, "Holographic Fidelity Susceptibility," *Phys. Rev. D* **96** no. 8, (2017) 086004, arXiv:1705.01834 [hep-th].
[35] A. Reynolds and S. F. Ross, "Complexity in de Sitter Space," *Class. Quant. Grav.* **34** no. 17, (2017) 175013, arXiv:1706.03788 [hep-th].
[36] D. Carmi, S. Chapman, H. Marrochio, R. C. Myers, and S. Sugishita, "On the Time Dependence of Holographic Complexity," *JHEP* **11** (2017) 188, arXiv:1709.10184 [hep-th].
[37] J. Couch, S. Eccles, W. Fischler, and M.-L. Xiao, "Holographic complexity and noncommutative gauge theory," *JHEP* **03** (2018) 108, arXiv:1710.07833

[hep-th].
[38] R.-Q. Yang, C. Niu, C.-Y. Zhang, and K.-Y. Kim, "Comparison of holographic and field theoretic complexities for time dependent thermofield double states," *JHEP* **02** (2018) 082, arXiv:1710.00600 [hep-th].
[39] R. Abt, J. Erdmenger, H. Hinrichsen, C. M. Melby-Thompson, R. Meyer, C. Northe, and I. A. Reyes, "Topological Complexity in $AdS_3/CFT_2$," *Fortsch. Phys.* **66** no. 6, (2018) 1800034, arXiv:1710.01327 [hep-th].
[40] B. Swingle and Y. Wang, "Holographic Complexity of Einstein-Maxwell-Dilaton Gravity," *JHEP* **09** (2018) 106, arXiv:1712.09826 [hep-th].
[41] A. P. Reynolds and S. F. Ross, "Complexity of the AdS Soliton," *Class. Quant. Grav.* **35** no. 9, (2018) 095006, arXiv:1712.03732 [hep-th].
[42] Y.-S. An and R.-H. Peng, "Effect of the dilaton on holographic complexity growth," *Phys. Rev. D* **97** no. 6, (2018) 066022, arXiv:1801.03638 [hep-th].
[43] Z. Fu, A. Maloney, D. Marolf, H. Maxfield, and Z. Wang, "Holographic complexity is nonlocal," *JHEP* **02** (2018) 072, arXiv:1801.01137 [hep-th].
[44] B. Chen, W.-M. Li, R.-Q. Yang, C.-Y. Zhang, and S.-J. Zhang, "Holographic subregion complexity under a thermal quench," *JHEP* **07** (2018) 034, arXiv:1803.06680 [hep-th].
[45] B. Chen, W.-M. Li, R.-Q. Yang, C.-Y. Zhang, and S.-J. Zhang, "Holographic subregion complexity under a thermal quench," *JHEP* **07** (2018) 034, arXiv:1803.06680 [hep-th].
[46] K. Hashimoto, N. Iizuka, and S. Sugishita, "Thoughts on Holographic Complexity and its Basis-dependence," *Phys. Rev. D* **98** no. 4, (2018) 046002, arXiv:1805.04226 [hep-th].
[47] M. Flory and N. Miekley, "Complexity change under conformal transformations in $AdS_3/CFT_2$," *JHEP* **05** (2019) 003, arXiv:1806.08376 [hep-th].
[48] J. Couch, S. Eccles, T. Jacobson, and P. Nguyen, "Holographic Complexity and Volume," *JHEP* **11** (2018) 044, arXiv:1807.02186 [hep-th].
[49] S. A. Hosseini Mansoori, V. Jahnke, M. M. Qaemmaqami, and Y. D. Olivas, "Holographic Complexity of Anisotropic Black Branes," *Phys. Rev. D* **100** no. 4, (2019) 046014, arXiv:1808.00067 [hep-th].
[50] S. Chapman, H. Marrochio, and R. C. Myers, "Holographic complexity in Vaidya spacetimes. Part I," *JHEP* **06** (2018) 046, arXiv:1804.07410 [hep-th].
[51] S. Chapman, H. Marrochio, and R. C. Myers, "Holographic complexity in Vaidya spacetimes. Part II," *JHEP* **06** (2018) 114, arXiv:1805.07262 [hep-th].
[52] E. Caceres, S. Chapman, J. D. Couch, J. P. Hernandez, R. C. Myers, and S.-M. Ruan, "Complexity of Mixed States in QFT and Holography," *JHEP* **03** (2020) 012, arXiv:1909.10557 [hep-th].
[53] O. Ben-Ami and D. Carmi, "On Volumes of Subregions in Holography and Complexity," *JHEP* **11** (2016) 129, arXiv:1609.02514 [hep-th].
[54] F. J. G. Abad, M. Kulaxizi, and A. Parnachev, "On Complexity of Holographic Flavors," *JHEP* **01** (2018) 127, arXiv:1705.08424 [hep-th].
[55] R.-G. Cai, S.-M. Ruan, S.-J. Wang, R.-Q. Yang, and R.-H. Peng, "Action growth for AdS black holes," *JHEP* **09** (2016) 161, arXiv:1606.08307 [gr-qc].
[56] L. Lehner, R. C. Myers, E. Poisson, and R. D. Sorkin, "Gravitational action with null boundaries," *Phys. Rev. D* **94** no. 8, (2016) 084046, arXiv:1609.00207 [hep-th].
[57] M. Moosa, "Evolution of Complexity Following a Global Quench," *JHEP* **03** (2018) 031, arXiv:1711.02668 [hep-th].
[58] M. Moosa, "Divergences in the rate of complexification," *Phys. Rev. D* **97** no. 10, (2018) 106016, arXiv:1712.07137 [hep-th].
[59] K. Hashimoto, N. Iizuka, and S. Sugishita, "Time evolution of complexity in Abelian gauge theories," *Phys. Rev. D* **96** no. 12, (2017) 126001, arXiv:1707.03840 [hep-th].
[60] M. Guo, J. Hernandez, R. C. Myers, and S.-M. Ruan, "Circuit Complexity for Coherent States," *JHEP* **10** (2018) 011, arXiv:1807.07677 [hep-th].
[61] M. Doroudiani, A. Naseh, and R. Pirmoradian, "Complexity for Charged Thermofield Double States," *JHEP* **01** (2020) 120, arXiv:1910.08806 [hep-th].
[62] T. Ali, A. Bhattacharyya, S. S. Haque, E. H. Kim, N. Moynihan, and J. Murugan, "Chaos and Complexity in Quantum Mechanics," *Phys. Rev. D* **101** no. 2, (2020) 026021, arXiv:1905.13534 [hep-th].
[63] S. Choudhury, A. Dutta, and D. Ray, "Chaos and Complexity from Quantum Neural Network: A study with Diffusion Metric in Machine Learning," *JHEP* **04** (2021) 138, arXiv:2011.07145 [hep-th].
[64] S. Chapman and H. Z. Chen, "Charged Complexity and the Thermofield Double State," *JHEP* **02** (2021) 187, arXiv:1910.07508 [hep-th].
[65] P. Bhargava, S. Choudhury, S. Chowdhury, A. Mishara, S. P. Selvam, S. Panda, and G. D. Pasquino, "Quantum aspects of chaos and complexity from bouncing cosmology: A study with two-mode single field squeezed state formalism," arXiv:2009.03893 [hep-th].
[66] J.-L. Lehners and J. Quintin, "Quantum Circuit Complexity of Primordial Perturbations," *Phys. Rev. D* **103** no. 6, (2021) 063527, arXiv:2012.04911 [hep-th].
[67] A. Bhattacharyya, S. Das, S. S. Haque, and B. Underwood, "Rise of cosmological complexity: Saturation of growth and chaos," *Phys. Rev. Res.* **2** no. 3, (2020) 033273, arXiv:2005.10854 [hep-th].
[68] S. Choudhury, S. Chowdhury, N. Gupta, A. Mishara, S. P. Selvam, S. Panda, G. D. Pasquino, C. Singha, and A. Swain, "Circuit Complexity From Cosmological Islands," *Symmetry* **13** (2021) 1301, arXiv:2012.10234 [hep-th].
[69] K. Adhikari, S. Choudhury, S. Chowdhury, K. Shirish, and A. Swain, "Circuit complexity as a novel probe of quantum entanglement: A study with black hole gas in arbitrary dimensions," *Phys. Rev. D* **104** no. 6, (2021) 065002, arXiv:2104.13940 [hep-th].
[70] K. Adhikari, S. Choudhury, S. Chowdhury, K. Shirish, and A. Swain, "Circuit complexity as a novel probe of quantum entanglement: A study with black hole gas in arbitrary dimensions," *Phys. Rev. D* **104** no. 6, (2021) 065002, arXiv:2104.13940 [hep-th].
[71] K. Adhikari, S. Choudhury, H. N. Pandya, and

R. Srivastava, "$\mathcal{PGW}$ Circuit Complexity," arXiv:2108.10334 [gr-qc].

[72] K. Y. Bhagat, B. Bose, S. Choudhury, S. Chowdhury, R. N. Das, S. G. Dastider, N. Gupta, A. Maji, G. D. Pasquino, and S. Paul, "The Generalized OTOC from Supersymmetric Quantum Mechanics—Study of Random Fluctuations from Eigenstate Representation of Correlation Functions," *Symmetry* **13** no. 1, (2020) 44, arXiv:2008.03280 [hep-th].

[73] J. Boruch, P. Caputa, D. Ge, and T. Takayanagi, "Holographic path-integral optimization," *JHEP* **07** (2021) 016, arXiv:2104.00010 [hep-th].

[74] J. Boruch, P. Caputa, and T. Takayanagi, "Path-Integral Optimization from Hartle-Hawking Wave Function," *Phys. Rev. D* **103** no. 4, (2021) 046017, arXiv:2011.08188 [hep-th].

[75] S. Choudhury, S. P. Selvam, and K. Shirish, "Circuit Complexity From Supersymmetric Quantum Field Theory With Morse Function," arXiv:2101.12582 [hep-th].

[76] P. Caputa, N. Kundu, M. Miyaji, T. Takayanagi, and K. Watanabe, "Liouville Action as Path-Integral Complexity: From Continuous Tensor Networks to AdS/CFT," *JHEP* **11** (2017) 097, arXiv:1706.07056 [hep-th].

[77] P. Caputa and J. M. Magan, "Quantum Computation as Gravity," *Phys. Rev. Lett.* **122** no. 23, (2019) 231302, arXiv:1807.04422 [hep-th].

[78] P. Caputa and I. MacCormack, "Geometry and Complexity of Path Integrals in Inhomogeneous CFTs," *JHEP* **01** (2021) 027, arXiv:2004.04698 [hep-th].

[79] R. Khan, C. Krishnan, and S. Sharma, "Circuit Complexity in Fermionic Field Theory," *Phys. Rev. D* **98** no. 12, (2018) 126001, arXiv:1801.07620 [hep-th].

[80] K. Adhikari, S. Choudhury, S. Kumar, S. Mandal, N. Pandey, A. Roy, S. Sarkar, P. Sarker, and S. S. Shariff, "Circuit Complexity in $\mathcal{Z}_2\mathcal{EEFT}$," arXiv:2109.09759 [hep-th].

[81] K. Adhikari, S. Choudhury, and A. Roy, "$\mathcal{K}$rylov $\mathcal{C}$omplexity in $\mathcal{Q}$uantum $\mathcal{F}$ield $\mathcal{T}$heory," arXiv:2204.02250 [hep-th].

[82] K. Adhikari, S. Choudhury, H. N. Pandya, and R. Srivastava, "PGW Circuit Complexity," arXiv:2108.10334 [gr-qc].

[83] S. Choudhury, A. Mukherjee, N. Pandey, and A. Roy, "Causality Constraint on Circuit Complexity from $\mathcal{COSMOEFT}$," arXiv:2111.11468 [hep-th].

[84] R. Jefferson and R. C. Myers, "Circuit complexity in quantum field theory," *JHEP* **10** (2017) 107, arXiv:1707.08570 [hep-th].

[85] L. Hackl and R. C. Myers, "Circuit complexity for free fermions," *JHEP* **07** (2018) 139, arXiv:1803.10638 [hep-th].

[86] S. Chapman, M. P. Heller, H. Marrochio, and F. Pastawski, "Toward a Definition of Complexity for Quantum Field Theory States," *Phys. Rev. Lett.* **120** no. 12, (2018) 121602, arXiv:1707.08582 [hep-th].

[87] A. Bhattacharyya, A. Shekar, and A. Sinha, "Circuit complexity in interacting QFTs and RG flows," *JHEP* **10** (2018) 140, arXiv:1808.03105 [hep-th].

[88] A. Polkovnikov, K. Sengupta, A. Silva, and M. Vengalattore, "Colloquium: Nonequilibrium dynamics of closed interacting quantum systems," *Rev. Mod. Phys.* **83** (Aug, 2011) 863–883.

[89] C. Gogolin and J. Eisert, "Equilibration, thermalisation, and the emergence of statistical mechanics in closed quantum systems," *Reports on Progress in Physics* **79** no. 5, (Apr, 2016) 056001.

[90] P. Calabrese, F. H. L. Essler, and G. Mussardo, "Introduction to 'quantum integrability in out of equilibrium systems'," *Journal of Statistical Mechanics: Theory and Experiment* **2016** no. 6, (Jun, 2016) 064001.

[91] P. Calabrese and J. Cardy, "Quantum Quenches in Extended Systems," *J. Stat. Mech.* **0706** (2007) P06008, arXiv:0704.1880 [cond-mat.stat-mech].

[92] P. Basu and S. R. Das, "Quantum Quench across a Holographic Critical Point," *JHEP* **01** (2012) 103, arXiv:1109.3909 [hep-th].

[93] A. Buchel, L. Lehner, R. C. Myers, and A. van Niekerk, "Quantum quenches of holographic plasmas," *JHEP* **05** (2013) 067, arXiv:1302.2924 [hep-th].

[94] S. R. Das, D. A. Galante, and R. C. Myers, "Universal scaling in fast quantum quenches in conformal field theories," *Phys. Rev. Lett.* **112** (2014) 171601, arXiv:1401.0560 [hep-th].

[95] S. R. Das, D. A. Galante, and R. C. Myers, "Universality in fast quantum quenches," *JHEP* **02** (2015) 167, arXiv:1411.7710 [hep-th].

[96] S. R. Das, D. A. Galante, and R. C. Myers, "Smooth and fast versus instantaneous quenches in quantum field theory," *JHEP* **08** (2015) 073, arXiv:1505.05224 [hep-th].

[97] S. R. Das, D. A. Galante, and R. C. Myers, "Quantum Quenches in Free Field Theory: Universal Scaling at Any Rate," *JHEP* **05** (2016) 164, arXiv:1602.08547 [hep-th].

[98] V. Alba and P. Calabrese, "Entanglement dynamics after quantum quenches in generic integrable systems," *SciPost Physics* **4** no. 3, (Mar., 2018) 017, arXiv:1712.07529 [cond-mat.stat-mech].

[99] S. Choudhury, R. M. Gharat, S. Mandal, N. Pandey, A. Roy, and P. Sarker, "Entanglement in interacting quenched two-body coupled oscillator system," *Phys. Rev. D* **106** no. 2, (July, 2022) 025002, arXiv:2204.05326 [hep-th].

[100] H. A. Camargo, P. Caputa, D. Das, M. P. Heller, and R. Jefferson, "Complexity as a Novel Probe of Quantum Quenches: Universal Scalings and Purifications," *Phys. Rev. Lett.* **122** no. 8, (Mar., 2019) 081601, arXiv:1807.07075 [hep-th].

[101] D. W. F. Alves and G. Camilo, "Evolution of complexity following a quantum quench in free field theory," *Journal of High Energy Physics* **2018** no. 6, (June, 2018) 29, arXiv:1804.00107 [hep-th].

[102] T. Langen, T. Gasenzer, and J. Schmiedmayer, "Prethermalization and universal dynamics in near-integrable quantum systems," *Journal of Statistical Mechanics: Theory and Experiment* **2016** no. 6, (Jun, 2016) 064009.

[103] T. Kinoshita, T. Wenger, and D. Weiss, "A quantum newton's cradle," *Nature* **440** (05, 2006) 900–3.

[104] S. Hofferberth, I. Lesanovsky, B. Fischer, T. Schumm, and J. Schmiedmayer, "Non-equilibrium coherence dynamics in one-dimensional Bose gases," *Nature (London)* **449** no. 7160, (Sept., 2007) 324–327,

arXiv:0706.2259 [cond-mat.other].
[105] S. Trotzky, Y. A. Chen, A. Flesch, I. P. McCulloch, U. Schollwöck, J. Eisert, and I. Bloch, "Probing the relaxation towards equilibrium in an isolated strongly correlated one-dimensional Bose gas," *Nature Physics* **8** no. 4, (Apr., 2012) 325–330, arXiv:1101.2659 [cond-mat.quant-gas].
[106] M. Gring, M. Kuhnert, T. Langen, T. Kitagawa, B. Rauer, M. Schreitl, I. Mazets, D. A. Smith, E. Demler, and J. Schmiedmayer, "Relaxation and prethermalization in an isolated quantum system," *Science* **337** no. 6100, (2012) 1318–1322.
[107] M. Cheneau, P. Barmettler, D. Poletti, M. Endres, P. Schauß, T. Fukuhara, C. Gross, I. Bloch, C. Kollath, and S. Kuhr, "Light-cone-like spreading of correlations in a quantum many-body system," *Nature (London)* **481** no. 7382, (Jan., 2012) 484–487, arXiv:1111.0776 [cond-mat.quant-gas].
[108] F. Meinert, M. J. Mark, E. Kirilov, K. Lauber, P. Weinmann, A. J. Daley, and H.-C. Nägerl, "Quantum quench in an atomic one-dimensional ising chain," *Phys. Rev. Lett.* **111** (Jul, 2013) 053003.
[109] T. Langen, R. Geiger, M. Kuhnert, B. Rauer, and J. Schmiedmayer, "Local emergence of thermal correlations in an isolated quantum many-body system," *Nature Physics* **9** no. 10, (Oct., 2013) 640–643, arXiv:1305.3708 [cond-mat.quant-gas].
[110] T. Fukuhara, P. Schauß, M. Endres, S. Hild, M. Cheneau, I. Bloch, and C. Gross, "Microscopic observation of magnon bound states and their dynamics," *Nature (London)* **502** no. 7469, (Oct., 2013) 76–79, arXiv:1305.6598 [cond-mat.quant-gas].
[111] T. Fukuhara, A. Kantian, M. Endres, M. Cheneau, P. Schauß, S. Hild, D. Bellem, U. Schollwöck, T. Giamarchi, C. Gross, I. Bloch, and S. Kuhr, "Quantum dynamics of a mobile spin impurity," *Nature Physics* **9** no. 4, (Apr., 2013) 235–241, arXiv:1209.6468 [cond-mat.quant-gas].
[112] H. R. Lewis and W. B. Riesenfeld, "An exact quantum theory of the time-dependent harmonic oscillator and of a charged particle in a time-dependent electromagnetic field," *Journal of Mathematical Physics* **10** no. 8, (1969) 1458–1473.
[113] K. Andrzejewski, "Dynamics of entropy and information of time-dependent quantum systems: exact results," *Quant. Inf. Proc.* **21** no. 3, (2022) 117, arXiv:2108.00975 [quant-ph].
[114] B. Khantoul, A. Bounames, and M. Maamache, "On the invariant method for the time-dependent non-Hermitian Hamiltonians," *European Physical Journal Plus* **132** no. 6, (June, 2017) 258, arXiv:1610.09273 [quant-ph].
[115] S. Choudhury, "Cosmological Geometric Phase From Pure Quantum States: A study without/with having Bell's inequality violation," arXiv:2105.06254 [gr-qc].
[116] J.-R. Choi, "Coherent and squeezed states for light in homogeneous conducting linear media by an invariant operator method," *International Journal of Theoretical Physics* **43** no. 10, (2004) 2113–2136.
[117] H. Kanasugi and H. Okada, "Systematic Treatment of General Time-Dependent Harmonic Oscillator in Classical and Quantum Mechanics," *Progress of Theoretical Physics* **93** no. 5, (05, 1995) 949–960.
[118] "Adiabatic evolution under quantum control," *Annals of Physics* **327** no. 5, (2012) 1293–1303.
[119] M.-Y. Ye, X.-F. Zhou, Y.-S. Zhang, and G.-C. Guo, "Two kinds of quantum adiabatic approximation," *Physics Letters A* **368** no. 1, (2007) 18–24.
[120] J. R. Choi, "Perturbation theory for time-dependent quantum systems involving complex potentials," *Frontiers in Physics* **8** (June, 2020) 189.
[121] V. P. Ermakov, "Second-order differential equations: conditions of complete integrability," *Applicable Analysis and Discrete Mathematics* (2008) 123–145.
[122] W. Milne, "The numerical determination of characteristic numbers," *Physical Review* **35** no. 7, (1930) 863.
[123] E. Pinney, "The nonlinear differential equation y+ p (x) y+ cy- 3= 0," in *Proc. Amer. Math. Soc*, vol. 1, pp. 681–681. 1950.
[124] M. A. Nielsen, "A geometric approach to quantum circuit lower bounds," quant-ph/0502070.
[125] M. A. Nielsen, "Quantum computation as geometry," *Science* **311** no. 5764, (Feb, 2006) 1133–1135, quant-ph/0603161.
[126] M. R. Dowling and M. A. Nielsen, "The geometry of quantum computation," *Quantum Info. Comput.* **8** no. 10, (Nov., 2008) 861–899, quant-ph/070100.
[127] M. A. Nielsen, M. R. Dowling, M. Gu, and A. C. Doherty, "Optimal control, geometry, and quantum computing," *Phys. Rev. A* **73** (Jun, 2006) 062323, quant-ph/0603160.
[128] M. A. Nielsen, "A Geometric Approach to Quantum Circuit Lower Bounds," arXiv:quant-ph/0502070 [quant-ph].
[129] R. A. Jefferson and R. C. Myers, "Circuit complexity in quantum field theory," *Journal of High Energy Physics* **2017** no. 10, (Oct., 2017) 107, arXiv:1707.08570 [hep-th].
[130] P. Caputa, S. R. Das, M. Nozaki, and A. Tomiya, "Quantum Quench and Scaling of Entanglement Entropy," *Phys. Lett. B* **772** (2017) 53–57, arXiv:1702.04359 [hep-th].
[131] H. A. Camargo, P. Caputa, D. Das, M. P. Heller, and R. Jefferson, "Complexity as a novel probe of quantum quenches: Universal scalings and purifications," *Phys. Rev. Lett.* **122** (Feb, 2019) 081601.
[132] K. H. Yeon, H. J. Kim, C. I. Um, T. F. George, and L. N. Pandey, "Wave function in the invariant representation and squeezed-state function of the time-dependent harmonic oscillator," *Phys. Rev. A* **50** no. 2, (Aug., 1994) 1035–1039.
[133] S. Mukherjee, A. Ghose Choudhury, and P. Guha, "Generalized damped Milne-Pinney equation and Chiellini method," *arXiv e-prints* (Mar., 2016) arXiv:1603.08747, arXiv:1603.08747 [nlin.SI].
[134] D. E. Parker, X. Cao, A. Avdoshkin, T. Scaffidi, and E. Altman, "A Universal Operator Growth Hypothesis," *Phys. Rev. X* **9** no. 4, (2019) 041017, arXiv:1812.08657 [cond-mat.stat-mech].
[135] P. Caputa and S. Liu, "Quantum complexity and topological phases of matter," arXiv:2205.05688 [hep-th].
[136] P. Caputa, J. M. Magan, and D. Patramanis, "Geometry of Krylov complexity," *Phys. Rev. Res.* **4** no. 1, (2022) 013041, arXiv:2109.03824 [hep-th].

[137] K. Adhikari, S. Choudhury, and A. Roy, "$\mathcal{K}$rylov $\mathcal{C}$omplexity in $\mathcal{Q}$uantum $\mathcal{F}$ield $\mathcal{T}$heory," arXiv:2204.02250 [hep-th].

[138] K. Adhikari and S. Choudhury, "$\mathcal{C}$osmological $\mathcal{K}$rylov $\mathcal{C}$omplexity," arXiv:2203.14330 [hep-th].

[139] K. Adhikari, S. Choudhury, S. Chowdhury, K. Shirish, and A. Swain, "Circuit complexity as a novel probe of quantum entanglement: A study with black hole gas in arbitrary dimensions," *Phys. Rev. D* **104** no. 6, (2021) 065002, arXiv:2104.13940 [hep-th].

[140] S. Choudhury, S. Chowdhury, N. Gupta, A. Mishara, S. P. Selvam, S. Panda, G. D. Pasquino, C. Singha, and A. Swain, "Circuit Complexity from Cosmological Islands," *Symmetry* **13** no. 7, (2021) 1301, arXiv:2012.10234 [hep-th].

[141] J. Eisert, "Entangling Power and Quantum Circuit Complexity," *Phys. Rev. Lett.* **127** no. 2, (July, 2021) 020501, arXiv:2104.03332 [quant-ph].

[142] S. D. Mathur, "Three puzzles in cosmology," *International Journal of Modern Physics D* **29** no. 14, (Jan., 2020) 2030013, arXiv:2009.09832 [hep-th].

FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 460-465 | e-ISSN: 2347-470X

# Comparative Analysis of Particle Swarm Optimization and Artificial Neural Network Based MPPT with Variable Irradiance and Load

**Spandan Srivastava[1], Charu Lata[2], Prateek Lohan[3] and Rinchin W. Mosobi[4]**

*[1,2,3,4]Electrical Engineering Department, Delhi Technological University (DTU), New Delhi, India, spandansrivastava_2k18ee208@dtu.ac.in, charulata_2k18ee056@dtu.ac.in, prateeklohan_2k18ee143@dtu.ac.in, wangzom@dtu.ac.in.*

***Correspondence:** Spandan Srivastava; Email: spandansrivastava_2k18ee208@dtu.ac.in

**ABSTRACT-** The escalating demands and increasing awareness for the environment, resulted in deployment of Photovoltaic (PV) system as a viable option. PV system are widely installed for numerous applications. However, the challenges in tracking the maximum power with intermittent atmospheric condition and varying load is significant. Maximum Power Point Tracking (MPPT) algorithms are employed and based on their convergence speed, control of external variations and oscillation, the output power efficiency, and other significant factors viz. the algorithm complexity and implementation cost, novel MPPT approach are preferable than the conventional approach. This paper presents an artificial intelligence-based optimization controller for MPPT in a PV system under varying load and irradiance conditions. Comparative analysis of Artificial Neural Network (ANN) and Particle Swarm Optimization (PSO) based MPPT is simulated and analysed. The PV system consisting of PV array and boost converter with MPPT controller feeds the DC load. The power conversion and panel efficiency are the significant factors to determine the effectiveness of tracking maximum power point. The simulation results show the performance of these controllers on the PV panel output power and the load side output power under changing loads and irradiance. In addition, the comparison of PV panel efficiency of ANN and PSO based MPPT techniques w.r.t changing loads is carried out. Based on the above analysis, PSO based MPPT algorithm marginally outperforms the ANN based MPPT algorithm. Further, the implementation of hybrid MPPT (ANN&PSO) for higher accuracy and tracking capability can be carried out as future work.

**Keywords:** Maximum Power Point Tracking (MPPT); Artificial Neural Network (ANN); Particle Swarm Optimization (PSO); Photovoltaic (PV) System.

ARTICLE INFORMATION

**Author(s)**: Spandan Srivastava, Charu Lata, Prateek Lohan and Rinchin W. Mosobi.
**Received**: 28/04/2022; **Accepted**: 20/07/2022; **Published**: 10/08/2022;
**e-ISSN**: 2347-470X;
**Paper Id**: IJEER100309;
**Citation:** 10.37391/IJEER.100309
**Webpage-link**:
https://ijeer.forexjournal.co.in/archive/volume-10/ijeer-100309.html

This article belongs to the Special Issue on **Recent Advancements in the Electrical & Electronics Engineering**

**Publisher's Note:** FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

## 1. INTRODUCTION

Energy employment and environmental issues resulting from global warming and pollution have paved the way for renewable energy sources. Photovoltaic (PV) system being unlimited, clean, and a free source of solar energy find various application. As a result, photovoltaic (PV) energy is the current trend in the electric power industry, and is widely installed globally. However, it suffers from two major issues: low energy conversion efficiency and a nonlinear change in the operating point on the I-V and P-V curves, known as the Maximum Power Point (MPP) [1]. As a result, Maximum Power Point Tracking (MPPT) approach with perturb and observe algorithm [2], artificial neural network (ANN) [3], Particle Swarm Optimization (PSO) [4,5], fuzzy logic control [6], and others are majorly used techniques to track MPP. The differences among these algorithms can be known by their convergence speed, convergence speed, control of external variations and oscillation, maximum output power efficiency, algorithm complexity and implementation cost.

In this paper, a comparative analysis of MPPT technique with ANN and PSO is performed under variable solar irradiance and load conditions. The PV system comprising of solar array and boost converter with MPPT controller feeds the DC load. The modelling and simulation are carried on MATLAB platform and the results indicates the effectiveness of this MPPT algorithm in terms of PV Panel efficiency.

## 2. PHOTOVOLTAIC MODELING

### 2.1 Basis of Photovoltaic Modeling

The ideal PV cell circuit is shown in Fig 1 which is further detailed in [1, 2] The ideal PV cell is mathematically represented by the following basic equation (1) and (2) and further explanation is detailed in [1, 2]:

$$I = I_{pv,cell} - I_{0,cell}\left[exp\left(\frac{qV}{akT}\right) - 1\right] \quad (1)$$

$$I_d = I_{0,cell}\left[exp\left(\frac{qV}{akT}\right) - 1\right] \quad (2)$$

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 460-465 | e-ISSN: 2347-470X

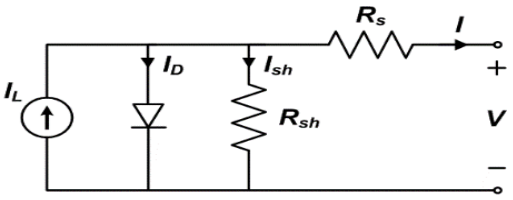



**Figure 1:** Equivalent PV circuit

## 2.2 Modeling of PV Array

Since practical arrays are made of several linked PV cells, their characteristics at the PV terminal demands addition of extra parameters to the fundamental as in (3) and (4)

$$I = I_{PV} - I_0\left[exp\left(\frac{V+R_SI}{V_t a}\right) - 1\right] \quad (3)$$

$$I_{PV} = \left(I_{PV,n} + K_I\Delta T\right)\frac{G}{G_n} \quad (4)$$

The overall stepwise mathematical modeling of PV array is further explained in [1,2,7]

# 3. MPPT ALGORITHMS

MPPT converters employs an algorithm to constantly identify the maximum instantaneous power of the PV array [8]. Since the input and output variables of the array (solar irradiation and load) are intermittent throughout, an MPPT algorithm is needed to extract and transmit the maximum instantaneous power to the load. PV array with DC-DC converter and MPPT control circuit [9] is shown in Fig. 2. MPPT [8-9] techniques are broadly classified into conventional and novel MPPT methods.

- Conventional Methods - Direct (offline) and Indirect (online) Methods
- Novel (soft computing) Methods – Artificial intelligence (ANN), biologically inspired (PSO) etc.

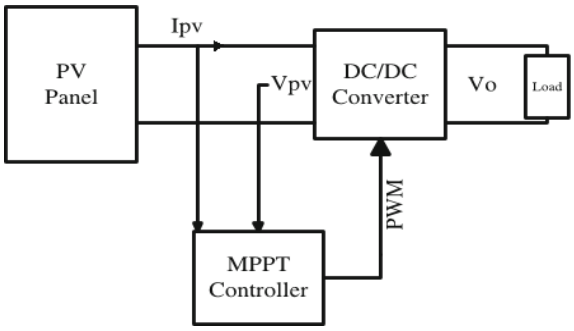



**Figure 2:** MPPT connected PV array feeding DC-DC converter [10]

By repeatedly perturbing the operational point of the PV array, the MPP is determined using direct procedures, also known as genuine searching methods. In this area, the Perturb and Observe, Hill Climbing [11] and Incremental Conductance methods are frequently utilized. Various novel (soft computing) optimization algorithms were employed to obtain the MPP including artificial intelligence and biologically inspired.

Artificial Intelligence and indirect approaches improve the dynamic performance of MPP tracking. Artificial Intelligence algorithms that focus on the nonlinear features of PV arrays provide a rapid, but computationally challenging solution to the MPPT problem. The indirect methods derive the array's MPP from its output properties. The MPP can be determined using the simple and effective fractional open-circuit voltage and short-circuit current approaches.

## 3.1 Artificial Neural Network (ANN)

ANN applications would require no particular facts regarding the device and mathematical modelling. In this manner it copes with more complicated issues by mapping the input-output of the device. ANN is the intelligence-primarily based superior MPPT technique, because of its inherent nature of the mastering method and organic nature of neurons. A three – layer neural network is used for MPPT, as shown in *Figure 3* [12-13]. Temperature (T) and Irradiance (G) are the input variables and the voltage of MPP (Vmpp) is the output variable.

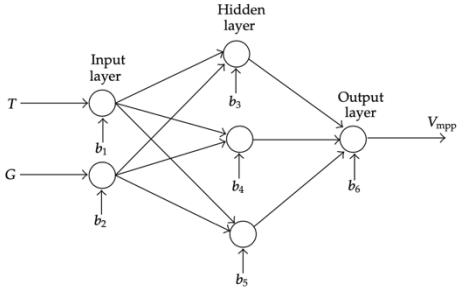



**Figure 3:** Neural Network Structure

The Levenberg-Marquardt backpropagation algorithm has been used to train the Neural Network. Trial and error technique is used to find the number of hidden layers. The inputs are 1000x2 matrix representing static data consisting of 1000 samples of 2 elements Temperature T and Irradiance G. The targets are 1000x1 matrix representing static data consisting of 1000 samples of 1 element $V_{mpp}$. In addition, current of MPP ($I_{mpp}$) can be determined by way of using the I-V curve of the modelled PV, and the maximum power $P_{max}$ is calculated by multiplying $V_{mpp}$ and $I_{mpp}$. At every instant of control, chopper with specified value of $V_{mpp}$, $I_{mpp}$, $I_{out}$ and $V_{out}$, the duty cycle is derived as:

$$D = 1 - \sqrt{\frac{V_{mpp}}{I_{mpp}} \times \frac{I_{out}}{V_{out}}} \quad (5)$$

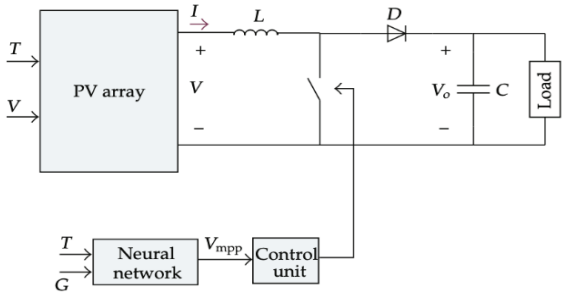



**Figure 4:** PV array with DC-DC converter and its control unit

## 3.2 Particle Swarm Optimization (PSO)

The PSO algorithm's movement is inspired by the behavior of flocking birds, which is based on the PSO optimizer's individual and neighboring experiences throughout each

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 460-465 | e-ISSN: 2347-470X

particle step. The PSO algorithm procedure is separated into four phases. The optimizer initiates the search with a random particle value in the first case. This particle value is chosen depending on the degree of solution space potential for numerous different optimizations. It compares the previous and next best fitness values (Pbi) and (Pli), respectively, in the second step to look for optimum solutions in the same space [13]. The global best positions (Gbi) are compared in the third stage to select the global fitness value. These positions are mathematically adapted and recorded for the subsequent phase, as stated in (6) and (7).

$$V_i^{k+1} = \omega \times V_i^k + r_1 \times c_1 \times \left(P_{bi} - X_i^k\right) + r_2 \times c_2 \times \left(G_{bi} - X_i^k\right) \quad (6)$$

$$X_i^{k+1} = X_i^k + V_i^k \quad (7)$$

where $X_i$ represents each particle's current position, $V_i$ represents the search space's speed, $i$ represents the optimization vector, $k$ represents the number of iterations, $\omega$ is the speed's inertia weight factor, $c_1$ is the single particle's cognitive coefficient, $c_2$ is the total particle's social coefficient, and $r_1$, $r_2$ are the search space's random velocity in the range of 0 to maximum value. The best particle in terms of fitness evaluation is selected and saved in the fourth step to improve the particle movement steps in each iteration. Those steps keep going until a stopping condition is reached or the number of iterations based on the system's necessary accuracy and control processing time, the stopping condition and iterations number are proposed. The algorithm is depicted in *Figures 5&6.*

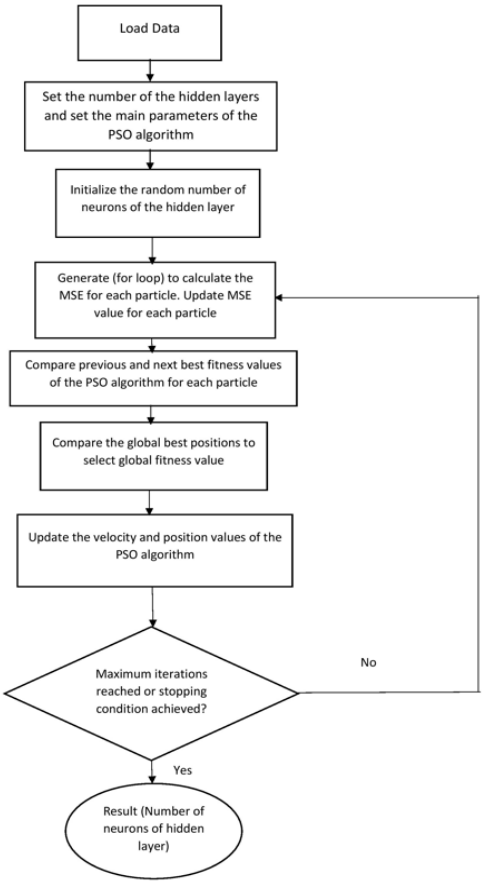



**Figure 5:** Algorithm to find the Hidden Layers of the PSO

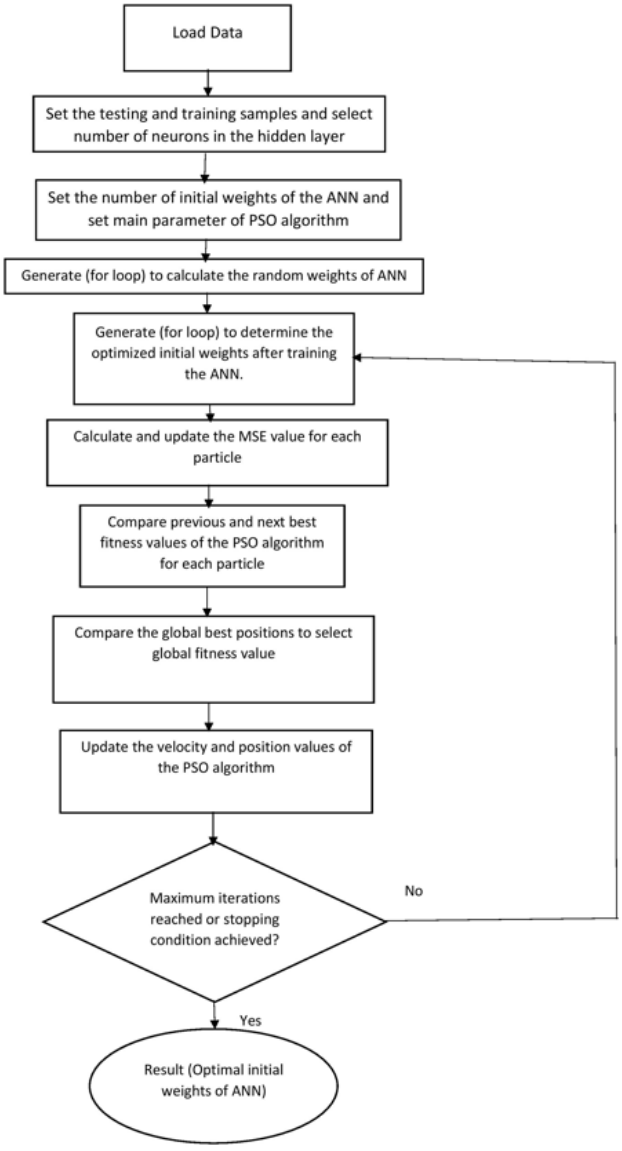



**Figure 6:** Algorithm to find the Initial Weights of the PSO

## 4. SIMULATION RESULT AND ANALYSIS

To analyse the implementation of the ANN and PSO MPPT algorithms in the PV system, a comparison under varying load and solar irradiance is performed. The power conversion and panel efficiency are the significant factor to determine the effectiveness of tracking maximum power point. The power output of PV panel and the load side power output, along with PV panel efficiency are determined in each case. *Table-I* lists the parameters of PV array. Fig. 9 depicts the I-V characteristics of PV array under varying irradiance i.e,1000 W/m$^2$, 800 W/m$^2$, 600 W/m$^2$, 400 W/m$^2$ and 200 W/m$^2$ at an interval of 0.2 sec with load remaining constant at 40 ohms. With increased irradiance, the PV current changes drastically with slight variation in PV voltage. Fig.10 depicts the P-V curve with variable irradiance and constant load at 40 ohms. The MATLAB Simulink of the PV system with ANN and PSO based MPPT are shown in *Figures 7&8* respectively.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 460-465 | e-ISSN: 2347-470X

**Table 1: Solar Array Parameters**

| Parameter | Duration (Sec) |
|---|---|
| Current of MPP, $I_{mpp}$ | 8.15 A |
| Voltage of MPP, $V_{mpp}$ | 30.7 V |
| Maximum power, $P_{max}$ | 250.205 W |
| Short circuit current, $I_{sc}$ | 8.66 A |
| Open circuit voltage, $V_{oc}$ | 37.3 V |
| Equivalent series resistance, $R_p$ | 224.1886 Ω |
| Equivalent parallel resistance, $R_s$ | 0.23724 |

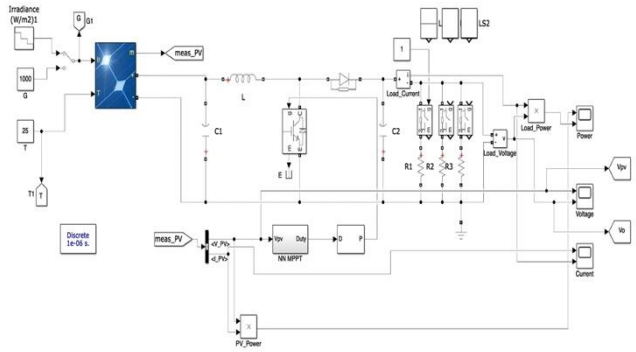


**Figure 7:** Simulation of MPPT using Neural Network

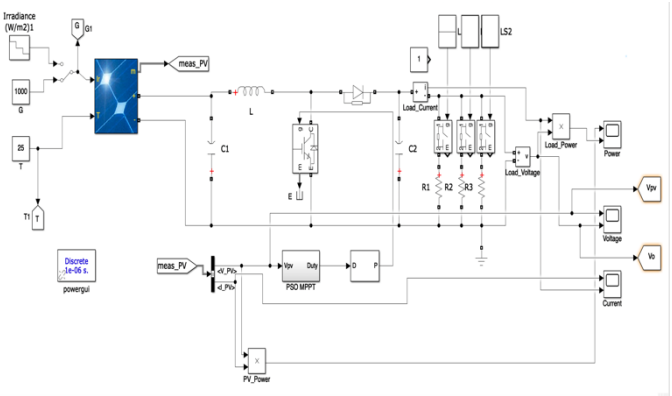


**Figure 8:** Simulation of MPPT using Particle Swarm Optimization

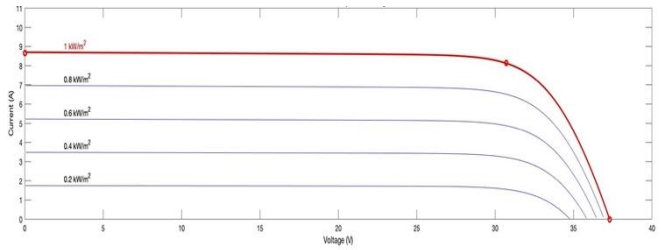


**Figure 9:** I-V characteristics of Solar Array

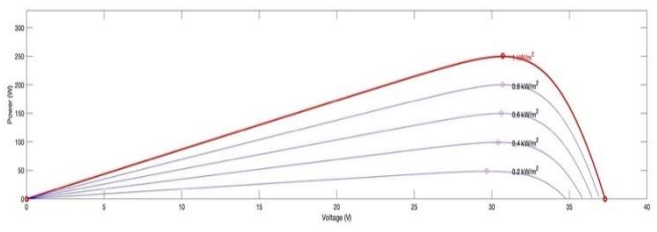


**Figure 10:** P-V characteristics of Solar Array

## 4.1 Artificial Neural Network

The performance under different loads with ANN has been illustrated in *Table 1* and the plots for power output, voltage and current in the load side is shown in *Figures 11-13*. The load is changed from 20 Ω to 30 Ω to 40 Ω at 0.3 second intervals, with the irradiance remaining constant at 1000 W/m$^2$. Similarly, *Table 3* illustrated the results for varying irradiances and constant load. *Figures 14-16* shows the variation of power output, voltage and current in the load side with irradiance changing from 1000 W/m$^2$ to 800 W/m$^2$ to 600 W/m$^2$ to 400 W/m$^2$ to 200 W/m$^2$ at an interval of 0.2 sec.

**Table 2: Performance at different loads for ANN**

| Load (Ω) | Power Output of PV Panel (W) | Power obtained at load side (W) | Efficiency of PV Panel |
|---|---|---|---|
| 20 | 245.7 | 245.7 | 98.19 |
| 30 | 246.2 | 245.3 | 98.03 |
| 40 | 246.6 | 244.9 | 97.87 |

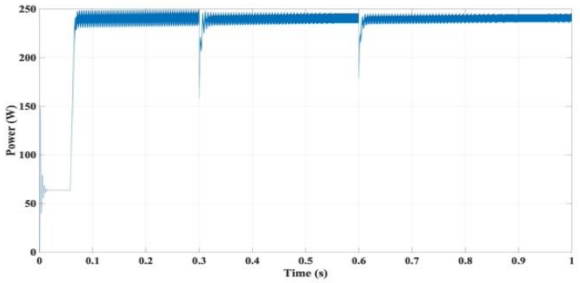



**Figure 11:** Load Side Power with Varying Loads

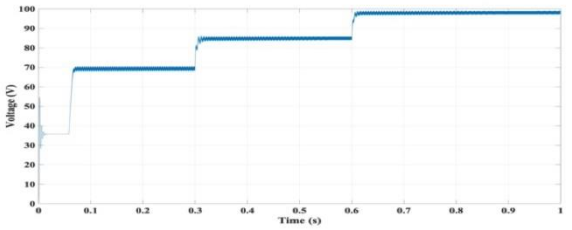



**Figure 12:** Load Side Voltage with Varying Loads

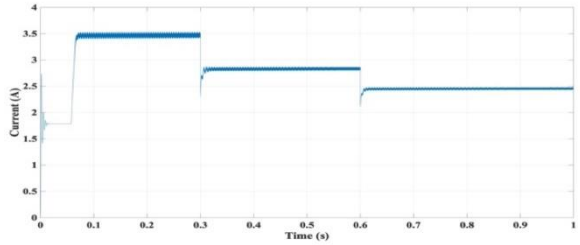



**Figure 13:** Load Side Current with Varying Loads

**Table 3. Performance at different irradiances for ANN**

| Irradiance (W/m$^2$) | Power Output of PV Panel (W) | Power Output of Load Side (W) | Rated Power (W) | Efficiency of PV Panel |
|---|---|---|---|---|
| 1000 | 246.2 | 242.3 | 250.2 | 96.84 |
| 800 | 196.1 | 193.9 | 199.9 | 96.99 |
| 600 | 146.8 | 144.5 | 149.6 | 96.59 |
| 400 | 96.93 | 94.68 | 98.97 | 95.66 |
| 200 | 46.22 | 45.06 | 48.37 | 93.15 |

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 460-465 | e-ISSN: 2347-470X

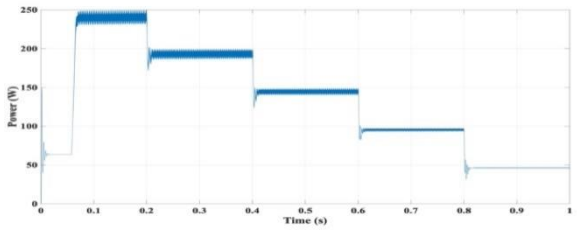



**Figure 14:** Load Side Power with Varying Irradiance

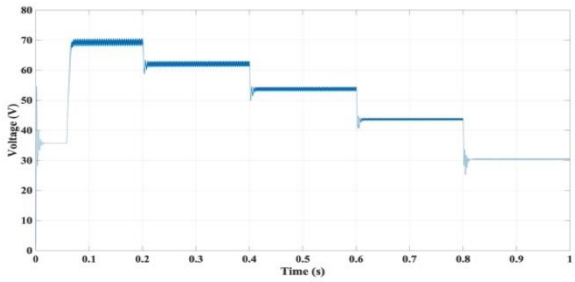



**Figure 15:** Load Side Voltage with Varying Irradiance

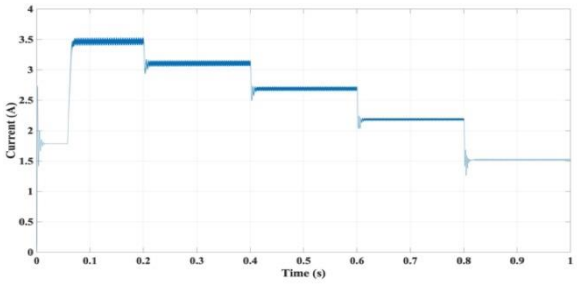



**Figure 16:** Load Side Current with Varying Irradiance

## 4.2 Particle Swarm Optimization (PSO)

The performance analysis of PSO under variable resistive loads has been shown in *Table IV* and the graphs for output power, voltage and current of load side with variable loads is displayed in *Figures 17-19*. Similarly, the results for varying irradiance are being illustrated in *Table V* and their respective plots in *Figures 20-22*.

**Table 4. Performance at different loads for PSO**

| Load (Ω) | Power Output of PV Panel (W) | Power obtained at load side (W) | Efficiency of PV Panel |
|---|---|---|---|
| 20 | 246.2 | 246.9 | 98.67 |
| 30 | 246.4 | 246.2 | 98.39 |
| 40 | 246.2 | 245.6 | 98.15 |

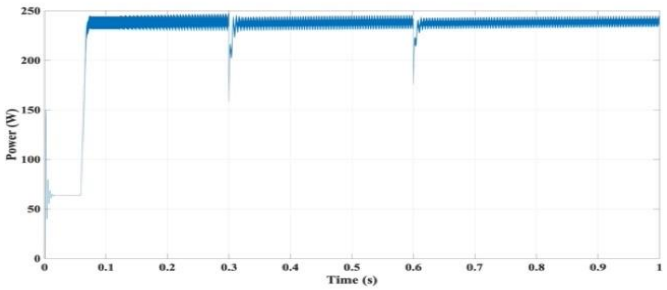



**Figure 17:** Load Side Power with Varying Loads

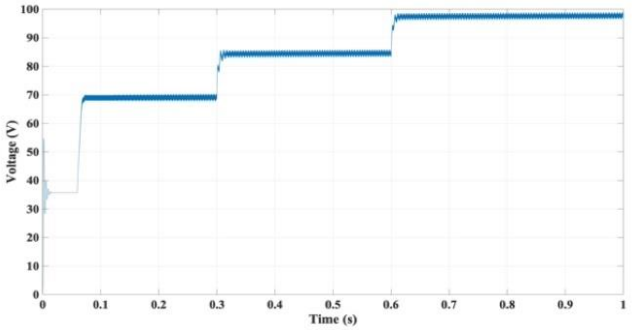



**Figure 18:** Load Side Voltage with Varying Loads

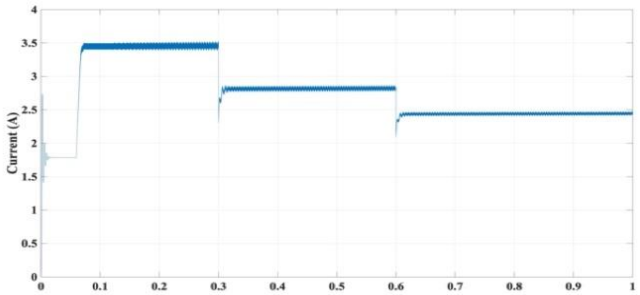



**Figure 19:** Load Side Current with Varying Loads

**Table 5: Performance at different Irradiances for PSO**

| Irradiance (W/m$^2$) | Power Output of PV Panel (W) | Power Output of Load Side (W) | Rated Power (W) | Efficiency of PV Panel |
|---|---|---|---|---|
| 1000 | 244 | 243.2 | 250.2 | 97.20 |
| 800 | 196.3 | 194.6 | 199.9 | 97.34 |
| 600 | 144.9 | 144.7 | 149.6 | 96.72 |
| 400 | 97.25 | 94.82 | 98.97 | 95.79 |
| 200 | 47.66 | 46.19 | 48.37 | 95.49 |

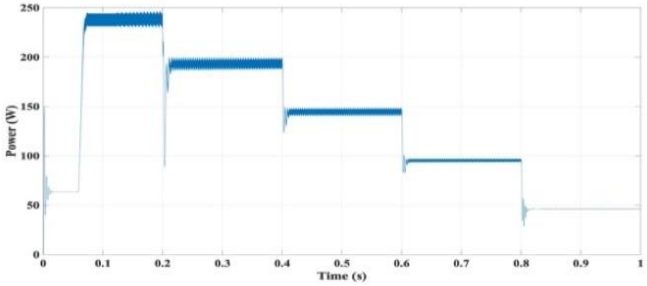



**Figure 20:** Load Side Power with Varying Irradiance

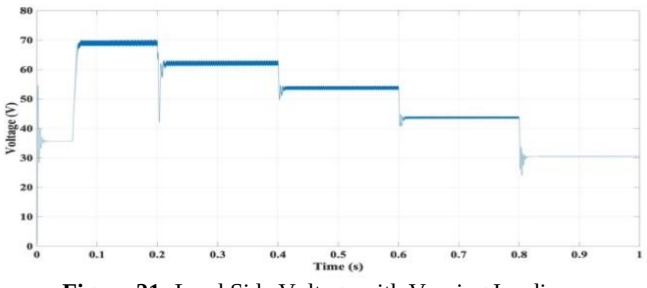



**Figure 21:** Load Side Voltage with Varying Irradiance

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 460-465 | e-ISSN: 2347-470X

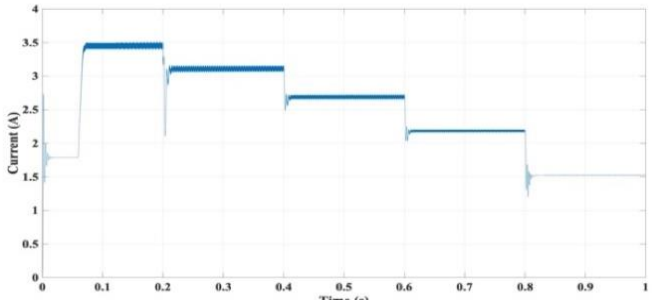



**Figure 22:** Plot of Load Side Current with Varying Irradiance

The panel efficiency w.r.t changing loads is plotted for both the techniques as shown in *Figure 23* and with varying irradiance as shown in *Figure 24*. As seen from the graphs, PSO based MPPT technique gives better performance compared to ANN algorithm.

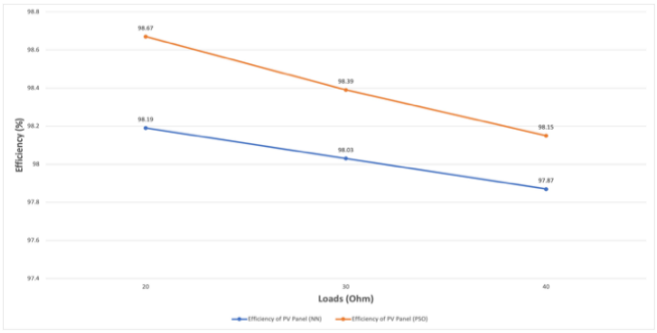



**Figure 23:** Comparison of Panel Efficiency for ANN and PSO for Varying Loads

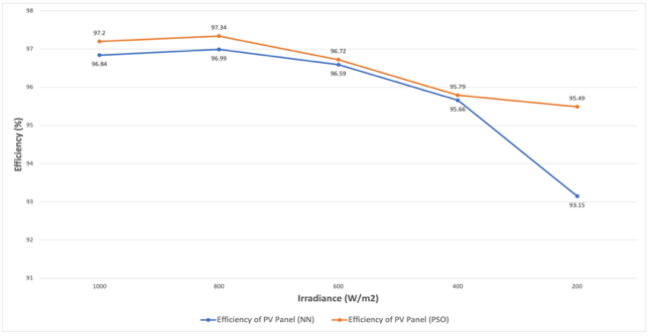



**Figure 24:** Comparison of Panel Efficiency for ANN and PSO for Varying Irradiance

## 5. CONCLUSION

In this paper, a comparative analysis on the implementation of ANN and PSO MPPT algorithms in the PV system is presented, under the condition of varying load and solar irradiance. The test system includes a PV array with boost converter and its control connected to a resistive load and modelled in MATLAB Simulink. With the results on power conversion and panel efficiency of the PV system for ANN and PSO based MPPT, PSO based MPPT technique marginally outperforms the ANN based MPPT technique. At an irradiance of 1000W/m$^2$, the panel efficiency with PSO based MPPT is 97.20%, while ANN based MPPT gave an efficiency of 96.84% and so on for different irradiance value. Similarly with variable load the maximum panel efficiency for PSO MPPT is 98.67% and for ANN MPPT is 98.19%.

Thus, PSO technique has proved to be a more efficient and effective to track the maximum power point of the PV system as compared to the ANN method. Future aspects of the proposed work include the implementation of hybrid MPPT (ANN+PSO) for higher accuracy and tracking capability.

## REFERENCES

[1] Li, X., Wen, H., Hu, Y., Du, Y., & Yang, Y., "A comparative study on photovoltaic MPPT algorithms under EN50530 dynamic test procedure". IEEE Transactions on Power Electronics, 36(4), 4153-4168, 2020.

[2] Chitransh, A., Kumar, S., "The Different Type of MPPT Techniques for Photovoltaic System", Indian Journal of Engineering and Materials Sciences, 2021

[3] Haq IU, Khan Q, Ullah S, Khan SA, Akmeliawati R, Khan MA, et al., "Neural network-based adaptive global sliding mode MPPT controller design for stand-alone photovoltaic systems", PLOS ONE 17(1): e0260480, 2022

[4] Podder, A. K., Roy, N. K., & Pota, H. R., "MPPT methods for solar PV systems: a critical review based on tracking nature", IET Renewable Power Generation, 13(10), 1615-1632, 2019

[5] W.Hayder, A.Abid, M. Hamed and L. Sbita, "Improved PSO Algorithms in PV System Optimisation",European Journal of Electrical Engineering and Computer Science, 2020.

[6] Abdellatif, W., Mohamed, M., Barakat, S., Brisha, Ayman., "A Fuzzy Logic Controller Based MPPT Technique for Photovoltaic Generation System", International Journal on Electrical Engineering and Informatics, 13, 394 – 417, 2021

[7] S. A. Nasirudeen, M. S. Haruna, M. A. Aminu and O. Osanaiye, "Mathematical Modelling, Simulation and Analysis of Solar PV module in Simulink," 2021 1st International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS), 2021, pp. 1-5, 2021

[8] Abo-Sennah M.A., El-Dabah M.A., Mansour A.E.-B., "Maximum power point tracking techniques for photovoltaic systems: A comparative study", International Journal of Electrical and Computer Engineering, vol. 11, no. 1, p. 57-73, 2021

[9] Prabhu, U., Ramesh Babu, Muthu., "Performance Study of MPPT Algorithms of DC-DC Boost Converters For PV Cell Applications", 7th International Conference on Electrical Energy Systems (ICEES), 201-205, 2021

[10] Raj, A., Arya, S. R., Gupta, J., "Solar PV array-based DC–DC converter with MPPT for low power applications", Renewable Energy Focus, 2020

[11] Mishra, S., Mohanty, D., Sahu, J., Mohanty, S., "Artificial Neural Network Based MPPT Controller for Stand-alone Solar PV System", International Conference on System, Computation, Automation and Networking, 2021

[12] Thangavel, S., "Neural Network Based MPPT Controller for Solar PV Connected Induction Motor", IAES International Journal of Robotics and Automation (IJRA), 2018

[13] Almajidi,S., Abbod,M., Al-Raweshidy, H., "A particle swarm optimisation-trained feedforward neural network for predicting the maximum power point of a photovoltaic array", Engineering Applications of Artificial Intelligence, 2020.

© 2022 by the Spandan Srivastava, Charu Lata, Prateek Lohan, Rinchin W. Mosobi. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 454-459 | e-ISSN: 2347-470X

# Comparative Study of NCM and NCA Electrode Material for Capacity-Fade Using 1-D Modeling

**Ebani Gupta[1], Vikas Verma[2], Gaurav Khulway[3] and Amrish K. Panwar[4]**

*[1,2,3,4]Department of Applied Physics, Delhi Technological University, Delhi, India, [1]ebani.ecelldtu@gmail.com, [2]technovikas06@gmail.com, [3]gauravkhulway16@gmail.com, [4]amrish.phy@dtu.ac.in*

***Correspondence:** Ebani Gupta; ebani.ecelldtu@gmail.com

**ABSTRACT-** Today, Lithium-ion (Li-ion) batteries are one of the most emerging power sources for almost all modern consumer electronic products. $LiNi_{0.8}Co_{0.15}Al_{0.05}O_2$ (NCA) and $LiNi_{0.3}Co_{0.3}Mn_{0.3}O_2$ (NCM) are projected to be utilized in lithium-ion power batteries as two typical layered nickel-rich ternary cathode materials. Moreover, there is still a need for systematic study from an industrial aspect as to the advantages and drawbacks of these two nickel-rich materials. Hence, a comparative study of NCM and NCA electrode material for capacity-fade has been explored using a 1-D simulated model constructed in the multi-physics software. The capacity of a battery depends on the cell potential, discharge rate, state of charge (SoC), and state of health (SoH). Therefore, the comparison of these parameters and the cycle number of a battery is extremely important. During this comparative study of NCM and NCA electrode material, the capacity fade based on discharge rate, SoC, and SoH over cycle number of a battery has been reported.

**Keywords:** Lithium-ion battery, NCM, NCA, Capacity-fade, Battery Modelling.

ARTICLE INFORMATION

**Author(s)**: Ebani Gupta, Vikas Verma, Gaurav Khulway and Amrish K. Panwar.
**Received**: 31/03/2022; **Accepted**: 18/07/2022; **Published**:10/08/2022;
**e-ISSN**: 2347-470X;
**Paper Id**: IJEER100308;
**Citation:** 10.37391/IJEER.100308
**Webpage-link**:
https://ijeer.forexjournal.co.in/archive/volume-10/ijeer-100308.html

This article belongs to the Special Issue on **Recent Advancements in the Electrical & Electronics Engineering**

**Publisher's Note:** FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

## 1. INTRODUCTION

Exploring energy storage and conversion device resources is critical to meet the growing demand for energy while also being environmentally friendly [1]. In recent years, the focus of rechargeable batteries has shifted to energy storage and conversion devices all around the world. Lithium-ion batteries (LIBs) are becoming more favored among rechargeable batteries because of their higher energy level and power densities when compared to other batteries. In addition to portable devices, these batteries are utilized in smart grids, the automobile industry, aeronautical engineering, communication, backup, and a range of other industries [2]. Battery usage is expanding on a massive scale across the globe for a variety of reasons, and battery dependency has been steadily increasing. Usually, Li-ion batteries undergo various electrochemical and mechanical changes during their working condition and are thus thought to be complex [1].

Electrochemical Energy Storage (EES) devices have become essential in contemporary times because of their high energy density, reliability, excellent life cycle, stability, etc. [3]. Lithium is the lightest metal with the most significant electrochemical potential, allowing for tremendous energy and power density. In LIBs, lithium metal is a promising anode material due to its small weight, high energy density, highest theoretical capacity, and low reduction potential. But uncontrolled reactivity with oxygen, moisture and interfacial interactions with the electrolyte limit the use of lithium metal an anode in LIBs for practical applications [2].

The application of these batteries is limited due to the physical limitations of the electrode materials whether it is for cathode or anode material. The electrodes in Li-ion batteries are made of various materials. Laptops and cell phones most commonly use the combination of Lithium cobalt oxide and graphite.

The battery industry is very much interested in finding ways to increase the energy, power, safety, life, and cost of the cathodes used in Li-ion batteries for automotive-propulsion applications. A novel strategy to build better batteries for hybrid electric, plug-in hybrid electric, and battery electric vehicles is to blend cathode materials. A 'blended' cathode's active material is a substantial composition of two or additional different lithium intercalation compounds. Blending of various compounds is motivated by creating a more balanced performance than is feasible with any single ingredient.

The ionic and electronic conduction properties of LIBs are portrayed by both the cathode and the anode electrode materials. It's difficult to improve the material's ionic conduction because it's so closely linked to the ionic transport species, which reveals the material's composition and structure. Specific technological procedures are utilized to boost the material's intrinsic electronic conductivity, such as adding a conduction additive or a carbon coating [3].

Among various developed cathode materials for lithium-ion batteries, NCM has a high energy density and capacity, which

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 454-459 | e-ISSN: 2347-470X

makes it an attractive candidate for usage in electric vehicles. The positive electrode's metal combinations offer stability to the cell while also making the electrode of high performance and cost-effective. Unlike other cathodes, NCM has a longer cycle life, better thermal stability, and a higher energy density [4,5]. In contrast to olivine structures with two phases, its layered cathode structure shows a single-phase intercalation process [6]. $Li(Ni_xMn_yCo_z)O_2$ is the cathode of NCM, with total molar fractions (x, y, and z) equaling one. [7]. In terms of similarity, the discharge chemistry of NCM is similar to $LiFePO_4$ (LFP) [8]. NCM is better for its low-temperature resistance [9]. Therefore, it may be better ordinate for low-temperature applications. However, lithium ion- batteries based on NCA cathode have aluminum instead of manganese in NCM cathode-based batteries. When compared to the NCM equivalent the results show that it improves the specific energy and lifespan of the previous [10].

Since, both the NCA and NCM have the same layered cathode structure and indicate many similar properties. When compared to the capacities of other structural cathodes $LiMn_2O_4$ (LMO) and LFP batteries, which are 148 $mAhg^{-1}$ and 170 $mAhg^{-1}$, respectively, the gravimetric capacity of NCA batteries is high with the value of 200 $mAhg^{-1}$. In general, the lifespan of NCA batteries are known to be long and lasting. NCA batteries are not considered to be safe when compared to other batteries and this becomes one of the major disadvantages for them. For usage in applications such as electric cars, these batteries require specific safety monitoring [11] NCA batteries have a similar discharge method or chemistry to other Li-ion batteries. For Lithium, oxidation occurs at the anode, while reduction occurs at the cathode, with charge reactions occurring from right to left and discharge reactions occurring from left to right. NCA allows for high energy density batteries due to the high voltage. Another advantage of NCA is its outstanding capacity to fast charging.

Both NCM and NCA have excessive applications in contemporary times. Their usage has been increasing and will show an upward trend exponentially in upcoming years. Thus, in this article, it is intended to compare the two in order to find the best-suited characteristics of each to further conclude best-suited applicability. The Multiphysics software used for the simulation of the two battery models is good to explore the cell potential versus time, that help in comparing the characteristics of NCM and NCA battery models.

In this investigation, a 1-D simulation has been performed to optimize the characteristics parameters of the battery using a computer model in multi-physics software. A theoretical model is constructed using experimental data to explore the lithium-ion battery's electrochemical, electrical, and thermal properties. The capacity variance with respect to a number of cycles, electrolyte potential, ion concentration on the surface, and electrolyte salt concentration are explored. Hence, a comparative study of NCM and NCA electrode material for capacity-fade has been explored using a 1-D simulated model.

## 2. BATTERY MODEL FORMULATION

Positive electrodes (PE) for lithium-ion coin cells have been studied using lithium nickel cobalt manganese oxide ($LiNi_{0.3}Co_{0.3}Mn_{0.3}O_2$) (NCM) and lithium nickel cobalt aluminum oxide ($LiNi_{0.8}Co_{0.15}Al_{0.05}O_2$) (NCA). One-dimensional (1D) model geometry has been conducted for simulation analysis. The main focus of this simulation in both cases is on the coin cell's capacity retention, cell potential during various load cycles, and cell discharge pattern. These three components are monitored and serve as the model's control unit. This simulation model also depicts the charging and discharging curves and explains the cell's state of charge (SoC). Multiphysics software was used to create the lithium-ion coin cell model. A current collector (CC) on both sides, a separator, a positive electrode, and a negative electrode make up this model's simulation battery (NE) [12].

### 2.1 Model Parameters of NCM and NCA Electrodes

Initially, a 1D model of the battery here a coin cell, geometry was used and various frameworks were set based on the conditions of the battery. *Table 1* shows the several material criterion for studying the capacity of both NCA and NCM batteries, while *Table 2* shows the same for NCM batteries. Conditions for the boundary were arranged for the battery, one of them is thermal insulation in the leftmost boundary of the battery as well as in the center of the battery. As a result, the heat flux condition depicts the battery's heating surface, which is located on the rightmost boundary.

The salt concentrations, diffusion coefficient, and minimum and maximum state-of-charge of each electrode are entered as input parameters in the simulation software to study capacity retention for both materials.

1D battery model parameters to study capacity-related properties

**Table 1: Material parameters for studying the capacity of NCA based batteries used in 1D battery simulation [13]**

| Property | Variable | Value | Unit |
|---|---|---|---|
| Diffusion coefficient | D_iso ; Dii = D_iso, Dij = 0 | 1.5e-15[$m^2/s$] | $m^2/s$ |
| Electrical conductivity | sigma_iso ; sigmaii = sigma_iso, sigmaij = 0 | 91[S/m] | S/m |
| Reference concentration | cEeqref | 48000[$mol/m^3$] | $mol/m^3$ |
| Maximum electrode state-of-charge | Socmax | 1 | 1 |
| Minimum electrode state-of-charge | Socmin | 0.25 | 1 |
| Transport number | transpNum | 1 | 1 |

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 454-459 | e-ISSN: 2347-470X

| | | | |
|---|---|---|---|
| Activity dependence | Fcl | 1 | 1 |
| Electrolyte conductivity | sigmal_iso ; sigmalii = sigmal_iso, sigmalij = 0 | 91[S/m] | S/m |
| Density | Rho | 4740[kg/m$^3$] | kg/m³ |

**Table 2. Material parameters for studying the capacity of NCM based batteries used in 1D battery simulation**

| Property | Variable | Value | Unit |
|---|---|---|---|
| Diffusion coefficient | D_iso ; Dii = D_iso, Dij = 0 | 5e-13[m$^2$/s] | m²/s |
| Electrical conductivity | sigma_iso ; sigmaii = sigma_iso, sigmaij = 0 | 100 [S/m] | S/m |
| Equilibrium potential | Eeq | Eeq_int1(SoC)+d EeqdT*(T-298[K]) | V |
| Temperature derivative of equilibrium potential | dEeqdT | -10[J/mol/K]/F_const | V/K |
| Reference concentration | cEeqref | 49000[mol/m$^3$] | mol/m³ |
| Max electrode SoC | SoHmax | 0.975 | 1 |
| Min electrode SoC | SoCmin | 0 | 1 |
| Electrolyte conductivity | sigmal_iso ; sigmalii = sigmal_iso, sigmalij = 0 | 9.65[S/m] | S/m |
| Transport number | transpNum | 1 | 1 |
| Activity dependence | Fcl | 1 | 1 |

## 2.2 Model Parameters used for Negative Electrode and Electrolyte

Graphite is used as the anode material in the proposed simulated Lithium-ion cells in 1-D. Therefore, material characteristics parameters taken as input parameters for the NE (Graphite) are noted in *Table 3.*

**Table 3: Material parameters used for Negative (graphite) electrode1-D battery simulation [12, 13]**

| Property | Variable | Value | Unit |
|---|---|---|---|
| Electrical conductivity | sigma_iso ; sigmaii = sigma_iso, sigmaij = 0 | 100[S/m] | S/m |
| Diffusion coefficient | D_iso ; Dii = D_iso, Dij = 0 | 1.4523e-13*exp (68025.7/8.314*(1/(T_ref/1[K])-1/(T2/1[K]))) [m$^2$/s] | m²/s |
| Reference concentration | cEeqref | 31507[mol/m$^3$] | mol/m³ |
| Max electrode SoC | SoCmax | 0.98 | 1 |
| Min electrode SoC | SoCmin | 0 | 1 |
| Electrolyte conductivity | sigmal_iso ; sigmalii = sigmal_iso, sigmalij = 0 | 468[S/m] | S/m |
| Transport number | transpNum | 1 | 1 |
| Activity dependence | Fcl | 1 | 1 |
| Thermal conductivity | k_iso ; kii = k_iso, kij = 0 | 1[W/(m*K)] | W/(m·K) |
| Heat capacity at constant pressure | Cp | 750[J/(kg*K)] | J/(kg·K) |
| Density | Rho | 2300[kg/m$^3$] | kg/m³ |

The most common electrolyte used by manufacturers and researchers in Li-ion batteries to manufacture or research is $LiPF_6$. The electrochemical properties of cathode, and anode have been evaluated using the LIPF-EC-EMC electrolyte and works for an optimum composition for Li-ion batteries [14]. The details of electrolyte used in this study is given in *Table 4.*

**Table 4: Input parameter used for electrolyte (LiPF6) [15]**

| Property | Variable | Value | Unit |
|---|---|---|---|
| Electrolyte salt concentration | cElsalt | 1200[mol/m$^3$] | mol/m³ |

# 3. RESULTS AND DISCUSSION

## 3.1 Related Capacity versus Cycle Number

The accurate estimation of the remaining battery life is critical for the system's reliable operation. As a result, SoH is introduced as a parameter. SoH is defined as a figure of standard for a battery's condition in comparison to its beginning conditions. SoH is measured in percent points, with 100 percent indicating the beginning state of a battery [16]. Cycle performance is used to determine the usable capacity of a battery. As the number of cycles increases, the usable capacity of the battery decreases [17].

In *Figure 1* for NCM, it can be noticed that for each cycle number the battery's usable capacity is changes. Based on cyclability, it can be observed that with an increase in cycle number the SoH decreases. At 400 cycles, cyclable lithium decreases from 1 to 0.98. It is observed that till 500 cycles, SoH drops more drastically than it decreases from 500 to 1800 cycles. For the next 400 cycles (400 to 800 cycles), It decreases mildly in comparison to the previous 400 cycles (0 to 400 cycles), and relative capacity decreases to 0.94. At 1800 cycles, it is decreased upto 0.91.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 454-459 | e-ISSN: 2347-470X

A C-rate is a measurement of how quickly a battery is depleted compared to its maximum capacity [18]. The discharge current at a 1C rate will totally deplete the battery in one hour. For the Nominal 1C rate the discharge capacity, it decreases throughout the cycles. It follows the same pattern as cyclable lithium. At 400 cycles, the Nominal 1C rate the discharge capacity decreases from 0.78 to 0.76. At 500 cycles, it decreases linearly with the cycle number. At 1800 cycles, it will decrease up to 0.67.

*Figure 2* shows the change in usable capacity of a battery with number of cycles for NCA materials used as anode. Based on cyclable lithium, it can be observed that with increase in cycle number, SoH decreases. At 500 cycles, Cyclable lithium decreases from 1 to 0.95. It is observed that till 500 cycles, SoH drops more drastically than it decreases from 500 to 1800 cycles. For next 500 cycles (500 to 1000 cycles), it decreases mildly in comparison to previous 500 cycles (0 to 500 cycles) and relative capacity decreases to 0.92. At 1800 cycles, it will decrease up to 0.87.For Nominal 1C rate the discharge capacity, it decreases throughout the cycles. It follows the same pattern as cyclable lithium. Till 500 cycles, Nominal 1C discharge rate the capacity decreases from 0.88 to 0.84. At 500 cycles, it decreases linearly with the cycle number. At 1800 cycles, it will decrease up to 0.76.

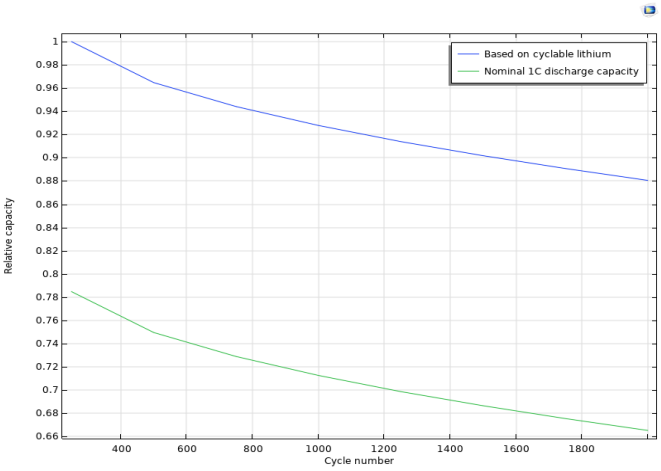



**Figure 1:** Variation of relative capacity with cycle number for NCM anode for SoH

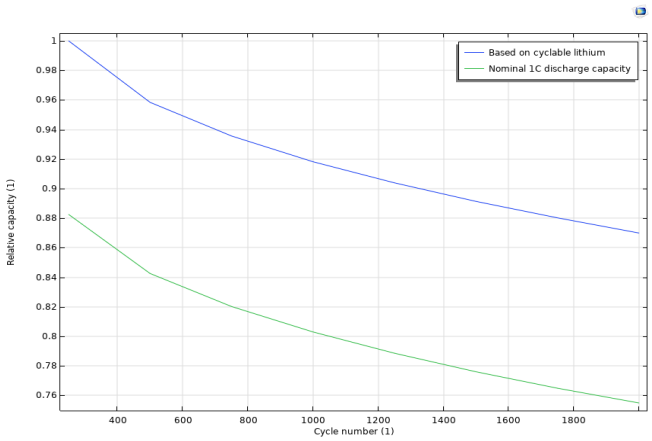



**Figure 2:** Variation of relative capacity with cycle number for NCA anode for SoH

## 3.2 Call Potential versus Time

To see the cyclic performance of the battery made of 1-D cathodes. The battery is run through 2000 cycles and the discharge curve is observed for both the NCM battery and the NCA battery model. The data of first and the last cycle is mentioned here. A linear drop has been observed for the first 500 seconds for both the cycles and the materials (NCM and NCA) [19]. In *Figure 3* for NCA, the decline follows a similar pattern till 2500 seconds after which a sudden drop in the cell potential is observed. For the first cycle, the steep decline is observed between 3.1V and 3.2V, whereas for the 2000th cycle, it is observed at 3.3V. Around 2700 seconds, the cell potential drops to a very low value for the 2000th cycle. In case of the 1st cycle, the same is observed post 3000s. A linear drop has been observed for the first 500 seconds for both the cycles as shown in *Figure 4* for NCM. The decline follows a similar pattern till 2000 seconds of the discharge (1C rate) after which a sudden drop in the cell potential is observed. For the first cycle, the steep decline is observed between 3.4V and 3.5V, whereas for the 2000th cycle, it is observed at 3.4V. Around 2200 seconds, the cell potential drops to a very low value for the 2000th cycle. In case of the 1st cycle, the same is observed post 2500s.

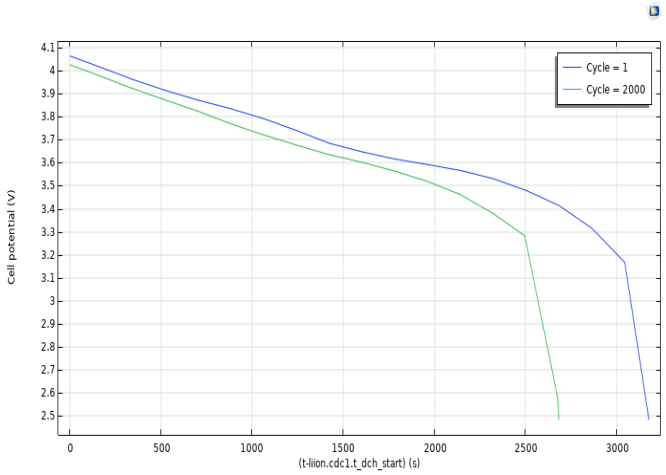



**Figure 3:** Discharge potential vs time for cycles 1st and 2000th for NCA

## 3.3 State of Charge versus Time

The SoC plays a crucial part in estimating the battery pack's remaining capacity. Because of the constrained battery model, accurate estimation of the SoC is quite complex and difficult to accomplish [20]. The battery's SoH is a crucial measure of its lifespan.

In *Figure 5* for NCM, the battery is run through 2000 cycles and the SoC curve is observed for both the NCM battery and the NCA battery model. The first and the last cycle is observed for the purpose of the study. The positive electrode in the battery for the first cycle starts from 0.73 whereas, for the last cycle, it starts from 0.63. At 3000 seconds, PE attains the lowest SoC for the first cycle as well as for the last cycle. After that, Positive electrode SoC at the first cycle remains closer to the last cycle with a minute difference.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 454-459 | e-ISSN: 2347-470X

For NE, SoC for the first cycle and last cycle is similar till the first 1000 seconds. After that, it can be observed that there is a minute difference between the first cycle and the last cycle. At 3000 seconds, both attain their highest SoC, for first cycle, it is 0.75 but for last cycle, it is closer to 0.6. Then there is a significant difference between first cycle SoC and last cycle SoC.

For NCA *Figure 6* shows that the battery is run through 2000 cycles and the SoC curve is observed for both the NCM battery and the NCA battery model. The positive electrode in the battery for the first cycle starts from 1 whereas for the last cycle, it starts from 0.88. At 3200 seconds, PE attain the lowest SoC for the first cycle as well as for the last cycle. After that, Positive electrode SoC at the first cycle remains closer to the last cycle with a minute difference.

For NE, SoC for the first cycle and last cycle is similar till the first 1000 seconds. Then it can be observed that there is a minute difference between the first cycle and the last cycle. At 3200 seconds, both attain their highest SoC, for first cycle it is 0.82 but for last cycle it is closer to 0.7. After that there is a significant difference between first cycle SoC and last cycle SoC.

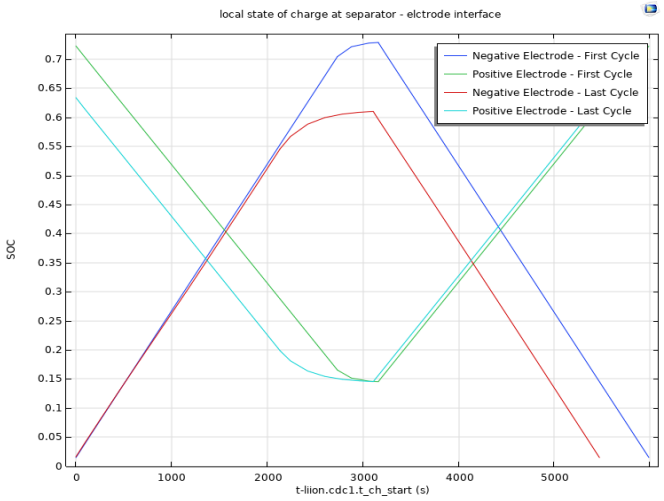



**Figure 5:** State of Charge vs Time for 5000s for NCM

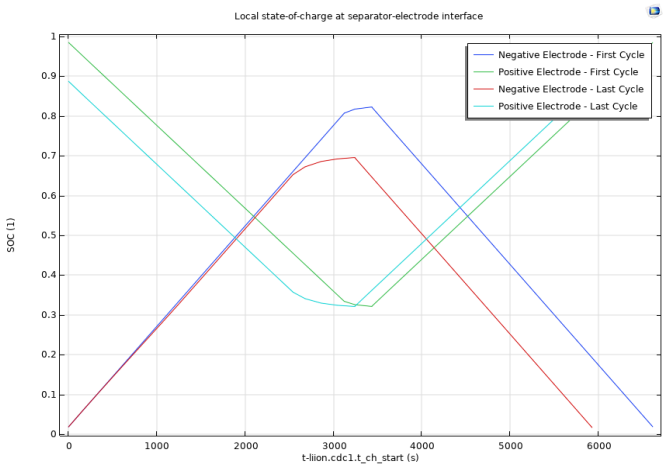



**Figure 6:** State of Charge vs Time for 5000s for NCA

## 4. CONCLUSION

This study reports the comparison of two typical layered nickel-rich ternary cathode materials (NCM and NCA electrode material) based on capacity fade of the battery. For capacity fade, the comparison is made in both the materials in terms of cell potential, discharge rate, SoC and SoH which affects the capacity of the battery. In the Relative capacity-cycle number curve, It has been observed that cyclable lithium capacity decreases by 1.23% (approx.) in NCA in comparison with NCM after completing 1800 cycles. While at Nominal 1C rate, discharge capacity decreases by 11.42% in NCM in comparison with NCA after completing 1800 cycles. In the SoC-time curve, the highest SoC for the NE (last cycle) reached 0.7 for NCA whereas between just crosses 0.6 for NCM, it indicates that higher charge retention in post multiple cycles for NCA is observed. Similarly, Cell potential shows a significant drop in the case of NCM for 2000 cycles as compared to NCA. Thus, the capacity fade models for the two positive electrode materials have been studied and summarized in a comparative manner.

## ACKNOWLEDGEMENTS

The authors are thankful to the Department of Applied Physics, Delhi Technological University, New Delhi for providing support to carry out this research work.

## REFERENCES

[1] Snigdha Sharma; Amrish K Panwar; Madan Mohan Tripathi. Investigation of Electrochemical, Thermal and Electrical Performance of 3D Lithium-Ion Battery Module in a High - Temperature Environment. Int. Journal of Renewable Energy Development 9 (2) 2020: 151-157

[2] Naoki Nitta; Feixiang Wu; Jung Tae Lee; Gleb Yushin1. Li-ion battery materials: present and future. Materials Today; Volume 18; Number 5; June 2015.

[3] Snigdha Sharma; Amrish K Panwar; Madan Mohan Tripathi. Thermal-Electrochemical Modeling and Analysis of Different Cathode-Anode Pairs for Lithium-ion Battery. Advances in Electrical and Computer Engineering. Volume 21, Number 3, 2021

[4] Capron, O.; Gopalakrishnan, R.; Jaguemont, J.; Van Den Bossche, P.; Omar, N.; Van Mierlo, J. On the Ageing of High Energy Lithium-Ion Batteries—Comprehensive Electrochemical Diffusivity Studies of Harvested Nickel Manganese Cobalt Electrodes. Materials 2018, 11, 176.

[5] Li, Z.; Chernova, N.A.; Roppolo, M.; Upreti, S.; Petersburg, C.; Alamgir, F.M.; Whittingham, M.S. Comparative Study of the Capacity and Rate Capability of LiNi y Mn y Co1–2y O2 (y = 0.5, 0.45, 0.4, 0.33). J. Electrochem. Soc. 2011, 158, A516.

[6] Forte, F.; Pietrantonio, M.; Pucciarmati, S.; Puzone, M.; Fontana, D. Lithium Iron Phosphate Batteries Recycling: An Assessment of Current Status. Crit. Rev. Environ. Sci. Technol. 2020, 1–28.

[7] Yang, H.; Savory, C.N.; Morgan, B.J.; Scanlon, D.O.; Skelton, J.M.; Walsh, A. Chemical Trends in the Lattice Thermal Conductivity of Li(Ni, Mn, Co)O2 (NCM) Battery Cathodes. Chem. Mater. 2020, 32, 7542–7550.

[8] He, H.; Xiong, R.; Guo, H.; Li, S. Comparison Study on the Battery Models Used for the Energy Management of Batteries in Electric Vehicles. Energy Convers. Manag. 2012, 64, 113–121

[9] Gurpreet Singh; Anjan Sil; Sudipto Ghosh ; Amrish Panwar. Effect of citric acid content on synthesis of LiNi1/3Mn1/3Co1/3O2 and its electrochemical characteristics. Ceramics International 36 (2010) 1831–1836.

[10] Li, K.; Zhang, J.; Chen, M.; Yang, Z.; Niu, Q. Advances in Green Energy Systems and Smart Grid; Springer: Singapore, 2018; ISBN 9789811323812.

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 454-459 | e-ISSN: 2347-470X

[11] Miao, Y.; Hynan, P.; von Jouanne, A.; Yokochi, A. Current Li-Ion Battery Technologies in Electric Vehicles and Opportunities for Advancements. Energies 2019, 12, 1074.

[12] Snigdha Sharma; Amrish K Panwar; Madan Mohan Tripathi. Enhanced Electrical and Mechanical Characteristics Monitored by Various Software for Lithium-Ion Battery Tests. Lecture Notes in Electrical Engineering , LNEE, volume 664.

[13] Simon Malifarge; Bruno Delobel; Charles Delacourt. Experimental and Modeling Analysis of Graphite Electrodes with Various Thicknesses and Porosities for High-Energy-Density Li-Ion Batteries. Journal of The Electrochemical Society, Volume 165, Number 7

[14] Willem Van Zyl. Analysis of Graphite for Lithium Ion Batteries. 07.07.2021

[15] S.S. Zhagn; T.R.Jow; K.Amine; G.L. Henriksen. LiPF6-EC-EMC electrolyte for Li-ion Battery. ELSAVIER, Journal of Power Sources 107 (2002) 18 - 23

[16] Türev Sarıkurt, Murat Ceylan; Abdulkadir Balıkçı. An Analytical Battery State of Health Estimation Method. Industrial Electronics (ISIE), 2014 IEEE 23rd International Symposium

[17] Du, Zhijia ORCiD logo; Wood, David L.; Daniel, Claus; Kalnaus, Sergiy; Li, Jianlin. Understanding limiting factors in thick electrode performance as applied to high energy density Li-ion batteries.Springer Science+Business Media Dordrecht 2017. J Appl Electrochem (2017) 47:405–415.

[18] Denny Schmidta; Marc Kamlahb; Volker Knoblauchc. Highly densified NCM-cathodes for high energy Li-ion batteries: Microstructural evolution during densification and its influence on the performance of the electrodes. Journal of Energy Storage 17 (2018) 213-233

[19] Ruhul Amin; Dorthe B. Ravnsbaek; Yet Ming Chian. Characterization of transport properties of LiNi0.8Co0.15Al0.05O2 (NCA). Abstract #260; 224th ECS Meeting; 2013 The Electrochemical Society.

[20] Duygu Karabelli; Kai Peter Birke; Max Weeber. A Performance and Cost Overview of Selected Solid-State Electrolytes: Race between Polymer Electrolytes and Inorganic Sulfide Electrolytes. MDPI Review; 5 March 2021.

© 2022 by Ebani Gupta, Vikas Verma, Gaurav Khulway and Amrish K. Panwar. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

*2022 2nd International Conference on Intelligent Technologies (CONIT)*
*Karnataka, India. June 24-26, 2022*

# Comparison of Different Coil Structures for Wireless Power Transfer Using Magnetic Resonant Coupling

Pranjal Bajpai
*Electrical Engineering Department*
*Delhi Technological University*
Delhi, India
pranjalbajpai_2k18ee139@dtu.ac.in

Pranav Saihgal
*Electrical Engineering Department*
*Delhi Technological University*
Delhi, India
pranavsaihgal_2k18ee138@dtu.ac.in

Pulkit Kapoor
*Electrical Engineering Department*
*Delhi Technological University*
Delhi, India
pulkitkapoor_2k18ee148@dtu.ac.in

Saurabh Mishra, *Member IEEE*
*Electrical Engineering Department*
Delhi Technological University
Delhi, India
saurabhmishra@dtu.ac.in

2022 International Conference on Intelligent Technologies (CONIT) | 978-1-6654-8407-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CONIT55038.2022.9848039

***Abstract*— Magnetic resonant coupling is preferred over magnetic induction because of its higher efficiency and ability to transfer power for longer distances. The main components of wireless power transfer system are a primary transmission coil connected to power source and a secondary receiving coil. Both coils are separated by some air gap. Primary coil generates an alternating magnetic field and flux, which triggers current to flow through a nearby receiver coil. This paper describes the power transfer between coils, coil position, size of coils, number of coil rotations, intermediate coils. This paper also deals with the variation of mutual inductance and coupling coefficient over a wide range of distance for each of the three coil structures namely - square spiral, circular spiral and helix. The analysis is performed with the assistance of ANSYS Electronics software.**

***Keywords—Wireless Power Transfer (WPT), Magnetic Resonant Coupling (MRC), Mutual Inductance, Coupling Coefficient, Coil separation.***

## I. INTRODUCTION

New technologies are emerging each and every day to make our lives simpler. The traditional wire system creates a great clutter when charging several devices simultaneously. This is where WPT (wireless power transfer) comes into play, a concept which provides an effective and simpler way to transfer power wirelessly. Inductive wireless power technology has gained attention from the growing number of startups in sectors like (Fulton Innovation, wise power, WiTricity, and others) and the consumer market titans like Intel and Sony which demonstrate that it has acquired substantial momentum for day-to-day usage for its prospective customers [1].

WPT is an effective method to transmit electric power between two points without the usage of wires in space through a vacuum or environment. Presently, the WPT technology is most commonly used in low power consuming devices to medium power-consuming devices such as mobile phones [2], medical implants and EVs to some extent [3]. It is used only for short air gap purposes since a fractional misalignment of coils leads to a substantial reduction in efficiency of the system. A technique of WPT technology, named magnetic resonance coupling or MRC, depends on a high frequency alternating magnetic field for transmission of power across the air gap between the two coils which are operating at the resonant frequency, leading it to be more efficient than conventional inductive coupling. It has been observed that MRC can achieve efficiency up to 90% within the range of one meter [4].

One application that has proven to be impactful is the use of wireless in-wheel motor (IWM) in EVs for better comfort and performance as this technology helps IWM receive its power wirelessly from the vehicle chassis [5]. For low power devices such as the charger of a mobile phone, the energy is transferred from the charging pad to the battery after making sure it is in place for charging to take place [6]. Another such application is popular in this era is the advent of electric vehicles replacing the traditional vehicles with internal combustion engines also leading to the rise in the need for well-connected charging infrastructure. Some benefits of using near-field WPT technology instead of plug-in EV charging includes the fact that it is more convenient to charge because the power ranges from 1 W to 3 kW, and the frequency ranges from 42 kHz to 148.5 kHz, and it is safer to use in repairing electric vehicles. It also prevents maintenance workers as well as the public from the possibility of potential danger or injury. In the near future, the inconvenience to stop the vehicle for charging will not be present. Instead, a wireless power transmitter will be linked to highways and congested regions, which will fetch power from grid and transmit that electricity as wireless power, to the Electric Vehicle moving on that highway/road. The EV itself consists of a wireless power receiver that uses this wireless electrical energy for the charging of the onboard battery inside the vehicle**.** At present, 97 percent of Indian vehicles are propelled by petrol and diesel which have an adverse impact on the environment. After a study conducted by Govt. of India NITI AYOG has mentioned in the corresponding report, that they aimed to reduce the emission levels of India by 33-35 percent in comparison to year 2005 level by the year 2030. To achieve this target a total of 1827 electric charging stations are being developed across the country by the government which will give a boost to the clean energy ecosystem.

The presented work is arranged in six sections, IInd section presents the system configuration. IIIrd section includes system analysis using MRC. Section IV covers design of coil various coil structures. Detailed results and analysis are discussed in section V followed by conclusion.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:51:19 UTC from IEEE Xplore. Restrictions apply.

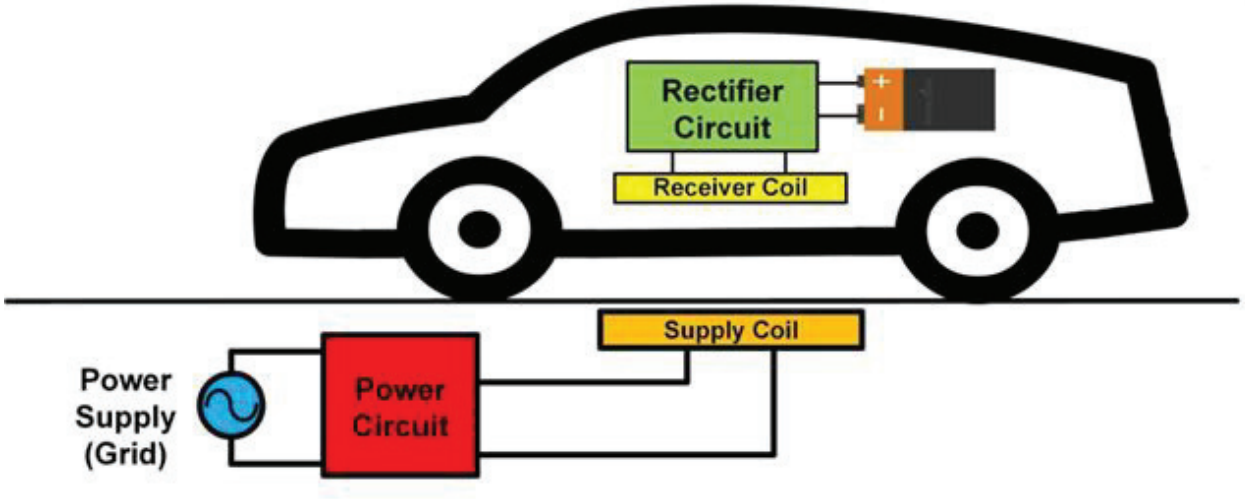



Fig. 1. Block Diagram of WPT system for Electric Vehicle

## II. SYSTEM CONFIGURATION

Wireless power transfer is the transmission of power from a source to a load without any need for wires or any other physical link. On the basis of the separation between source and the load, WPT techniques are classified into two types: one is near field and other is far field power transfer. In near field power transfer, a significant amount of power can be transmitted between source and load but only for small distances (a few hundred millimeters). In far-field power transfer, only a small amount of power can be transmitted using Electromagnetic waves, i.e. use of high-frequency waves in radio and satellite communication.

Near field power transfer occurs using phenomena like capacitive coupling between metal electrodes and inductive power transfer between metal coils. Wireless power transfer between two coils via resonant coupling can be explained by the principle of Electromagnetic Induction which states that "Change in current passing through a coil produces EMF in another coil placed near it". It has potential for application in various domains. One such application is the charging of electric vehicles [7]. The schematic diagram for wirelessly charging electric vehicles is illustrated in Fig 1. An AC-DC converter converts AC electricity from the grid into a DC supply. The DC supply is then converted into an AC supply of required frequency by feeding it to an inverter. The output coming from the inverter has high-frequency AC, which is fed into the transmission circuit, that contains majorly a resistor, capacitor and transmission coil. On the receiving side, EMF is induced in the load coil by the alternating magnetic field. Alternating EMF from the load coil is rectified in DC by passing it through a rectifier. Then the DC output of the rectifier is conditioned for storing in the battery.

## III. SYSTEM ANALYSIS USING MRC

Magnetic resonant coupling is a more efficient version of Inductive Power Transfer (IPT) [8]. During IPT, when the primary side containing voltage source $V_S$ (as shown in Fig. 2) is operated at a particular frequency such that the magnetic field of primary coil starts resonating with the magnetic field of secondary coil, then magnetic flux linkage between the two coils induces maximum EMF in the secondary coil. Thus, the two coils are called to be magnetically coupled. This phenomenon is called magnetic resonant coupling.

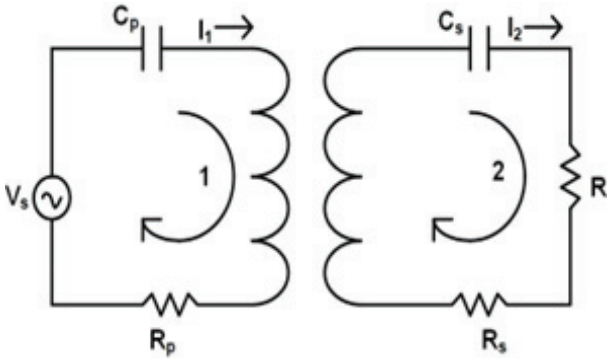



Fig. 2. Power Transfer Circuit

From loop 1,

$$V_s = I_1\left\{R_p + jL_p\omega + \frac{1}{j\omega C_p}\right\} - I_2\left(jL_m\omega\right) \quad (1)$$

From loop 2,

$$0 = I_2\left\{jL_s\omega + \frac{1}{j\omega C_s} + z_0 + R_s\right\} - I_1\left(jL_m\omega\right) \quad (2)$$

$$C_p = C_s = C$$

$$L_p = L_s = L$$

$$R_p = R_s = R$$

Equating Equations and 1 and 2

$$V_s = I_1\left\{R_p + jL_p\omega + \frac{1}{j\omega C_p}\right\} - I_1\left\{\frac{\left(jL_m\omega\right)^2}{jL_s\omega + \frac{1}{j\omega C_s} + z_0 + R_s}\right\} \quad (3)$$



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:51:19 UTC from IEEE Xplore. Restrictions apply.

$$Z_i = \left\{ R + jL\omega + \frac{1}{j\omega C} \right\} + \left\{ \frac{L_m^2 \omega^2}{jL\omega + \frac{1}{j\omega C} + z_0 + R} \right\} \quad (4)$$

$$Z_i = R + j(L - L_m)\omega + \left\{ \frac{A}{jL\omega + \frac{1}{j\omega C} + z_0 + R} \right\} \quad (5)$$

Where,

$$A = -jL_m^2\omega^2 + L_m L\omega^2 + jL_m\omega(z_0 + R) + jL_m\left(\frac{1}{j\omega C}\right)\omega$$

$$\eta = \frac{P_0}{P_i} = \frac{I_2^2 z_0}{I_1^2 z_i} \quad (6)$$

From above equations,

$$\eta = \left\{ \frac{jL_m\omega}{jL\omega + \frac{1}{j\omega C} + z_0 + R} \right\}^2 \left\{ \frac{z_0}{B} \right\} \quad (7)$$

Where,

$$B = R + jL\omega + \frac{1}{j\omega C} + \frac{L_m^2}{jL\omega + \frac{1}{j\omega C} + z_0 + R}$$

Condition for optimal power transfer,

$$L_m^2 = \frac{z_0^2 - R^2}{\omega^2}$$

The charging of EVs with MRC WPT technique is feasible for those vehicles having ratings lying in the range from 1 kW to up to 3 kW, and the resonance frequency can vary between 42 kHz to 148.5 kHz or even higher [9].

### *A. Magnetic Coupling between Coils*

#### *1) Mutual Inductance*

The flux linkage of one coil is changed due to the current flowing through the other coil.

$$\phi_{21} = MI_1$$

#### *2) Coupling Coefficient*

The ratio of flux passing through the secondary coil to flux passing through the primary coil is called coupling coefficient.

$$k = \frac{M}{\sqrt{L_1 L_2}}$$

'k' can vary from 0 to 1 with being tightly coupled and 0 being not coupled at all.

### *B. Factors Affecting Magnetic Coupling between coils*

Aside from frequency, the shape of the coils, the medium between them, the orientation of the coils with regard to each other, the material of the coils, and the distance between coils are all factors that influence the coupling between two coils.

To analyze the effect of coil shapes on magnetic coupling between the coils the three most widely used coil shapes are chosen namely - Circular coil, Helical coil and Square coil.

## IV. DEISIGN OF DIFFERENT COIL STRUCTURES

Coil geometry is one of the most important criteria of MRC WPT design. Different uses of WPT systems use various configurations of a coil with different parameters. The three most widely used coil structures are Circular, Helical and Square.

Since there are many factors that affect the magnetic coupling between the coils, to focus on the effect of coil shapes only, the following parameters have been kept the same for each of the three coil shapes.

TABLE I. COMMON DESIGN PARAMETER

| S No. | Parameter | Value |
|---|---|---|
| 1 | No. of Turns | 4 |
| 2 | Area of Cross-section | 1 cm x 1 cm |
| 3 | Distance between consecutive loops | 1 cm |
| 4 | Diameter | 20 cm |
| 5 | Excitation Current | 5 A |
| 6 | Material | Copper |
| 7 | Medium between two coils | Vacuum |

The two main characteristics of magnetic coupling are coupling coefficient and mutual inductance, so variation of these two over a range of coil separation will be analyzed.

The most anticipated application of WPT is in the domain of wireless charging of electric vehicles and the distance of ground to the floor of a car ranges from 150-200 mm. Thus, the range of coil separation is varied from 100 mm to 200 mm. All the simulations are performed in ANSYS Maxwell software.

### *A. Circular Coil Structure*

The variation of coupling coefficient and mutual Inductance of coils is shown in the below Fig. 3 when the distance between them changes. For this purpose, a 3D Model of the circular coil is designed on Ansys Maxwell software. In four steps, the coil separation is increased from 100 to 200 mm. At each step, coils are moved 25 mm apart. The power transmission takes place from the primary coil to the secondary coil wirelessly through an alternating magnetic field, in other words, the magnetic field is delivering power from one coil to another instead of wires. Thus, visualization of the magnetic field is crucial to analyze the extent of magnetic coupling taking place between coils. Magnetic flux is the number of magnetic field lines travelling through a given cross-section area. The magnetic flux that originates out from source coil and goes through the load coil is referred to as the flux interaction between these two coils.

#### *1) Flux Linkage:*

The magnetic flux passing between the coils is shown in Fig. 3. In this Fig. 3, it can be seen that a very strong magnetic field is present along the axis passing through the center of the coil. The magnetic field decreases radially outwards from the center of coils.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:51:19 UTC from IEEE Xplore. Restrictions apply.

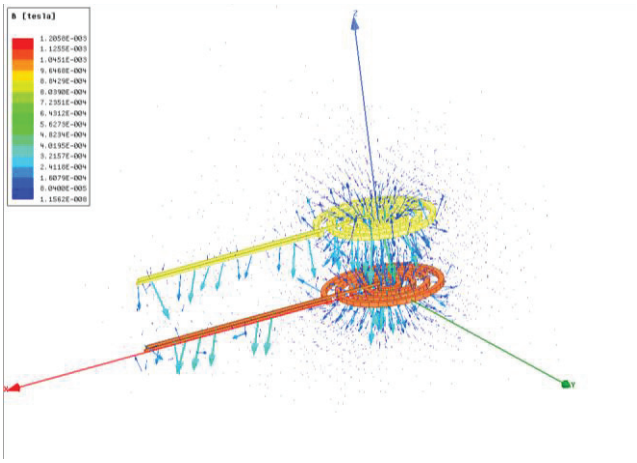


Fig. 3. Magnetic field present between circular coils.

### B. Helical Coil Structure

The mutual inductance and coupling coefficient of the coils fluctuate as the distance between them changes. Ansys Maxwell is used to create a 3D model of the square coil for this purpose. The coil separation is extended from 100 to 200 mm in four increments, with each step increasing by 25 mm. At a distance of 100 mm, the flux linkage between the coils can clearly be seen. An alternating magnetic field wirelessly transfers electricity from the primary coil to the secondary coil, or in other words, the magnetic field delivers power from one coil to another instead of wires. Due to which, visualizing the magnetic field is critical for determining the extent of magnetic coupling between coils. Controlling the number of turns and diameter of each coil, thus changing the setup, can produce adequate magnetic field. Furthermore, owing to its simple architecture, this structure is indifferent to the effects of the ground plane and is intuitive to manufacture.

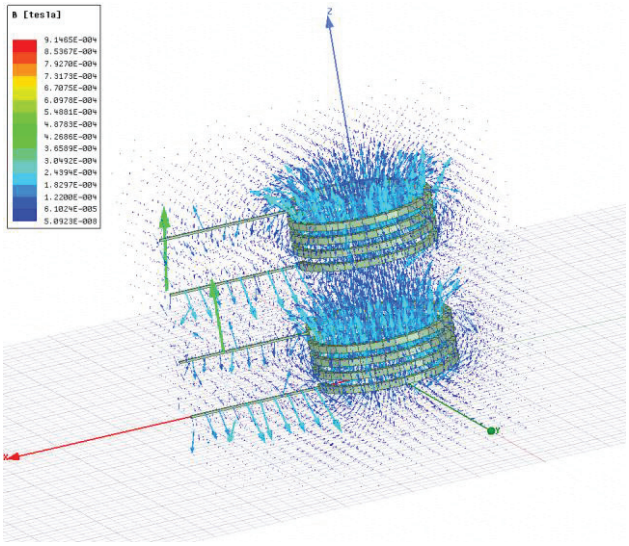


Fig. 4. Helical Coil Structure

#### 1) Flux Linkage:

The magnetic field between the coils is shown in Fig. 4. In this Fig. 4, it can be seen that a very strong magnetic field is present along the axis passing through the center of the coil. The magnetic field decreases radially outwards from the center of coils.

### C. Square Coil Structure

The mutual inductance and coupling coefficient of the coils changes as the distance between them changes. For Simulation purpose, a 3D model of the square coil is constructed in Ansys Maxwell software. The coil separation is increased in four successive steps from 100 to 200 mm, with each step rising by 25 mm. The flux coupling between the coils can be clearly observed at a distance of 100 mm in Fig 5. An alternating magnetic field wirelessly transmits energy between the primary coil and secondary coil, or in other words, the magnetic field instead of wires is being used to transport power from one coil to another. As a result, to determine the extent of magnetic coupling between coils, visualization of the magnetic field becomes important.

#### 1) Flux Linkage:

Magnetic flux is the number of magnetic field lines travelling through a given cross-section area. This flux interaction between these two coils refers to the magnetic flux that originates from the source coil and passes through the load coil. The magnetic field between the coils is shown in Fig. 5. The magnetic field decreases radially outwards from the center of coils.

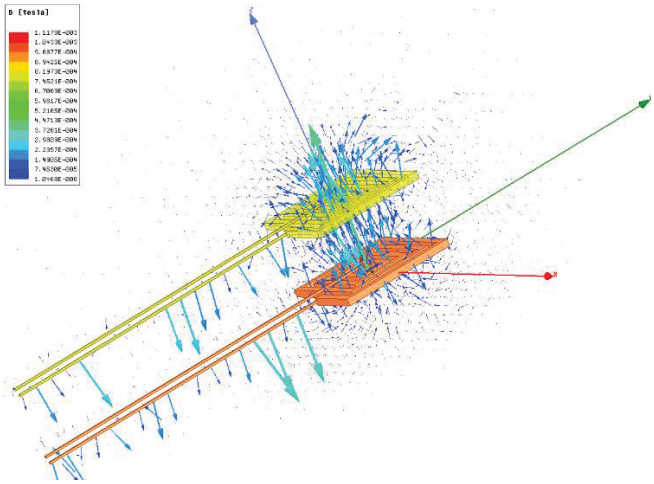


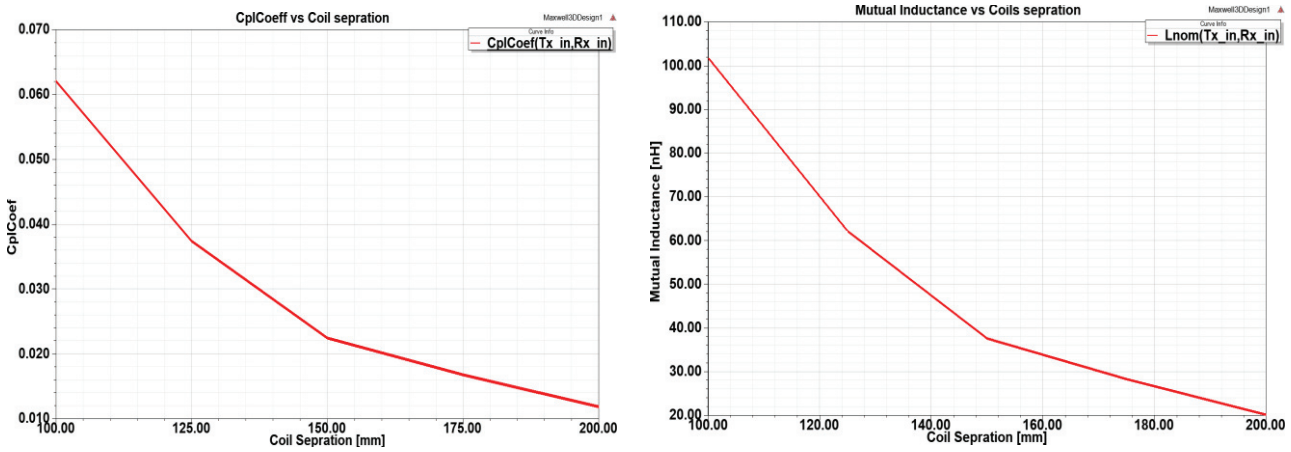
Fig. 5. Square Coil Structure

## V. RESULTS AND DISCUSSION

The amount of power transfer taking place from primary coil to secondary coil depends on the extent of magnetic coupling present between the coils. As discussed above the magnetic coupling present between coils is determined through two parameters namely – Mutual Inductance and Coupling Coefficient. Mutual Inductance is the ratio of flux passing through one coil due to the current flowing in the other coil. Coupling Coefficient is the ratio of flux passing through secondary coil to primary coil. Since, the most prominent and anticipated application of MRC WPT technology is in the field of wireless charging of static and dynamic electric vehicles, thus variation of coupling coefficient and mutual inductance between different coil pairs is analyzed at various separation intervals starting from 100 mm to up to 200 mm. Fig 6, Fig 7 and Fig 8 show the variation of Mutual Inductance and Coupling Coefficient for Circular, Square and Helix coil structure respectively. All models are created and analyzed on Ansys Maxwell 3-D.

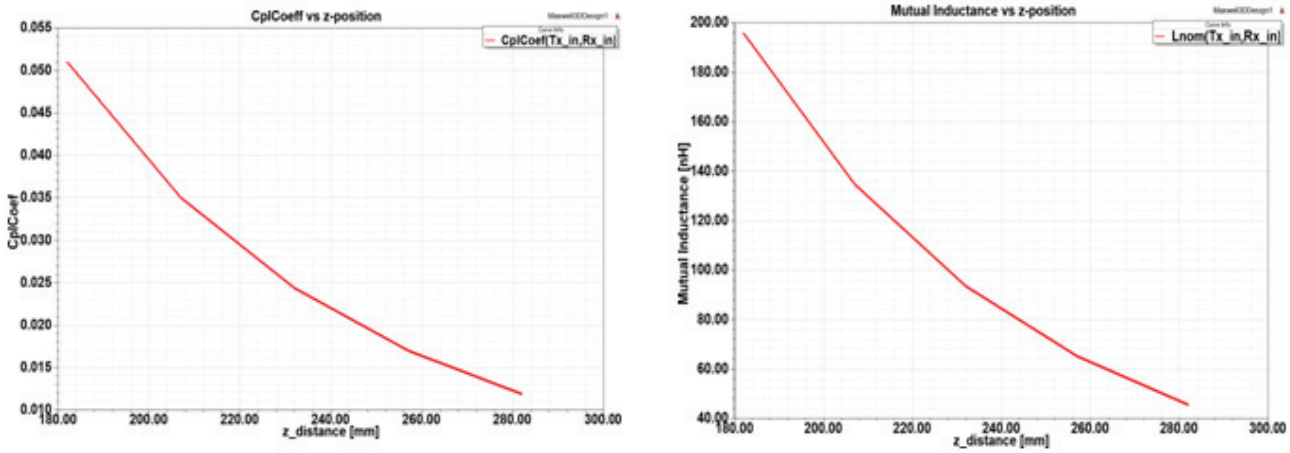


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:51:19 UTC from IEEE Xplore. Restrictions apply.



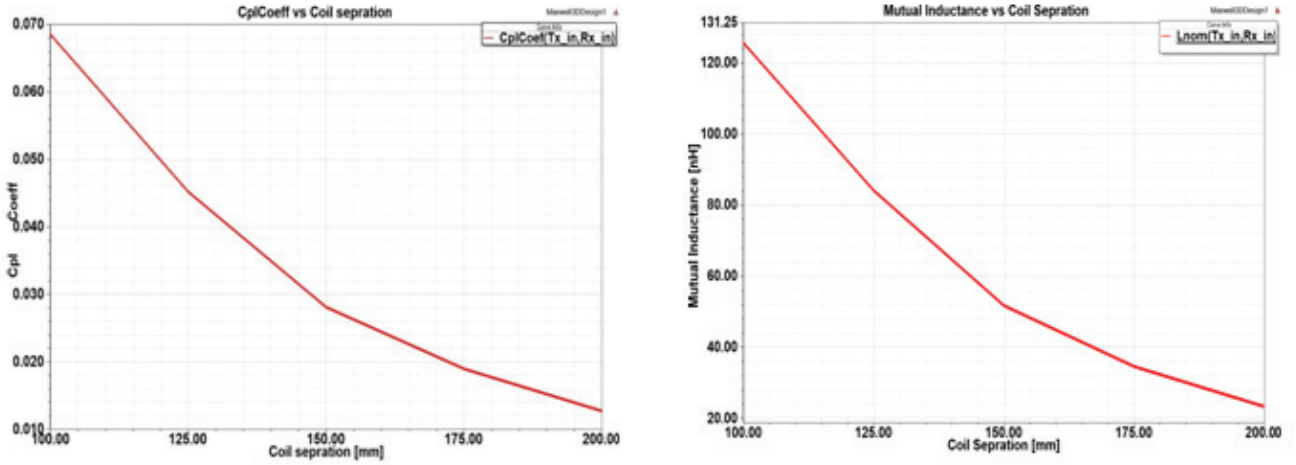

(a) (b)

Fig. 6. (a) Coupling coefficient vs Coil separation for circular coil structure (b) Mutual inductance vs Coil separation for Circular Coil Structure




(a) (b)

Fig. 7. (a) For Helical Coil Structure (coupling coefficient VS. Z-Position of the bottom of the secondary helix) (b) For Helical coil structure (Mutual inductance vs z-position of the bottom of the secondary helix)




(a) (b)

Fig. 8. (a) Coupling coefficient vs coil separation for square coil structure (b) Mutual inductance vs Coil separation for square coil structure



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:51:19 UTC from IEEE Xplore. Restrictions apply.

## VI. CONCLUSIONS

Nowadays, the area of wireless power transmission techniques has reawakened interest among both industrial and academic researchers. This technique also aids in the reduction of energy shortages and power outages. This paper explains the well-known formula and methods for analyzing WPT with magnetic resonance. Out of all the various WPT techniques MRC has emerged as the most preferred power transfer technique, due to its high operating range of several KWs, and high efficiency over a larger air gap. Although the idea of WPT using magnetic resonance is well-known presently, only a little amount of study has been done to compare and analyze performance of coil having different coil structures and geometries. The amount of power transfer takes place between coils depends on various factors, but in this analysis, focus has been laid on the effects of coil shape and geometry in MRC WPT and all the remaining parameters of the system are kept same. For this purpose, three most commonly used coil structures are analyzed. For each of the three coils used, the magnetic field is shown at 100 mm separation. Variation of coupling coefficient and mutual inductance with coil separation is also plotted at every 50mm interval starting from 100 mm to up to 200 mm. Fig. 9 display the value of the coupling coefficient for all the coils at different intervals.

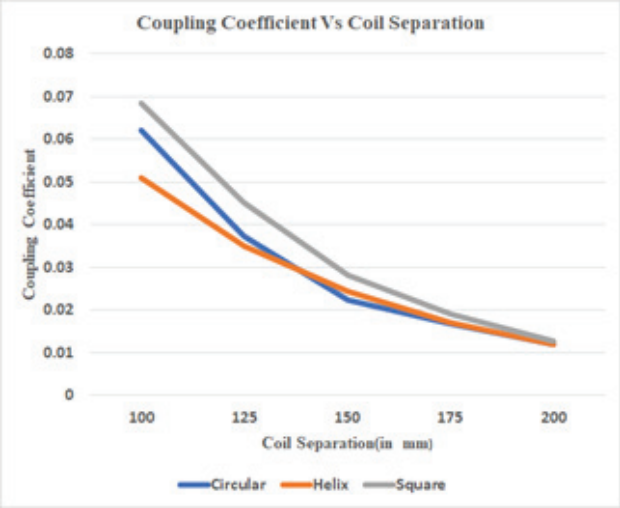



Fig. 9. Coupling coefficient vs coil separation for all the coils

From Fig. 9, it can be seen that at each separation, the value of coupling coefficient for the square coil is highest among all three types of coils. Thus, it is concluded that the square coil structure is the most efficient coil structure for power transfer.

## APPENDICES

POWER RATING FOR EV BATTERY = 3.3 kW

Voltage Applied at Transmission coil = 132.5V AC,

Current Applied at Transmission coil = 24.67 A

## ACKNOWLEDGEMENT

We thank CONIT Publisher, annexed sponsors, and Delhi Technological University for providing us with such a wondeful opportunity to present our findings. We'd also like to extend our profound gratitude for enabling us in accomplishing our research by offering an enlightening opportunity.

## REFERENCES

[1] Suzhi Bi, Chin Keong Ho, Rui Zhang, "Wireless Powered Communication: Opportunities and Challenges", *IEEE Communications Magazine,* Singapore, April 2015

[2] Lala Bhaskar, Pradeep Kumar, Kishore Naik Mude, "Wireless Power Transfer to Low Power Devices", IEEE 1st International Conference on Advances in Information Technology (ICAIT), Chikmagalur, India,July 2019

[3] Jeong S; Jang .J.; Kum D.; Lee M.S., "Charging Automation for Electric Vehicles: Is a Smaller Battery Good for Wireless Charging Electric Vehicles?" IEEE Transactions on Automation Science and Engineering, Madrid, Spain, April 2018

[4] Chih-Jung Chen, Tah-Hsiung Chu, Chih-Lung Lin, Zeui-Chown Jou, "A Study of Loosely Coupled Coils for Wireless Power Transfer", IEEE Transactions on Circuits and Systems—II: Express Briefs,Taiwan, July 2010

[5] Motoki Sato, Gaku Yamamoto, Daisuke Gunji, Takehiro Imura, Hiroshi Fujimoto, "Development of Wireless In-Wheel Motor Using Magnetic Resonance Coupling", IEEE Transactions on Power Electronics, Tokyo, Japan, July 2016

[6] Farhana Haque Sumi, Lokesh Dutta, Dr. Farhana Sarker, "Future with Wireless Power Transfer Technology", J Electr Electron Syst, Bangladesh, Nov. 2018

[7] Subhadeep Bhattacharya, Y.K.Tan, "Design of Static Wireless Charging Coils for Integration into Electric Vehicle", IEEE International Conference on Science, Engineering and Technology(ICSET), Nepal, October 2012

[8] Jianghua Lu, Guorong Zhu, Chunting Chris Mi, "Foreign Object Detection in Wireless Power Transfer Systems", IEEE Transactions on Industry Applications, China, Feb. 2021

[9] Werachet Khan-ngern, Heinz Zenkner, "Wireless power charging on electric vehicles", IEEE International Electrical Engineering Congress (iEECON), Chonburi, Thailand, March 2014



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:51:19 UTC from IEEE Xplore. Restrictions apply.

*2022 2nd International Conference on Intelligent Technologies (CONIT)*
*Karnataka, India. June 24-26, 2022*

# Comparison of Neural Networks and XGBoost Algorithm for Music Genre Classification

Ritik Gusain
*Department of Electrical Engineering*
*Delhi Technological University*
New Delhi, India
gusainritik16@gmail.com

Sagar Sonker
*Department of Electrical Engineering*
*Delhi Technological University*
New Delhi, India
sagarsonker590@gmail.com

Sachin Kumar Rai
*Department of Electrical Engineering*
*Delhi Technological University*
New Delhi, India
srai47727@gmail.com

Ankita Arora
*Department of Electrical Engineering*
*Delhi Technological University*
New Delhi, India
arora.ankita09@gmail.com

S.T. Nagarajan
*Department of Electrical Engineering*
*Delhi Technological University*
New Delhi, India
stnagarajan@dtu.ac.in

***Abstract*—The purpose of this research paper is to implement Neural Networks and XGBoost based models in order to generate music genre classifiers & analyze their performance. This analysis is based on their accuracy, speed and various other factors. The dataset is taken from the kaggle website. The dataset includes audio files, mel spectrogram & CSV with extracted features. For the Neural Network (NN) model, audio files are used as the dataset. This dataset is analyzed thoroughly and different waveforms are plotted using the python package called Librosa. The dataset is preprocessed and its MFCC features are extracted. The MFCC values are treated as the input and their respective music genres are supposed to be the output for the Neural Network model. The model is trained with four neural network layers. Similarly for the XGBoost algorithm, CSV with extracted features is taken as the main dataset. Some fraction of the dataset is trained with the model and for the rest of the data is analyzed for comparison between the two. The accuracy/ error plot of test/training data and a confusion matrix, is generated for both and models. Models are trained using open-source libraries- Tensorflow, Keras, Xgboost, Sklearn, etc**

***Keywords—Neural Networks, XGBoost, Mel Frequency Cepstral Coefficients, Machine Learning, Digital Signal Processing.***

## I. INTRODUCTION

This Music is an art form in which one expresses himself through the elements of sound i.e. melody and harmony (other forms of art includes writing, painting, cinema, etc). The music can be divided into different genres and every genre gives a different feel. Some of the famous genres of music are blues, classical, rock, pop, hip-hop and country.

Each genre has some unique features and these features can be extracted using speech recognition techniques. Using machine learning, a classifier can be generated which is helpful for getting the idea of listeners' music taste and placing further song recommendations. Previously, Y Jeong and K Lee [1] applied deep neural networks for feature learning of audio and generated the classifier.. Bob L Sturm [2] analyzed the GTZAN dataset that is available on kaggle website. Panagakis, Benetos and Kotropoulos [3] used the multilinear approach for the classifier inspired by the auditory cortical processing model. Tao Li [4,7] used DWCHs that captures the global and local information of music dataset and gives histogram on its coefficients. Xueyu Zhou [5] used the XGBoost and Node2vec BP model to check the composer's influences on the revolution of music. Benjamin Murauer [6] classified music using CNN for classification using spectrograms and other methods like DNN, XGBoost. Stepan Evstifeev and Ivan Shanin [8] extracted music features like rhythm, timbre, tone, etc and used signal processing methods for classification.

We have compared Neural Network and XGBoost algorithms on many factors and determined which one should be preferred for a particular data for music genre classification. We also used signal processing techniques for feature extraction.

## II. PROPOSED ALGORITHMS

### *A. Neural Networks*

Basically, the neural networks model is inspired from the human neural system, which consists of the brain and it is connected by several neurons which give signals to different parts of the body.

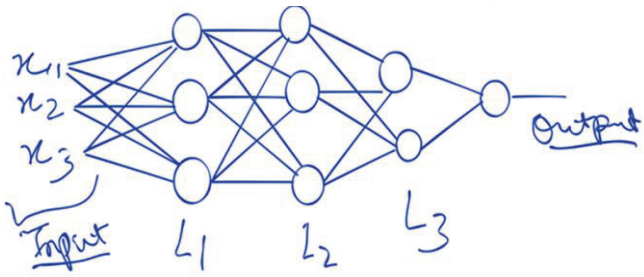



Fig. 1. Neural Networks model

A neuron is the basic building block in a neural network. They are further connected to other neurons through some weight, shown as lines in the figure. The data is taken from the input layer and is forward propagated through these perceptrons. Consider the circle shown in the figure as an activation function.

Let 'x' be the output from previous step and 'w' be the weight between the two neurons, then the input value for a layer will be,

$$\sum = (x_1 \times w_1) + (x_2 \times w_2) + \cdots + (x_n \times w_n) \qquad (1)$$

Let this summation value be x.w hence,

2022 International Conference on Intelligent Technologies (CONIT) | 978-1-6654-8407-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CONIT55038.2022.9847814



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:53:19 UTC from IEEE Xplore. Restrictions apply.

$$x.w = (x_1 \times w_1) + (x_2 \times w_2) + \cdots + (x_n \times w_n) \quad (2)$$

Then a bias term 'b' is added for better generalizing of activation function,

$$z = x.w + b \quad (3)$$

Now this 'z' is passed through activation function to get the estimated output 'y' for that layer,

$$\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}} \quad (4)$$

Now, the difference between the actual output 'y' and estimated output ' $\hat{y}$ ' calculated, with the help of this difference, the cost function 'E' is calculated.

$$E = M.S.E. = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (5)$$

Now derivation of this cost function is calculated wrt every weight and bias using chain rule,

$$\frac{\partial E}{\partial w_i} = \frac{\partial E}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial z} \times \frac{\partial z}{\partial w_i} \quad (6)$$

Now the new value of weights and bias is updated. How much of value is updated is determined by a parameter (alpha) called learning rate.

$$w_i = w_i - (\propto \times \frac{\partial E}{\partial w_i}) \quad (7)$$

and updated bias b,

$$b = b - (\propto \times \frac{\partial E}{\partial b}) \quad (8)$$

The whole process of going backwards and updating the values is known as backpropagation. The data set taken for the Neural Network model is

TABLE I. DATASET 1 INFORMATION

| | |
|---|---|
| Total size | 126.27 MB |
| File format | .wav |

It contains 3 directories, each representing a music genre. Each directory contains 32 audio files, each having 30 sec length.

TABLE II. DATASET 1 CONTENT

| **Directory's Name** | **Number of files** | **Length** |
|---|---|---|
| blues | 32 | 30 seconds |
| classical | 32 | 30 seconds |
| country | 32 | 30 seconds |

*B. XGBoost*

XGBoost algorithm is a type of ensemble learning algorithm. Basically, ensemble learning trains different M.L. models. It averages the prediction of all models, hence gives a better overall model. Its two main types are: 1) Bagging 2) Boosting. In bagging, the dataset is divided into different subsets and each subset is used to train a unique decision tree model. The final model is average of all the weak models. Hence, it is a parallel kind of model.

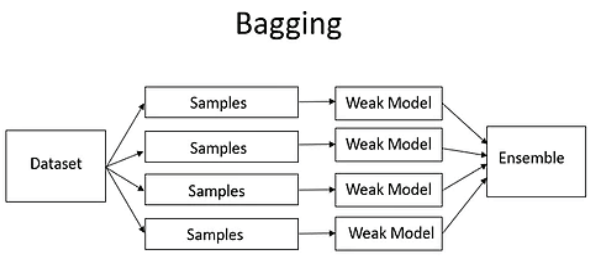



Fig. 2. Bagging

In boosting, firstly the dataset is used to train the first decision tree model. Then the second model is trained using the previous model and a parameter (alpha) called learning rate. Similarly, all the models are trained and one gets a final model. Hence, it is a sequential kind of model.

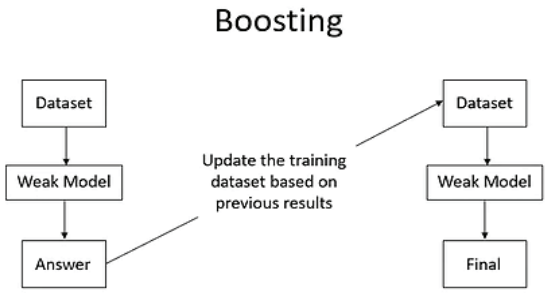



Fig. 3. Boosting

A decision tree is a kind of tree built considering all possible conditions. Based on the truth and falseness of the condition, it splits into further subtrees. Xgboost is a boosting algorithm and it uses decision trees as its sequential models. The dataset taken for XGBoost Algorithm is

TABLE III. DATASET 2 INFORMATION

| | |
|---|---|
| Total size | 11.04 MB |
| File format | .csv |

It contains mathematical features of audio files. Each 30 second song is splitted into 3 second samples, hence the dataset is increased by 10 times. The file includes pre-calculated mean and variance of multiple features of audio samples.Some of the features given in Table IV below

TABLE IV. DATASET 2 CONTENT

| | |
|---|---|
| chroma_stft | zero_crossing_rate |
| rms | harmony |
| spectral_centroid | perceptr |
| spectral_bandwidth | tempo |
| rolloff | mfcc1 |



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:53:19 UTC from IEEE Xplore. Restrictions apply.

## III. METHODOLOGY FOR PREPARING DATASET

### A. Data Visualization

In order to build a machine learning model, there are two most imp steps: 1) Analyzing and the pre-processing the data set. 2) Model selection and tuning its parameters.

It's important to bear in mind that pre-processing cannot be done unless and until the data has been analyzed. In order to analyze the dataset, one must study and implement everything about the audio dataset. So firstly, the audio waveform was extracted from the wav file.

#### 1) Audio Waveform

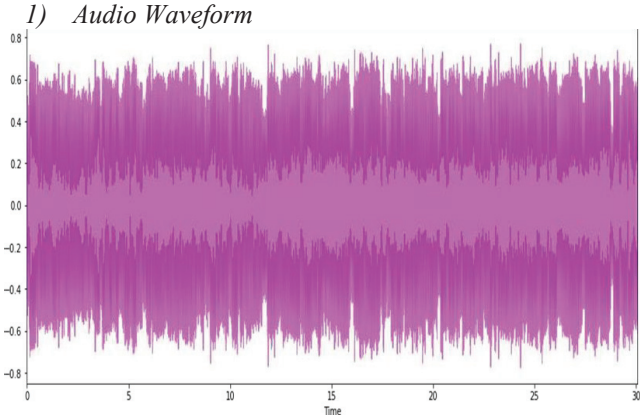



Fig. 4. Sampled audio wave plot

The dataset was loaded with a sampling rate of 22050 Hz and a bit depth of 16 bits. It means it will take 22050 samples in 1 second & it will quantize the waveform in 2^16 i.e. 65,536 values.

#### 2) F.F.T. Taken

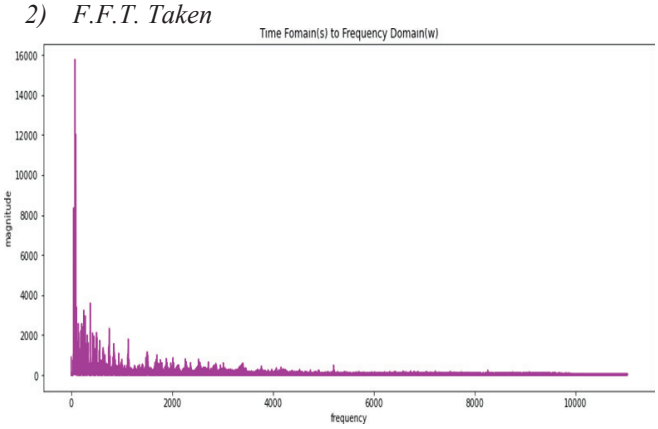



Fig. 5. Frequency domain spectrum

Time domain is converted into frequency domain. X-axis represents frequency.

#### 3) Spectrogram Generation

In the frequency spectrum, we lost the information about the contribution of each frequency at a particular time. Hence, a short-time fourier transform (STFT) is used. It divides the 30 seconds longer audio into smaller segments (some milliseconds) of equal length. Hence, it computes changing spectra as a function of time called Spectrogram for each and every segment. Using this spectrogram is produced which is a function of both time and frequency. The color represents the respective magnitude.

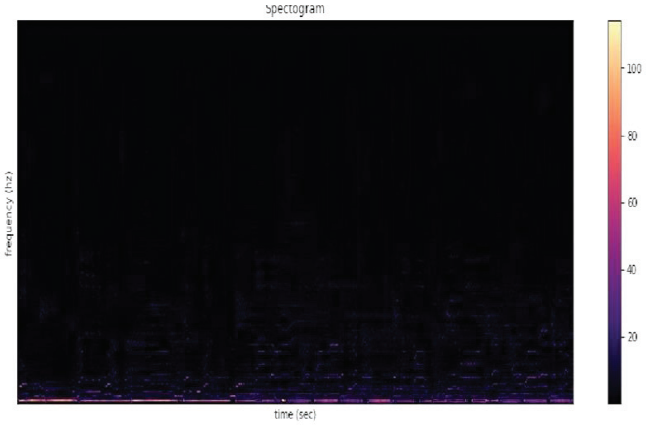



Fig. 6. Spectrogram generated

Here,the concept of removing discontinuity at the edges is removed. For non-overlapping segments, at the ends of each segment there is discontinuity present which makes aur model faulty and can't be ignored. Due to this discontinuity, the extracted features will differ from real ones. To avoid this, we used overlapping segments. We defined the parameters of window and hop length.

#### 4) On Logarithmic Scale

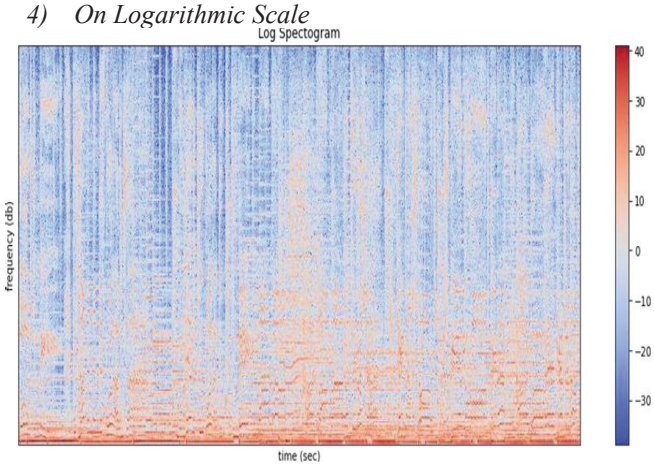



Fig. 7. Spectrogram in decibels(db)

The Spectrogram color produced previously was not distinguishable. For that log the absolute value of STFT is being taken. Now, the colorbar scale has changed to decibels. Using logarithmic operation, a more visible and significant spectrogram is generated.

#### 5) Mel Frequency Ceptrum

These are the steps to follow in order to calculate MFCCs

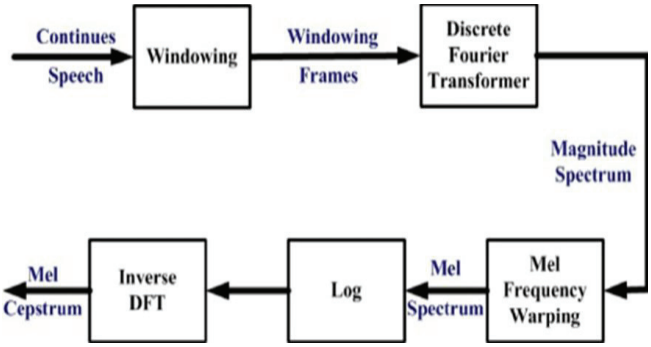



Fig. 8. Steps involved for calculating MFFCs



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:53:19 UTC from IEEE Xplore. Restrictions apply.

The MFCC features are used in the Neural Networks model as the input layer. With the desired hop length and widow length, MFCC spectrum generated for a single audio file is as follows:

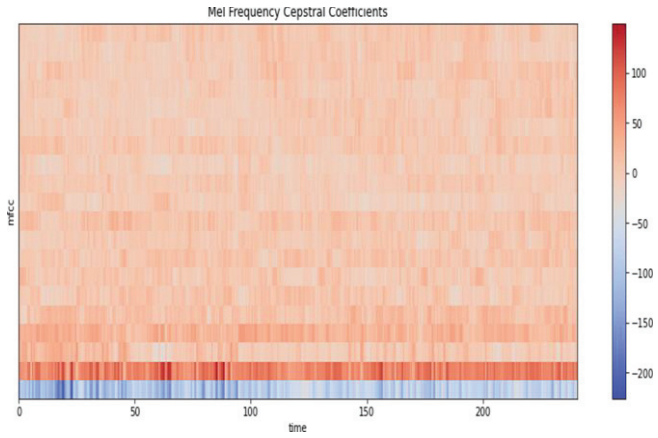



Fig. 9. Mel Cepstrum

### B. MFCC Extraction

While calculating mfccs, we sampled the audio dataset using the following parameters:

TABLE V. MFCC PARAMETERS

| | |
|---|---|
| Sampling Rate | 22050 Hz |
| Hop Length | 64 samples |
| Window Length | 1024 samples |
| Number of segments | 5 |

In order to maintain continuity, we introduced the hop and window concept. And in order to increase dataset size, we divided the 30 second audio file into segments of 6 seconds each.

### C. Preprocessing

Preprocessing plays a very important role in machine learning. This is a step that is used to put the data in a format that can gain a massive amount of information during model training. In most of the cases, the data one will use to train the model may not be in the right format. It might contain wrong data types, NULL values, duplicate values, etc. All these things are being taken care of in the preprocessing step. In most cases the softwares such will throw an error when trying to analyze the dirty data. In conclusion, pre-processing is necessary in any analytics task.

The dataset (audio files & csv with extracted features) was taken for GZTAN kaggle website and is stored in google drive. The audio dataset was preprocessed using the audio analysis library - Librosa. Basically, code was implemented in Python in order to process the music dataset and to extract the MFCC features. The MFCC values are treated as the input and their respective music genres are supposed to be the output for the Neural Network model. The csv with extracted features are treated as the dataset for training the XGBoost model.

### D. Division Of Dataset

For both the models, 70% of the complete dataset is used for training the model. Rest of the dataset i.e. 30% is used to test the model accuracy.

## IV. RESULT AND ANALYSIS

The performance analysis for classification of musical genres for the proposed algorithms are discussed below in detail. The performance is analyzed in terms of terms of factors such as accuracy, speed, processing power, dataset size, dataset type, etc.

### A. Neural Networks

These are the four layers used for modeling:

TABLE VI. ALGORITHM 1 SUMMARY

| TYPE OF LAYERS | OUTPUT SHAPE | NUMBER OF PARAMETERS |
|---|---|---|
| lstm_1 (LSTM) | (None, 32) | 6784 |
| dense_2 (Dense) | (None, 32) | 1056 |
| dropout_1 (Dropout) | (None, 32) | 0 |
| dense_3 (Dense) | (None, 3) | 99 |

Layer 1 (LSTM) parameters: activation= 'tanh', recurrent activation= 'sigmoid', kernel initializer= 'glorot_uniform', recurrent initializer= 'orthogonal'.

Layer 2 parameters: activation = 'relu', kernel initializer = 'glorot_uniform'.

Layer 3 (Dropout): 20%.

Layer 4 (Dense): activation='softmax', kernel initializer=' glorot_uniform'.

Additionally for all the layers, bias is used and initialized as 0. On layer 2 (dense), we introduced L2 regularization with {lambda} as 0.01. Dropout and regularization are used in order to avoid overfitting. Basically in dropout, we randomly drops some of the neurons layer during training the model. And in regularization, we penalize the cost function by adding a regularsation parameter.

Model is compiled as: optimizer=optimizers.Adam(), loss='sparse_categorical_crossentropy', metrics=['accuracy'], learning rate= 0.001.

When we used half (0.5x) and quarter(0.25x) the amount of data used, the accuracy was lower than current. It implies the accuracy for the test can further be increased by using more data (5-10 times) but it takes comparably more RAM and processing power.

For 40 epochs, the accuracy and loss is plotted for training as well as testing data set. Following are the loss and accuracy plots calculated for each epoch



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:53:19 UTC from IEEE Xplore. Restrictions apply.

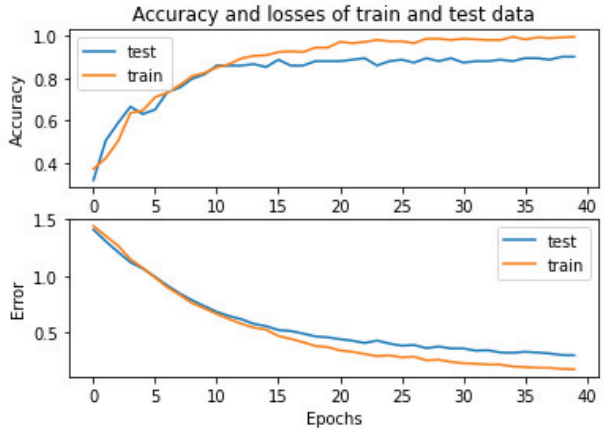



Fig. 10. Accuracy and error plots

After 40 epochs, the accuracy of training data reached 99.7%. For test data, it reached 90.28%.

*B. XGBoost*

XGBoost is implemented using 1024 sequential models with a learning rate of 0.1. Following are the major parameters:-

TABLE VII. ALGORITHM 2 SUMMARY

| Parameters | Values |
|---|---|
| Learning Rate | 0.1 |
| Eta | 0.2 |
| Gamma | 0 |
| No. of Estimators | 1024 |
| objective | 'binary:logistic' |
| booster | 'gbtree' |
| maximum depth | 3 |

The accuracy reached for the XGBoost algorithm using 1024 sequential decision trees is 89.52%.

Following is the confusion matrix generated for the XGBoost model with respect to the musical genres:

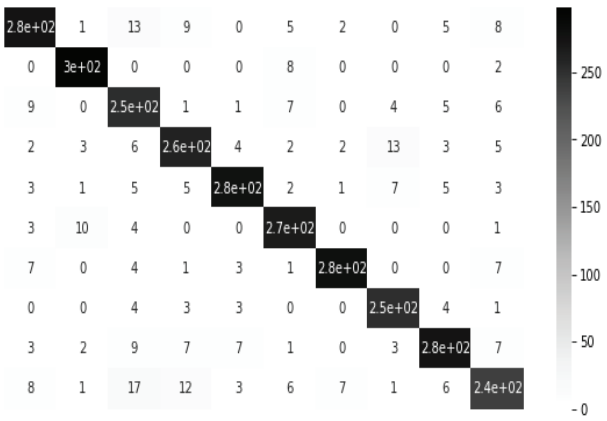


Fig. 11. Confusion matrix

It clearly shows that the models using xgboost algorithm are trained accurately.

By increasing the gamma value, the accuracy was decreasing. This is due to the fact that the dataset is already pre-processed quite well. So by increasing the regularization parameter, it won't affect that much. The most optimal value of eta is 0.2 to handle overfitting.

The accuracy for both the algorithms is similar but there are many factors that determine which algorithm to use. It is preferred to initially pre-process the data, so that the important features for the model can be extracted. If one can extract different mathematical features and store them in tabular form, he can apply XGBoost. For more complex datasets neural networks are preferred, one can extract the features and use them as the input layer. XGBoost takes comparatively less time, less processing power to train the model than Neural Networks. But Neural Networks are preferred for large datasets as it gives better accuracy when data size is huge. For classification where data is based on time series or image like datasets, some modified neural networks eg-LSTM, CNN are preferred.

## V. CONCLUSION

In this paper, we studied and implemented digital signal processing for extracting audio data features & machine learning techniques for generating the music genre classifier. We sampled and pre-processed the .wav data and extracted its MFCC features for the Neural Network model. We implemented LSTM (long short term memory) and ANN (artificial neural network) as Neural Network model layers. We implemented the XGBoost algorithm on mean and variance of different audio features. We compared both the algorithms based on many factors and determined which one should be preferred for a particular data. The results we got from both the models are quite impressive. The accuracy of the Neural Networks model can be further improved by increasing the size of the dataset but it takes more processing power which is limited in our computer. The music classifier we generated can be used in music streaming platforms such as spotify, wynk etc for placing recommendations to the listeners.

## REFERENCES

[1] Young Jeong and Kyogu Lee,"Learning temporal features using a deep neural network " ,Proceedings of the 17th ISMIR Conference, New York City, USA, pp. 434-440, 2016.

[2] Bob L Sturm,"An analysis of the GTZAN music genre dataset," Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies, pp. 7–12, 2012.

[3] Panagakis, Benetos and Kotropoulos, "Music genre classification: a multilinear approach". presented at the International Symposium Music Information Retrieval, pp. 583-588, 2008.

[4] Tao Li and M. Ogihara,"A comparative study on content-based music genre classification",Presented at 26th annual international ACM SIGIR conference on Research and development ,pp. 282–289, 2003.

[5] Xueyu Zhou,"Music Exploration Based on XGBoost Algorithm and Node2vec-BP Model", Presented at 6th conference on Machine Learning Technologies, ICMLT, pp.96, 2021.

[6] Benjamin Murauer, Günther Specht,"Detecting Music Genre Using Extreme Gradient Boosting", WWW Companion Proceedings of The Web Conference, pp. 1923–1927, 2018.

[7] Tao Li and M. Ogihara,"Music genre classification with taxonomy", Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.197-200, 2005..



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:53:19 UTC from IEEE Xplore. Restrictions apply.

[8] Stepan Evstifeev and Ivan Shanin, “Music Genre Classification Based on Signal Processing”, Proceedings of the 20th International Conference in Data Analytics and Management in Data Intensive Domains (DAMDID), pp. 157-161, 2018.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:53:19 UTC from IEEE Xplore. Restrictions apply.

*2022 2nd International Conference on Intelligent Technologies (CONIT)*
*Karnataka, India. June 24-26, 2022*

# Constructing Sentiment Sentence Embedding Model Using Transfer Learning

Rajesh Kumar Yadav
*Department of Computer Science and Engineering*
*Delhi Technological University*
Delhi, India
rkyadav@dtu.ac.in

Shivam Singhal
*Department of Computer Science and Engineering*
*Delhi Technological University*
Delhi, India
shivamsinghal1012@gmail.com

Shashank Chugh
*Department of Computer Science and Engineering*
*Delhi Technological University*
Delhi, India
shshankchugh@gmail.com

Shivam Jaiswal
*Department of Computer Science and Engineering*
*Delhi Technological University*
Delhi, India
shivamjaiswal82@gmail.com

2022 International Conference on Intelligent Technologies (CONIT) | 978-1-6654-8407-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CONIT55038.2022.9847927

***Abstract*—For Sentiment classification a suitable embedding to train a machine learning model is needed. A lot of research has been done on generating embeddings that are capable of storing the context of words, but they are inefficient for sentiment extraction as they cannot differentiate words used in the same situation but contrasting sentiment polarity. To address this issue, a lot of researchers have proposed different embedding models, but most of them are word embeddings where lot of semantic information is lost. We have constructed a model by fine-tuning Google's Universal Sentence Encoder using a Deep Neural Network. The embedding layer will be allowed to retrain through backpropagation on the Sentiment dataset. Once training is completed, the hidden and output layers will be discarded, leaving only the USE layer, the output of this layer will be used as our Sentiment Embedding.**

***Keywords*—Sentiment Embedding, Sentence Embedding, Fine tuning, Universal Sentence Encoder, Transfer Learning**

## I. Introduction

Large Corporates get a lot of unstructured data about their users through various means like tweets, blogs, comments etc. Customers post reviews, issues and seek help from product owners through social media. For example, a user who's facing issues in using Windows 11 may post a tweet seeking help to solve the issue. These reviews are necessary for companies to analyze how their product is performing in the public market. But to extract relevant information from the large amount of data we need to perform Machine Learning Analysis such as Sentiment. With the help of Sentiment Analysis, companies can analyze the sentiment of a user from social media posts, extract critical product reviews from all over the internet.

For the Sentiment Classification Machine Learning Model, we need to convert text to its vector representation that can be used for further numerical computations. The first approach is Word Embeddings. Word Embeddings in simple terms, is a way to convert words into its numerical representation, usually real valued vectors that portray the meaning of a word such that words with identical meaning are expected to be close in vector space.

To use Word Embedding for text, one simple method is to take the average of all word embeddings to generate a vector for sentence. This approach has many disadvantages, one of them is no respect for order, that means the average would be the same irrespective of the ordering of words, if we write the text in different order. In the recent study, several models have been proposed to figure out the best way to combine the word embedding vectors to obtain sentence embedding such that there is minimal loss of semantic information.

Fig. 1. Sentences in Sentence Encoder Vector Space.

One of the famous sentence embedding models is Universal Sentence Encoder (USE), which focuses on the semantic meaning of the word but neglects the sentiment in texts. This is a typical problem while using these embedding directly for sentiment analysis, as two sentences with the same situation but different sentiment polarity are placed closer in the vector space. For example, "Won a match" and "Lost a match" would be placed nearby in Sentence Embedding vector space as shown in Fig. 1, but they should be largely distant in the sentiment embedding vector space as shown in Fig. 2.

To solve this problem, one simple approach is to retrain the Universal Sentence Encoder model on sentiment data set. This



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:58:28 UTC from IEEE Xplore. Restrictions apply.

Fig. 2. Sentences in Sentiment Embedding Vector Space.

method has several disadvantages, one is that these models are very complex and large and thus to train these models, a very large dataset is necessary. Secondly, to train such a complex model on a very large dataset, heavy computing machines are required which are very costly. We propose a method for fine tuning the Universal Sentence Encoder on a large dataset using which we can generate sentiment embeddings and transfer this knowledge to other datasets. This approach eliminates the need of having a large dataset and heavy computing machines every time. Along with this , this approach allows us to use algorithms other than deep learning algorithms according to demands of the problem.

## II. Related Work

The research in word and sentence embeddings with deep learning has been advancing after T. Mikolov et al. [1], introduced the word2vec algorithm in 2013. Word2vec is a method to convert words to its vector representation using the Skip-gram and Common Bag of Words (both involve Neural Network). The vector representation from the word2vec model contains the semantic meaning of words that can be helpful in various NLP problems.

D. Cer et al. [2] introduced Universal Sentence Encoder which transforms sentences into its numerical vectors which depicts the similar meaning as the sentence itself does. They show that results on sentence embedding surpass word level. They show effective performance and efficient results on this model. They present two variants of the model, one with a Transformer design and another with Deep Averaging Network trained on huge data from Wikipedia, news, QA, and discussion blogs.

Word embedding trained by classical neural networks is not efficient in sentiment classification. Those methods generally model the semantic information of words, but cannot differentiate words with the same situations but opposite sentiment (e.g. won and loss). To address this issue, Tang et al. [3], presented the Sentiment Specific Word Embedding model extending the Collobert et al. [4] model, a word representation for Twitter sentiment analysis. They trained SSWE by combining the sentiment information into a neural network model. Tang et al. [5] trained sentiment-specific word embedding by introducing a few neural network layers to efficiently encode semantic and sentiment information. On word level, sentiment embeddings were found to be efficient for determining the relationship between words. But they do not perform well on sentences.

From the related work papers, we found out that a lot of research has been done on embeddings that describe the situation of words, but they are not suitable for sentiment analysis as they cannot differentiate words with the same sense but dissimilar sentiment polarity (e.g. “won” and “loss”). To address this issue, a lot of researchers have proposed models, but most of them are word embeddings. We also noted that sentence embeddings are better than word embeddings because there is no respect for order and loss of semantic information in word embeddings.

## III. Dataset

### A. Twitter Airline

This data set is used for sentiment analysis about the Airline facilities or services using customer’s tweet. The tweets were collected from Crowdflower’s Data for Everyone library. Texts are categorized into positive, negative and neutral sentiments. The whole data set consists of 14,461 tweets. We used this dataset for training our embedding models.

### B. IMDB movie review

This data set is developed for sentiment analysis on movie’s review (A. Maas, 2011) [18]. Review of different movies were collected from the official IMDB website to construct this data set. Reviews were divided into different sentiment such as negative and positive on the basis of rating of the review while neutral reviews were ignored. The whole data set contains 50 thousand movies reviews with equal distribution for training and testing data, both having 25 thousand each. We used this dataset for testing our model.

## IV. Experimental Design

### A. Preprocessing

Before we move on to train a model , we need to perform some preprocessing steps like cleaning , data quality assessment and data transformation. For example converting emoticons, emojis to texts which are closest to their context. Other examples are conversion of chat words like ‘AFK’ to their real full form ‘Away From Keyboard’ and transformation of hashtags into a meaningful sentence.

### B. Universal Sentence Encoder

Universal Sentence Encoder has two architectures, both take sentences as input and outputs 512 dimensional similar meaning vectors known as embeddings. One is based on Transformer Architecture which compute vector of the sentence taking both identity and ordering into account. Other is Deep Averaging Network, In DAN architecture a deep averaging network is used to produce sentence embeddings. Input Embeddings for words are averaged together to fed a deep neural network with these input embeddings. Time Complexity of DAN architecture is linear in length of input.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:58:28 UTC from IEEE Xplore. Restrictions apply.

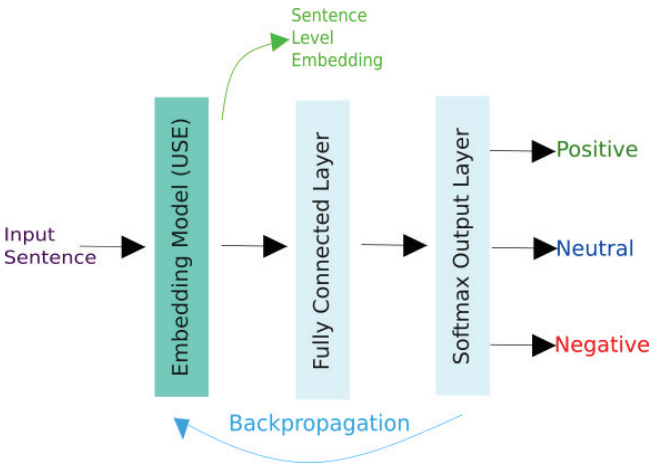



Fig. 3. Proposed Deep Learning Model Architecture.

### C. Model

As shown in Fig. 3 Our Model comprises a Sentence Encoder at the first layer to encode textual input to its vector representation , which generates 512 vectors, followed by one ReLU layer and an output layer having three nodes each representing a sentiment (positive, negative, neutral). The embedding layer will be allowed to retrain through backpropagation on the Sentiment data set. Once training is completed, the hidden and output layers will be discarded, leaving only the embedding layer which is the required output of our study.

### D. Evaluation

We have trained our model on Twitter Airline Dataset (described in section 3.1). For testing our model we have used IMDB Movie Review (described in section 3.2). In this experiment, we compare four models:

- Original USE DAN Architecture Model
- Original USE Transformer Architecture Model
- Fine tuned USE DAN Architecture Model on Sentiment Dataset
- Fine tuned USE Transformer Architecture Model on Sentiment Dataset

To measure effectiveness of a model we generate random pairs of text of same and different classes and then calculate distance between them using equation below,

$$c(M, N) = \frac{M \cdot N}{|M| \times |N|} \tag{1}$$

where M and N are the two vectors and c(M,N) is the cosine distance between M and N.

In the next step, we can take average of distance for same and different classes. Similarity is defined as,

$$similarity = \frac{\sum_{s=0}^{S} c(A, B)}{S} - \frac{\sum_{d=0}^{D} c(C, D)}{D} \tag{2}$$

where, c(A,B) is the cosine distance between A and B , S and D are the number of random samples of same class and different class respectively.

## V. Result and Analysis

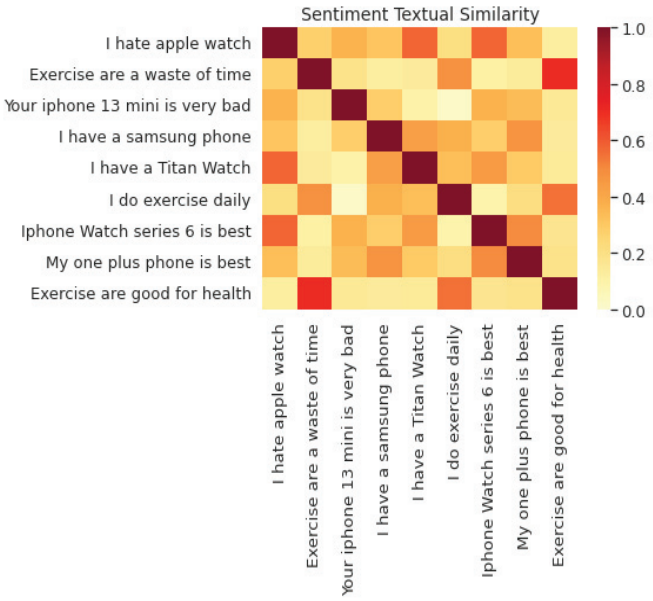



Fig. 4. Confusion Matrix of Original USE Embeddings.

Fig. 5. Confusion Matrix of Fine tuned USE Embeddings (Sentiment Embedding)

For testing our model results, we first take a few samples, three each of negative, positive and neutral and form a confusion matrix for these samples where each cell of confusion matrix depicts the cosine distance between vectors of corresponding text pairs. Fig 4, 5 show confusion matrix for Original USE (Transformer Architecture) and Fine tuned USE (Transformer Architecture) respectively. It can be observed from the confusion matrix diagrams that clear clusters are getting formed in our sentiment model whereas there are no



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:58:28 UTC from IEEE Xplore. Restrictions apply.

distinct clusters in the confusion matrix of USE. Formation of clusters signifies that in our sentiment model positive , negative and neutral data points can be distinctively classified and distinguished from each other unlike traditional USE model where texts having different context may have very little cosine distance or texts having same context may have high cosine distance.

We evaluate four models: Original USE (DAN Architecture), Original USE (Transformer Architecture), Fine Tuned USE (DAN Architecture), and Fine Tuned USE (Transformer Architecture). The original USE models are not trained on sentiment data. All these models are evaluated on IMDB Review Dataset by using the method described in section 4.5. We calculate the difference in cosine distance between tweets from different class and from same class to evaluate our model. If cosine similarity between two vectors is close to 1 then it means they are highly correlated, and if it is close to 0 then they are not correlated.

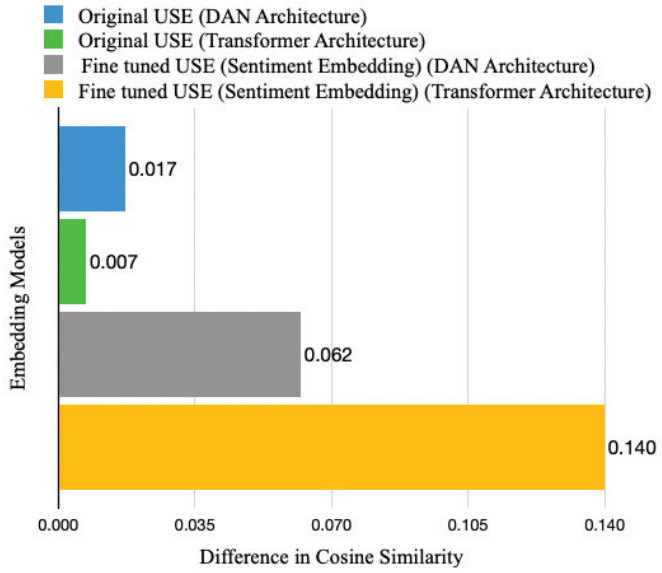



Fig. 6. Difference of Cosine Similarity compared for original and fine tuned model.

Fig. 6 shows that fine tuned USE with Transformer architecture has highest similarity score of 0.140 followed by Fine tuned USE with DAN architecture with a score of 0.062 subsequently followed by conventional USE models, both having scores less than 0.02. We can observe from the above data that similarity scores for Fine tuned models is significantly higher than that of conventional models which signifies that former will have better classification accuracy than the later for a given model.

## VI. Conclusion

We have presented a method to construct sentiment embeddings that can be directly used for Sentiment Analysis by fine tuning Universal Sentence Encoder on Sentiment Dataset. Our method is based on Deep Learning Architecture which classifies the embeddings into its sentiments, later we discarded the hidden and output layer keeping the fine tuned universal sentence encoder layer which is our final sentiment embedding model. This fine tuned sentiment embedding saves training time whenever we use it for any sentiment classification task without compromising with accuracy of classification. Thus, the model that we have proposed is better than conventional USE model both in terms of speed and accuracy.

## VII. Future Work

We have fine tuned Universal Sentence Encoder in our work. However other encoders such as BERT can be fine tuned using the same approach. Also, we have used the English version of Universal Sentence Encoder, and an English Dataset was used for this but the Multilingual Universal Sentence Encoder can be used for a multilingual dataset using the same method discussed in this paper. This approach is used for fine tuning on sentiment data, but similar approach can be used for other domains as well where USE is failing to provide efficient results.

## References


[1] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," arXiv [cs.CL], 2013.

[2] D. Cer et al., "Universal Sentence Encoder for English," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2018.

[3] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014.

[4] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in Proceedings of the 25th international conference on Machine learning - ICML '08, 2008.

[5] D. Tang, F. Wei, B. Qin, N. Yang, T. Liu, and M. Zhou, "Sentiment embeddings with applications to sentiment analysis," IEEE Trans. Knowl. Data Eng., vol. 28, no. 2, pp. 496–509, 2016.

[6] D. Tang, F. Wei, B. Qin, T. Liu, and M. Zhou, "Coooolll: A deep learning system for twitter sentiment classification," in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014.

[7] L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Refining Word Embeddings for Sentiment Analysis," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017.

[8] L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Refining word embeddings using intensity scores for sentiment analysis," IEEE ACM Trans. Audio Speech Lang. Process., vol. 26, no. 3, pp. 671–681, 2018.

[9] W. Xu, C. Callison-Burch, and B. Dolan, "SemEval-2015 task 1: Paraphrase and semantic similarity in twitter (PIT)," in Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015.

[10] S. Rosenthal, A. Ritter, P. Nakov, and V. Stoyanov, "SemEval-2014 Task 9: Sentiment Analysis in Twitter," in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014.

[11] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," arXiv [cs.CL], 2002.

[12] A. Severyn and A. Moschitti, "Twitter sentiment analysis with deep convolutional neural networks," in Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15, 2015.

[13] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

[14] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015.




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:58:28 UTC from IEEE Xplore. Restrictions apply.

[15] A. Olteanu, S. Vieweg, and C. Castillo, "What to expect when the unexpected happens: Social media communications across crises," in Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15, 2015.
[16] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: Understanding rating dimensions with review text," in Proceedings of the 7th ACM conference on Recommender systems, 2013.
[17] A. Go, R. Bhayani, and L. Huang, Twitter sentiment classification using distant supervision. Stanford, 2009.
[18] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:58:28 UTC from IEEE Xplore. Restrictions apply.

*2022 2nd International Conference on Intelligent Technologies (CONIT)*
*Karnataka, India. June 24-26, 2022*

# Contrast Aware Image Dehazing using Generative Adversarial Network

Anil Singh Parihar
*Department of Computer Science*
*Delhi Technological University*
Delhi, India
parihar.anil@gmail.com

Kavinder Singh
*Department of Computer Science*
*Delhi Technological University*
Delhi, India
kavinder85@gmail.com

Aryan Ganotra
*Department of Computer Science*
*Delhi Technological University*
Delhi, India
aryanganotra7@gmail.com

Aviral Yadav
*Department of Computer Science*
*Delhi Technological University*
Delhi, India
aviralx10@gmail.com

Devashish
*Department of Computer Science*
*Delhi Technological University*
Delhi, India
devashishyadav2531@gmail.com

2022 International Conference on Intelligent Technologies (CONIT) | 978-1-6654-8407-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CONIT55038.2022.9847710

***Abstract*—One of the most demanding aspects of computer vision is image dehazing. This paper presents a learning-based method to generate a dehazed image from a hazy input image using a densely connected end-to-end encoder-decoder-based GAN (Generative Adversarial Network) to facilitate feature extraction from images and their utilization. We added a Contrast-aware channel attention module to enhance and utilize the features extracted for generating more natural appealing dehazed images with better lightness computation, edges and structure detection. We used a traditional patch-based CNN type discriminator for pushing the generator to yield realistic results with regularized contrast and lower deformation of images. Moreover, we used cosine distance, a non-euclidean distance, and a weighted loss function to improve the perceptual quality of our output. Due to the unavailability of ground truth in real-world foggy images, we synthesized hazy images using a transmission map of the clear images for our dataset. On both synthetic and real-world foggy pictures, the results of our proposed model compete well with existing state-of-the-art techniques.**

***Index Terms*—Image Dehazing,Atmospheric Scattering Model, Generative Adversarial Networks, CNN Contrast Aware Channel Attention**

Fig. 1. Generated dehazed images from hazy images using our proposed model. Left: Input hazy images, Right: Output haze-free images

## I. Introduction

In the domain of image processing and enhancement, dehazing image is one of the most challenging and time-consuming jobs. The haziness in photos could be due to the rising pollution levels in metropolitan cities, which causes the presence of aerosol particles in the atmosphere, thus leading to the degradation of images' quality. Aerosol particles (fog, haze, mist, smoke, and so on) are tiny solids or drops of liquid floating in the environment or another gas. Image quality deteriorates with the presence of such aerosol particles, posing problems during surveillance and object detection.

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{1}$$

$$t(x) = e^{-\beta d(x)} \tag{2}$$

The renowned Atmospheric Scattering Model formula, proposed by Mccartney in 1977 [1], is given by (1), where I stands for the hazy picture, J for the ground truth image, A for ambient light, and t for the coefficient of transmission. It can be observed from (1) that transmission map(T) and atmospheric light(A) are the two essential parameters for generating a clear image(J) from the hazy input image(I). These two parameters lead to a broad division of two methods for dehazing images: prior and learning-based approaches. [2, 3]. With the medium of this paper, we propose the following characteristics of our model.

1. A learning-based method to dehaze single images using a complete and densely connected end-to-end encoder-decoder based generator (Fig. 4) [11] consisting of a sequence of dense blocks (Fig. 5) and transition blocks [14] for better-facilitating features extraction of images and their utilization, along with contrast-aware channel attention module [19] to generate natural appealing dehazed images with better resolution. To reduce the learning parameters and thus the computation time, we have used pooling layers after convolution and ReLU layers.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:52:43 UTC from IEEE Xplore. Restrictions apply.

The generator is pushed to provide more realistic dehazed outputs with regularised contrast and lesser distortion of pictures using a typical patch-based CNN-type discriminator [15].

2. Moreover, we used the weighted loss function and cosine-similarity, a non-euclidean distance [16] for measuring the perceptual loss between the dehazed output and the ground-truth target photos to make resultant images more natural and visually appealing.

3. It is observed that the training dataset has a significant influence on the results of the learning-based methods; therefore, we created a custom synthetic training and testing dataset consisting of a variety of indoor and outdoor scenes taken from NYU Depth [17] and OTS (Outdoor Training Set) by RESIDE-$\beta$[18] respectively to increase our model's accuracy and encompass the majority of hazy scenarios.

Fig. 2. Dark Channel Prior. Left: Input hazy image, Middle: Dehazed image using DCP, Right: Estimated transmission map

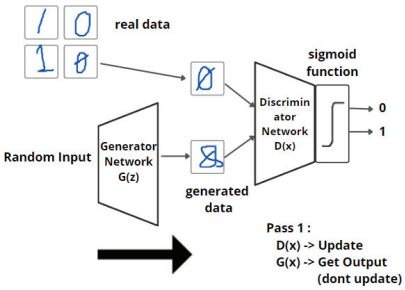



Fig. 3. Generative Adversarial Network

## II. Related Works

Initially, for dehazing images, prior-based methods were used. Dark Channel Prior(DCP) [4] produces the dehazed clear image with the contrast's lowest upper bound according to Koschmieder's law. **Disadvantage:** Prior-based dehazing methods cannot be generalized to all scenes as they dehaze images at the cost of introducing artifacts due to a lack of efficient and accurate priors.

Learning-based algorithms [6, 7] perform better than prior-based since the architecture of a neural network is enough for capturing a wide range of low-level image statistics before any learning. Learning algorithms tend to generate comparatively better and natural results. For example, AODNet [6] proposed using a CNN to generate haze-free images resulting in much better visual effects but at the cost of more haze appearing than the counterpart algorithms. **Disadvantage:** Since the learning methods are supervised learning-based, they require pair of foggy as well as the ground truth pictures of the same scene. Such training samples are difficult to achieve, and hence, their abilities are constrained due to inaccuracies in relations between dehazed image [8], atmospheric light, and transmission map.

Generative Adversarial Networks [9, 10] have contributed enormously to various image reconstruction and translation works [23], such as de-hazing [21] and low-light enhancement [22], and have shown remarkable results in generating realistic images [11]. GAN incorporates two neural architectures, Generator and Discriminator, competing with each other to become more accurate in their predictions. The Generator creates an output image from a random input image. This output image, together with the subsequent training set image, is used by the Discriminator to help categorize the picture as real or fake.

In [9], The problem of image dehazing is addressed by constructing a Network based on the conditional GAN. An Encoder and Decoder is introduced and developed in the architecture of the CGAN network to retain more detailed, informative, and dehazed images. A loss function containing pre-trained VGG features and a gradient prior, which is L1-regularized, is developed. H. Chen et al. Reference [10] have developed a network where the Atmospheric Scattering Model is employed, and CGAN is taken as a training framework. The ambient light value and transmission map values are estimated simultaneously rather than individually, and they are combined in image K to boost image detail.

In [13], the authors proposed that if we combine the prior-based and learning-based algorithms, we can achieve the benefits of both the visibility restoration methods alongside images appearing natural and not artificially manipulated. RefineD-Net Architecture proposed dividing the DHA into two subtasks: the Dark Channel Prior-based Model and CycleGAN backbone-based model. In [11], the authors have proposed a Fusion Discriminator in GANs for removing haze from images. The Fusion Discriminator uses frequency parameters as additional features, resulting in an outstanding performance on a simulated dataset encompassing a variety of indoor and outdoor photos.

## III. Proposed Model

In this section, we discuss the proposed densely connected end-to-end encoder-decoder-based Generative Adversarial Network capable of extensive dehazing and can compete with existing notable research methodologies. In the subsections following, the network's components, particularly the components of Generator and Discriminator along with Loss Functions, are addressed. Our model has three types of blocks -

### A. Dense Block

Dense Block contains a sequence of Batch Normalization, ReLU, Convolutional layers, and dropout layer. In convolutional neural networks, a Dense Block is a module that directly connects all layers (with matching feature-map sizes) to one



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:52:43 UTC from IEEE Xplore. Restrictions apply.

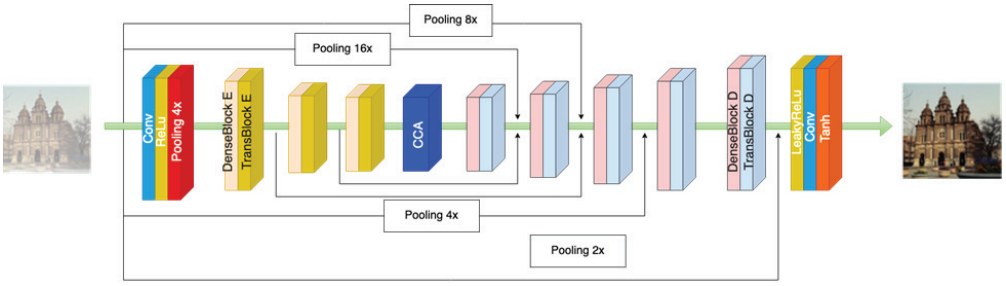



Fig. 4. Generator Architecture

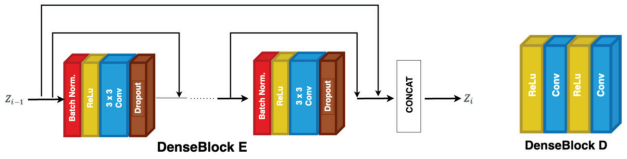



Fig. 5. DenseBlock Architecture

another. To retain the feed-forward nature through the dropout layer, each layer accepts extra inputs from all preceding levels and passes on its feature maps to all succeeding layers.

### B. Transition Block

The Transition Block is made up of batch normalization, ReLu, convolutional, and avg pooling. Concatenating feature maps of different sizes would be impractical. Therefore, the feature maps of each layer are the same size in each dense Block. Transition blocks assure the role of applying batch normalization using downsampling where the number of channels is reduced to half. In the decoder, a transition block (Transition Block D) is employed to upscale the images to restore the size of the feature maps progressively.

### C. Contrast Aware Channel Attention Module

We have used the Contrast Aware Channel Attention (CCA) Module proposed by Zheng Hui et al. [19] to enhance and utilize the features extracted by the dense blocks in the encoder module. Pooling can increase the PSNR (Point to Signal Noise Ratio). However, it ignores some information concerning perceptual tasks, including edge identification, picture segmentation, lightness calculation, and three-dimensional structure estimate. The CCA module helps us achieve and utilize this information to generate more natural, appealing, and contrast regularized dehazed images.

### D. Generator

The input image is fed to the encoder module of the Generator containing a series of 3 dense blocks and transition blacks (DenseBlock E and TransitionBlock E) from the pre-trained DenseNet121 [14] model responsible for extracting features and downscaling the input image, respectively. These layers are coupled with a CCA module to utilize and enhance the information extracted to generate results with better lightness computation, edges, and structure detection. Output from the encoder module after features extraction and utilization is fed to the decoder module consisting of 5 dense (DenseBlock D) and transition blocks (TransitionBlock D), in which the size of the features map is progressively recovered along with pooling to minimize the computational power and number of parameters to train, hence preventing overfitting as well. To make the model training faster and more accurate, output from the decoder block is passed through activation layers to increase its non-linearity.

### E. Discriminator

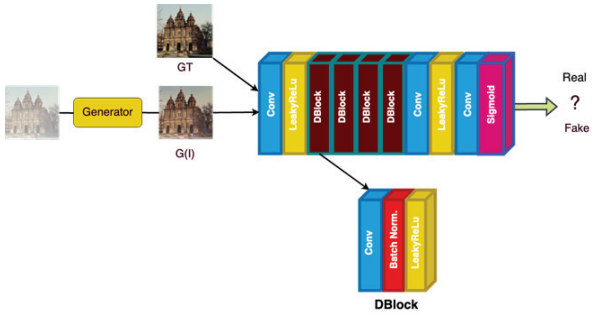



Fig. 6. Discriminator Architecture

The discriminator is a classifier based on a neural network and is used to classify images from generator as real or fake. As long as the categorization is accurate, the generator will try to deceive the discriminator by learning from it. To push the generator to create more realistic dehazed results with regularised contrast and lesser distortion of pictures, we employed a standard CNN patch-based discriminator [15] that only penalizes structure at the size of local image patches. Our discriminator consists of a four-layer stridden convolutional neural network with batch normalisation (except for the input layer) and leaky ReLU activations.LeakyReLU makes it simpler for gradients to flow across the architecture. Based on our training settings, we utilized a 14 × 14 patch size. Each 14 × 14 patches in the input image is classified as genuine or fake by our patch discriminator. This convolutionally scans the picture, averaging all answers to produce the discriminator's final output.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:52:43 UTC from IEEE Xplore. Restrictions apply.

### F. Loss Functions

*a) Discriminator Loss:* The Discriminator classifies the images from Generator as real or not by maximising the below function

$$Loss = \nabla_{\theta g}(1/m)\Sigma[logD(x_i) + log(1 - D(G(z_i)))] \quad (3)$$

where log(D(x)) is the probability of the generator correctly classifying the real picture. For the discriminator loss, we computed both the real image and fake image loss using BCE(Binary Cross Entropy) equation.

$$Loss = -(1/N)\Sigma(y_i log(y_i) + (1 - y_i) log(1 - y_i)) \quad (4)$$

*b) Reconstruction Loss:* Reconstruction loss examines the pixel-wise difference between the dehazed and clear image and determines how well the picture has been recreated from the input. In our model, we have used L1 loss as the primary loss function for computing the Reconstruction Loss. The generator's output is $G(I_i)$, and the ground truth is $J_i$, given an input hazy image $I_i$. Then, across N samples, the L1 norm loss can be represented as:

$$L_1 = \Sigma \|G(I_i) - J_i\| \quad (5)$$

*c) Perceptual Loss:* When comparing two identical pictures, such as the same photo but displaced by one pixel, perceptual loss functions are utilized to deliver more accurate style transfer findings. We have used the cosine distance as the perceptual loss function in our model. Even if two comparable images are separated by the Euclidean distance (according to the characteristics of the images), they are likely to be oriented closer together in the vector space owing to cosine similarity.

The cosine similarity is given by -

$$\cos(x, y) = \frac{x.y}{||x||.||y||} \quad (6)$$

and cosine distance, a function of cosine similarity is given as

$$1 - \cos(x, y) = 1 - \frac{x.y}{||x||.||y||} \quad (7)$$

where x and y represent the features map of the output (dehazed) image and target (clear) image respectively, extracted from pre-trained VGG-19 model.

$$L_2 = \cos(a, b) \quad (8)$$

where a is the output dehazed image and b is the clear target image from VGG.

*d) GAN loss:* The generator loss is calculated depending upon the conditions; if it successfully tricks the discriminator, it is rewarded; otherwise, it is penalized. In our model, we have used BCE loss as the primary function for computing the GAN loss. To train the generator, the following equation is used:

$$L_3 = \nabla_{\theta g}(1/m)\Sigma[log(1 - D(G(z^i)))] \quad (9)$$

We calculated the total loss , L, as a weighted loss function is given by

$$L = L1 + W2 * L2 + W3 * L3 \quad (10)$$

where, W2 and W3 are the perceptual loss and GAN loss weights, respectively.

## IV. Methodology

Fig. 7. Synthesized hazy images using transmission map of ground truth image. Left: Clear image, Middle: Transmission map, Right: Synthesized hazy image

### A. Dataset

Since there is no publicly available dataset for image dehazing, we have manually synthesized fog/haze into available clear indoor and outdoor image dataset using their depth maps. We have used the following publicly available datasets:

1) **Indoor:** NYU Depth Indoor Dataset [17] containing 1449 indoor images along with their transmission map depth information.
2) **Outdoor:** OTS (Outdoor Training Set) by RESIDE-β[18] is a standard benchmark dataset which contains outdoor images (Outdoor Training Dataset) along with their transmission map depth information from which we filtered out 1500 images that covered the majority of hazy scenarios.

We used depth information of the images to generate hazy images from them using the atmospheric scattering model (1) where the ambient light factor was set to 1, transmission map (t) was generated using one minus normalized depth. Moreover, we randomly scaled the transmission map for a given set of clear images to generate haziness of different intensities. A total of 2949 clear/hazy pairs of images constituted our custom synthetic dataset.

### B. Training Details

We trained our model on an Intel Xeon Silver 4114 CPU @2.20GHz with Titan RTX GPU and 128 GB RAM. We started with a 1e-5 learning rate for the generator and discriminator and then stepped down the rate by 2 on every 100 epochs completed. The value of the perceptual loss weight $W_{per}$=0.5,



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:52:43 UTC from IEEE Xplore. Restrictions apply.

TABLE I
QUANTITATIVE COMPARISON OF OUR MODEL WITH DIFFERENT EXISTING NOTABLE METHODS ON THE BASIS OF SSIM AND PSNR METRICS

| Metrics | Our Dataset | | | | SOTS | | | |
|---|---|---|---|---|---|---|---|---|
| | He et. al [4] | Q.Zhu et. al [5] | Meng et. al [12] | **Ours** | He et. al [4] | Q.Zhu et. al [5] | Meng et. al [12] | **Ours** |
| **PSNR** | 14.20 | 13.35 | 11.69 | **23.90** | 18.64 | 18.69 | 16.14 | **22.90** |
| **SSIM** | 0.79 | 0.76 | 0.68 | **0.94** | 0.84 | 0.85 | 0.78 | **0.92** |
| **Score** | 1.50 | 1.43 | 1.27 | **2.13** | 1.77 | 1.78 | 1.59 | **2.05** |

*The above results were obtained by testing the models against our custom and common SOTS dataset.

Fig. 8. Visual comparison between different notable dehazing methods on real world hazy images

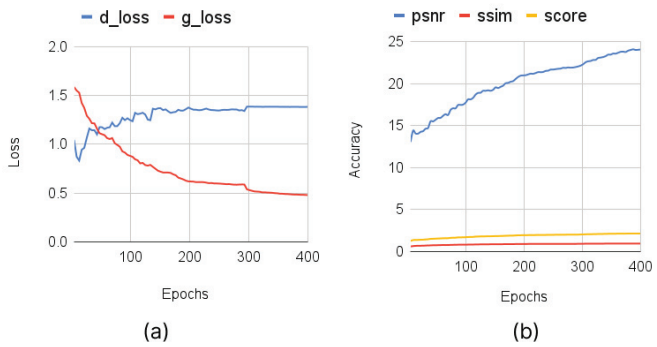



Fig. 9. (a) Relationship in d_loss and g_loss with each epoch in training. (b) Relationship in psnr, ssim and score with each epoch in training

and GAN loss weight $W_{gan}$=0.5, were set to gain optimal results. Adam optimizer was used to modify the attributes of our network on each step. Our custom dataset consisted of 2500 training, 249 validation, and 200 testing pair of clear/hazy images. To increase the computation speed, we scaled the training and validation dataset to 512 x 512 and image batch size to 8. Due to small size of our dataset and to improve our model efficiency and accuracy on various intensities of light and haziness, on each iteration, we augmented the pair of images by flipping them horizontally and vertically.

### C. Performance Metrics

The SSIM and PSNR, two extensively used metrics, are two examples of measurements that may be used to assess the effectiveness of a haze extraction technique. PSNR (Peak Signal to Noise Ratio) is a statistic for determining how well an algorithm can eliminate noise from an image. The Structural Similarity Index Measure (SSIM) is a tool that compares two pictures to determine how similar they are. Both PSNR and SSIM measurements alone do not guarantee visually acceptable outcomes; therefore, measures that combine the two are essential. To create a composite score measure, we employed the weighted sum of the image's PSNR and SSIM values, as described by [20].

$$Score = W_{PSNR} * PSNR + W_{SSIM} * SSIM \qquad (11)$$

Here the value of $W_{SSIM}$ was set as 1 while the value of $W_{PSNR}$ was equal to 0.05 [20]. During the training phase, we



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:52:43 UTC from IEEE Xplore. Restrictions apply.

continuously stored the evaluation metrics (PSNR, SSIM) and losses values using which we stored the model checkpoints.

### D. Qualitative and Quantitative Evaluation

By comparing the SSIM, PSNR, and our overall score metric generated by evaluating the approaches on our custom test and SOTS (Synthetic Objective Testing Set) dataset, we have offered a quantitative and visual comparison of our model to other significant dehazing procedures in this section. The following models were chosen for our quantitative analysis:

1) Dark channel prior (He et. al [4])
2) AOD-Net (B. Li et al [6])
3) Boundary constraint and context regularization (Meng et. al [12])
4) Color attenuation prior (Q.Zhu et. al [5])

As we can observe from Table I, our model performs at par with the above-stated methods quantitatively and from Fig. 8 the qualitative results are found to be naturally appealing, with regularized contrast, and better edges and structure detection.

## V. CONCLUSION

Our paper developed a generative adversarial network based on encoder-decoders that takes a hazy input image and produces a clear dehazed output image.. We made use of a connected encoder-decoder based generator for facilitating features extraction along with the CCA module for enhancement of the extracted information. Using the patch-based discriminator, and cosine distance for calculating the perceptual loss, we significantly reduced the presence of artifacts and generated more visually appealing natural results with better lightness computation, edges, and structure detection. Our model was quantitatively and visually comparable to other common dehazing algorithms, demonstrating that the proposed model performs better than other novel methods for dehazing image. In addition, to cover a variety of situations, we generated a custom synthetic dataset. It can be further modified by training the model on various other datasets, such as O-HAZE, I-HAZE and NTIRE to broaden its capabilities and generate better results on non-homogeneous and densely foggy pictures. However, our technique yields dehazed results that are more aesthetically appealing and have less color distortion. On both synthetically created datasets as well as real-world foggy photos, comprehensive experimental findings have shown that the approach surpasses several state-of-the-art techniques.

## REFERENCES


[1] A. Cantor, "Optics of the atmosphere–Scattering by molecules and particles," in IEEE Journal of Quantum Electronics, vol. 14, no. 9, pp. 698-699, September 1978, doi: 10.1109/JQE.1978.1069864.

[2] A. S. Parihar, Y. K. Gupta, Y. Singodia, V. Singh and K. Singh, "A Comparative Study of Image Dehazing Algorithms," 2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 766-771, doi: 10.1109/ICCES48766.2020.9138037.

[3] A. S. Parihar and G. Gupta, "Prior based Single Image Dehazing using Decision Image," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 960-965, doi: 10.1109/ICECA49313.2020.9297597.

[4] He, Kaiming & Sun, Jian & Tang, Xiaoou. (2011). Single Image Haze Removal Using Dark Channel Prior.. IEEE Trans. Pattern Anal. Mach. Intell.. 33. 2341-2353. 10.1109/CVPRW.2009.5206515.

[5] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," IEEE Transactions on image processing, vol. 24, no. 11, pp. 3522–3533, 2015.

[6] B. Li, X. Peng, Z. Wang, J. Xu and D. Feng, "AOD-Net: All-in-One Dehazing Network," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4780-4788, doi: 10.1109/ICCV.2017.511.

[7] B. Cai, X. Xu, K. Jia, C. Qing and D. Tao, "DehazeNet: An End-to-End System for Single Image Haze Removal," in IEEE Transactions on Image Processing, vol. 25, no. 11, pp. 5187-5198, Nov. 2016, doi: 10.1109/TIP.2016.2598681.

[8] Jisnu, K & Meena, Gaurav. (2020). Image DeHazing Using Deep Learning Techniques Ravi Raj Choudhary. Procedia Computer Science. 167. 1110-1119. 10.1016/j.procs.2020.03.413.

[9] R. Li, J. Pan, Z. Li and J. Tang, "Single Image Dehazing via Conditional Generative Adversarial Network," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8202-8211, doi: 10.1109/CVPR.2018.00856.

[10] K. Gan, J. Zhao and H. Chen, "Multilevel Image Dehazing Algorithm Using Conditional Generative Adversarial Networks," in IEEE Access, vol. 8, pp. 55221-55229, 2020, doi: 10.1109/ACCESS.2020.2981944.

[11] Dong, Yu, Yihao Liu, He Zhang, Shifeng Chen, and Yu Qiao. "FD-GAN: Generative adversarial networks with fusion-discriminator for single image dehazing." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 10729-10736. 2020.

[12] Meng, Gaofeng & Wang, Ying & Duan, Jiangyong & Xiang, Shiming & Pan, Chunhong. (2013). Efficient Image Dehazing with Boundary Constraint and Contextual Regularization. IEEE International Conference on Computer Vision (ICCV). 617-624. 10.1109/ICCV.2013.82.

[13] S. Zhao, L. Zhang, Y. Shen and Y. Zhou, "RefineDNet: A Weakly Supervised Refinement Framework for Single Image Dehazing," in IEEE Transactions on Image Processing, vol. 30, pp. 3391-3404, 2021, doi: 10.1109/TIP.2021.3060873.

[14] Huang, Gao, et al. "Densely Connected Convolutional Networks." ArXiv:1608.06993 [Cs], Jan. 2018. arXiv.org, http://arxiv.org/abs/1608.06993.

[15] P. Isola, J. Zhu, T. Zhou and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5967-5976, doi: 10.1109/CVPR.2017.632.

[16] A. R. Lahitani, A. E. Permanasari and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," 2016 4th International Conference on Cyber and IT Service Management, 2016, pp. 1-6, doi: 10.1109/CITSM.2016.7577578.

[17] Silberman, N., Hoiem, D., Kohli, P., Fergus, R. (2012). Indoor Segmentation and Support Inference from RGBD Images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds) Computer Vision – ECCV 2012. ECCV 2012. Lecture Notes in Computer Science, vol 7576. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33715-4_54

[18] B. Li et al., "Benchmarking Single-Image Dehazing and Beyond," in IEEE Transactions on Image Processing, vol. 28, no. 1, pp. 492-505, Jan. 2019, doi: 10.1109/TIP.2018.2867951.

[19] Hui, Z., Gao, X., Yang, Y. and Wang, X., 2019, October. Lightweight image super-resolution with information multi-distillation network. In Proceedings of the 27th acm international conference on multimedia (pp. 2024-2032).

[20] N. , Bharath Raj, and Venkateswaran N. "Single Image Haze Removal Using a Generative Adversarial Network." ArXiv:1810.09479 [Cs, Eess], 2020. arXiv.org, http://arxiv.org/abs/1810.09479.

[21] K. Singh, A. S. Parihar, "Variational optimization based single image dehazing", Journal of Visual Communication and Image Representation, Vol. 79, 2021.

[22] Parihar, A.S., Singh, K, Rohilla, H, Asnani, G. Fusion-based simultaneous estimation of reflectance and illumination for low-light image enhancement. IET Image Process. 2021; 15: 1410– 1423.

[23] Bhowmik, A., Kumar, S. & Bhat, N. Evolution of automatic visual description techniques-a methodological survey. Multimed Tools Appl 80, 28015–28059 (2021).




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:52:43 UTC from IEEE Xplore. Restrictions apply.

arXiv:2209.09651v1 [cs.LG] 18 Sep 2022

# DEEP CONVOLUTIONAL ARCHITECTURES FOR EXTRAPOLATIVE FORECASTS IN TIME-DEPENDENT FLOW PROBLEMS

**Pratyush Bhatt, Yash Kumar**
Department of mechanical engineering
Delhi Technological University
P4X9+Q8X, Bawana Rd, Shahbad Daulatpur Village, Rohini, New Delhi, 110042, Delhi, India
bhattpratyush906@gmail.com, yashk8481@gmail.com

**Azzeddine Soulaïmani**
Department of mechanical engineering
École de technologie supérieure
1100 Notre-Dame St W, Montreal, Quebec H3C1K3, Canada
Azzeddine.Soulaimani@etsmtl.ca

**ABSTRACT**

Physical systems whose dynamics are governed by partial differential equations (PDEs) find applications in numerous fields, from engineering design to weather forecasting. The process of obtaining the solution from such PDEs may be computationally expensive for large-scale and parameterized problems. In this work, deep learning techniques developed especially for time-series forecasts, such as LSTM and TCN, or for spatial-feature extraction such as CNN, are employed to model the system dynamics for advection dominated problems. These models take as input a sequence of high-fidelity vector solutions for consecutive time-steps obtained from the PDEs and forecast the solutions for the subsequent time-steps using auto-regression; thereby reducing the computation time and power needed to obtain such high-fidelity solutions. The models are tested on numerical benchmarks (1D Burgers' equation and Stoker's dam break problem) to assess the long-term prediction accuracy, even outside the training domain (extrapolation). Non-intrusive reduced-order modelling techniques such as deep auto-encoder networks are utilized to compress the high-fidelity snapshots before feeding them as input to the forecasting models in order to reduce the complexity and the required computations in the online and offline stages. Deep ensembles are employed to perform uncertainty quantification of the forecasting models, which provides information about the variance of the predictions as a result of the epistemic uncertainties.

***Keywords*** non-intrusive reduced-order modelling · deep autoencoders · LSTM · TCN · CNN · time-dependent flow problems · deep ensembles

## 1 Introduction

Efficient numerical simulations of complex dynamical systems are needed in order to seek solutions at different times or parameter instances, especially in fluid dynamics. These systems are typically described by a set of parameterized nonlinear partial differential equations (PDEs). Obtaining numerical solutions using a high-fidelity (finite element, finite volume, or finite difference type) computational solver may be extremely expensive, as they must create high-dimensional renderings of the solution to precisely resolve the spatial-temporal multifolds and inherent non-linearities. This method thus becomes inefficient for applications such as optimization and uncertainty quantification, where numerous simulations are required to arrive at the desired solution. Reduced-order models (ROMs) are suitable substitutions for computationally-expensive numerical solvers, as these methods generate a low-ranked structure of the high-dimensional snapshots, which are then utilized to model the spatio-temporal dynamics of the PDE system. Among the various ROM techniques that have been developed, projection-based ROMs are the type employed most

extensively. The method involves the generation of a reduced set of basis functions or modes such that their linear superposition effectively overlaps a low-rank approximation of the solutions. Proper Orthogonal Decomposition (POD) is the most popular method among the reduced basis class. POD utilizes singular value decomposition (SVD) to generate an empirical basis of dominant, orthonormal modes to obtain an optimum linear subspace in which to project the system-governing PDEs [1, 2]. Availability of the governing equations is necessary to employ intrusive ROM techniques such as the Galerkin projection [3], or the Petrov–Galerkin projection [4], which produce an interpretable ROM defined by high-energy or dominant modes. However, scenarios where the governing equations are unavailable require the application of data-driven methods, such as non-intrusive ROM (NIROM)[5, 6]. In a NIROM, the expansion coefficients for the reduced solution are obtained via interpolation on the reduced basis space spanned by the set of dominant modes. However, since the reduced dynamics generally belong to nonlinear, matrix manifolds, a variety of interpolation and regression methods have been proposed, capable of enforcing the constraints characterizing those manifolds. Some of the methods most often employed are dynamic mode decomposition [7, 8, 9], radial basis function interpolation [10, 11] and Gaussian process regression [12, 13]. The recent advancements in machine learning (ML) methods[14] have given rise to revolutionary approaches that effectively evaluate and expedite existing numerical models or solvers by using online-offline computational stages. In the offline stage, the ML model updates its weights or coefficients (training) to learn the system dynamics by using the high-fidelity solutions obtained by the numerical solver, hence requiring computational power and time. In the online stage, the model uses the pre-computed/optimized weights (from the training) to obtain the solution (prediction) for a new set of input instances, and does so almost instantly with minimal computational cost. Various data-driven ML-based frameworks have been proposed to model the propagation of system dynamics in latent space. Some of the more highly successful examples involve the use of deep neural networks (DNNs) [15], long-short-term memory (LSTM) networks [16, 17, 18, 19] , neural ordinary differential equations (NODE) [20, 21, 19], and temporal convolutional networks (TCNs) [22, 23].

Significant work has been carried out recently on predicting solution instances outside the training domain for a variety of fluid problems with discontinuities, wave propagation, and advection-dominated flows. Liu et al [24] presented a predictive data assimilation framework based on the Ensemble Kalman Filter (EnKF) and the DDROM model, which uses an autoencoder network for the compression of high-dimensional dynamics to lower dimensional space and then the LSTM method to model the fluid dynamics in latent space. The model capabilities were estimated using 2D Burgers' equation and flow past a cylinder test case. Maulik et al. [25] proposed a Convolutional Autoencoder (CAE) for compression and a recurrent LSTM network for time evolution on the reduced space. The CAE-LSTM model was capable of reconstructing the sharp profile of the advecting Burgers' equation more accurately than the POD-Galerkin technique. Dutta et al. [18] utilized an advection-aware (AA) autoencoder network that learns nonlinear embeddings of the high-fidelity system snapshots using an arbitrary snapshot from the dataset, and then models the latent space dynamics using an LSTM network to make predictions for the linear advection and Burgers' problem. Cheng et al. [26] used the POD-ANN model, in which they performed a priori dimension reduction on the high-fidelity dataset and parameterization with an artificial neural (ANN) network to solve the strongly non-linear Allen-Cahn equations and the cylinder flow problem. Heaney et al.[27] proposed an AI-DDNIROM framework, capable of making predictions for spatial domains, significantly larger than the training domain, using a domain decomposition approach, an autoencoder network for low-rank representation, and an adversarial network for making the predictions for flow past a cylinder and slug flow problems. Fatone et al. [19] introduced a μt-POD-LSTM ROM framework that is capable of extrapolation for time windows around 15% those of the training domain on unsteady advection-diffusion and unsteady Navier-Stokes equation for new parameter instances. Xu et al. [23] proposed a multi-level framework comprising a convolution autoencoder (CAE), a temporal CAE (TCAE) and a multilayer perceptron (MLP), for the purpose of parameterization, and a TCN network for auto-regressive future state predictions, and evaluated the results on problems such as Sod's-shock tube and transient ship airwakes. Wu et al. [22] developed a POD and TCN-based neural network for making predictions on the viscous periodic flow past a cylinder case. Abdedou et.al [28] proposed two CAE architectures to compress the high-dimensional snapshot matrices obtained from numerical solvers for the Burgers', Stoker's, and dam-break equations in space and time, and performed parameterization on the compressed latent space. Jacquier et al. [29] employed uncertainty quantification methods - Deep Ensembles and Variational Inference-based Bayesian Neural Networks on the POD-ANN order-reduction method to perform predictions within and outside of the training domain on problems such as shallow water equations for flood prediction, and generated probabilistic flooding maps aware of model uncertainty. Geneva et al. [30] presented a physics-constrained Bayesian auto-regressive CAE network that models non-linear dynamical systems (Kuramoto-Sivashinsky equation, 1D Burgers', 2D Burgers') devoid of training data, using only the initial conditions. This reduces the computation cost tremendously and provides uncertainty quantification at each time-step.

The caveat that remains is long-term temporal extrapolation for fluid problems marked by sharp gradients and discontinuities. Our study explores forecasting convolutional architectures (LSTM, TCN, and CNN) to obtain accurate solutions for time-steps distant from the training domain, on advection-dominated test cases. The high-dimensional input snapshots matrix is first compressed in space to obtain the reduced latent vectors before they are passed as a sequence

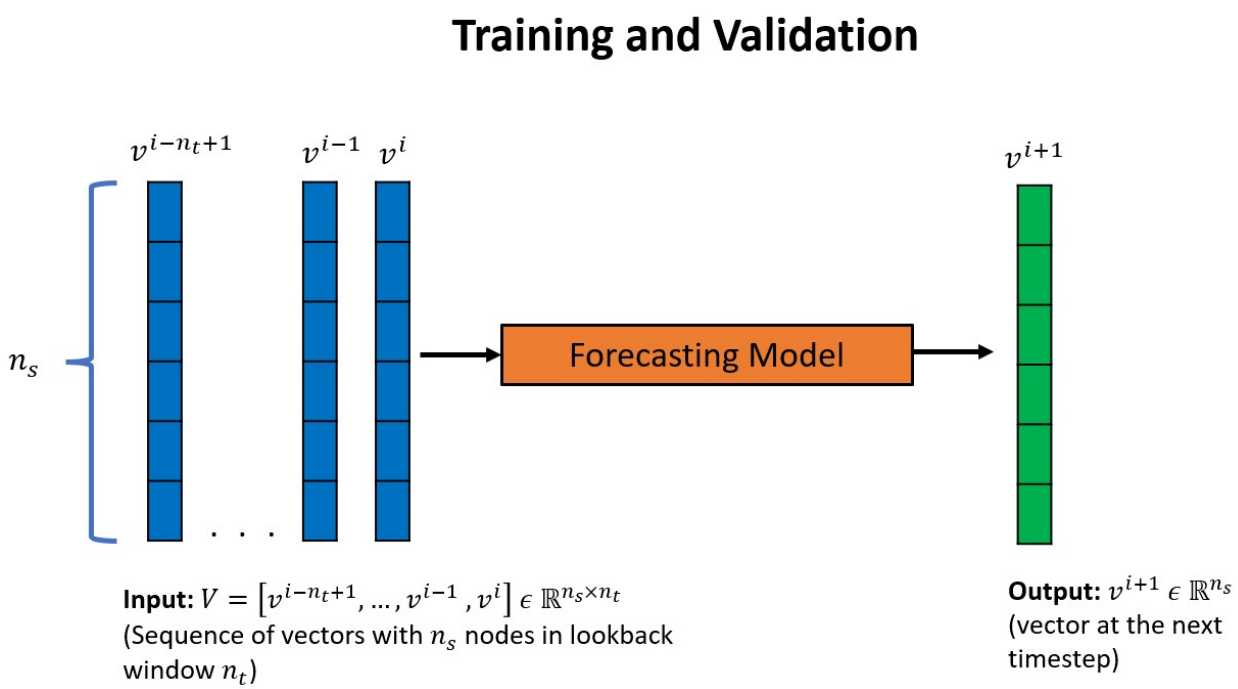



Figure 1: Training and validation method

to the forecasting models. Two types of architectures are first evaluated for space compression - MLP autoencoder and CAE autoencoder, to identify the one that is more accurate in terms of the reconstruction and preservation of the input information. A simple convolutional architecture is then proposed and shown to provide accurate results for the forecasts. To evaluate the epistemic uncertainties in the solutions, the methodology of deep ensembles is adopted.

The subsequent sections of the paper are organized as follows. Section 2 describes the dataset structure along with the training and testing strategies, followed by a presentation of the autoencoders for space compression and the forecasting convolutional architectures. In Section 3, the models are tested on two numerical cases - one-dimensional Burgers' and Stoker's equations, which are representative of advection-dominated flows. Finally, section 4 presents a summary of the results obtained by the models, and some concluding remarks.

## 2 Methodology

### 2.1 Dataset

The dataset is comprised of $T$ solution vectors/snapshots: $v^i$ with $n_s$ nodes ($v^i \in \mathbb{R}^{n_s}$) at time-steps $i \in \{1, 2, \ldots, T\}$ obtained using a high-fidelity PDE solver. For the autoencoder models, the output is the reconstruction of the input, therefore the training and validation input and output data are snapshot vectors $v^i$. For the forecasting models (Figure 1), N samples are used for training; in each sample, the input is a sequence of $n_t$ snapshots (lookback window = $n_t$): $V = [v^{i-n_t+1}, \ldots, v^{i-1}, v^i]$, with $V \in \mathbb{R}^{n_s \times n_t}$, and the corresponding output is the vector at the time-step immediately after the sequence end - $v^{i+1} \in \mathbb{R}^{n_s}$. For extrapolative testing (Figure 2), a sequence of $n_t$ vectors from the start of the dataset, $V = [v^1, \ldots, v^{n_t-1}, v^{n_t}] \in \mathbb{R}^{n_s \times n_t}$ is fed to the model to produce the vectors at all the subsequent time-steps: $[v^{n_t+1}, v^{n_t+2}, \ldots, v^T] \in \mathbb{R}^{n_s \times (T-n_t)}$ in an auto-regressive manner, i.e, first only a single subsequent snapshot $v^{n_t+1}$ is predicted, which is then concatenated with previous $n_t - 1$ vectors and passed to the forecasting model to produce vector $v^{n_t+2}$. This process is repeated in accordance with the desired number of subsequent solution vectors.

### 2.2 Non-intrusive reduced-order modelling

Non-intrusive ROMs (NIROMs) bypass the governing equations and utilize the full-order model solutions to develop a data-driven model, which compresses the full order data (snapshot) into a reduced-order (latent) space. The method most widely adopted to perform this utilizes deep neural network architectures called autoencoders [31].

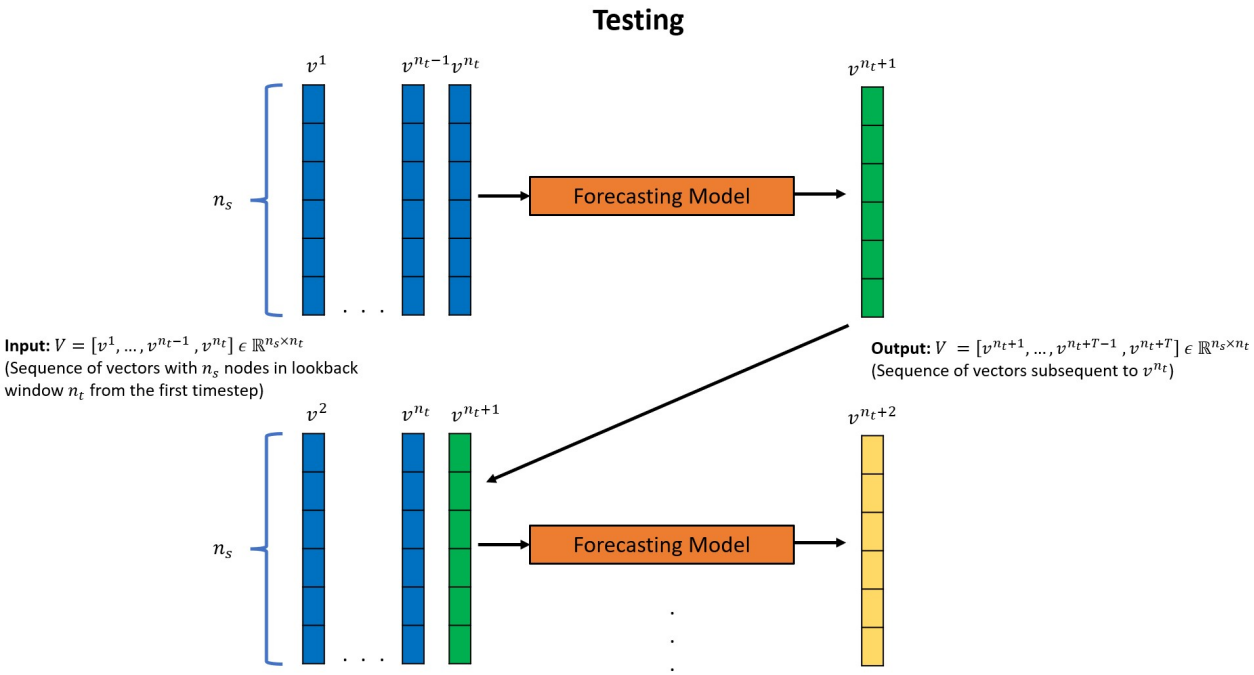



Figure 2: Autoregressive testing method for forecasting models

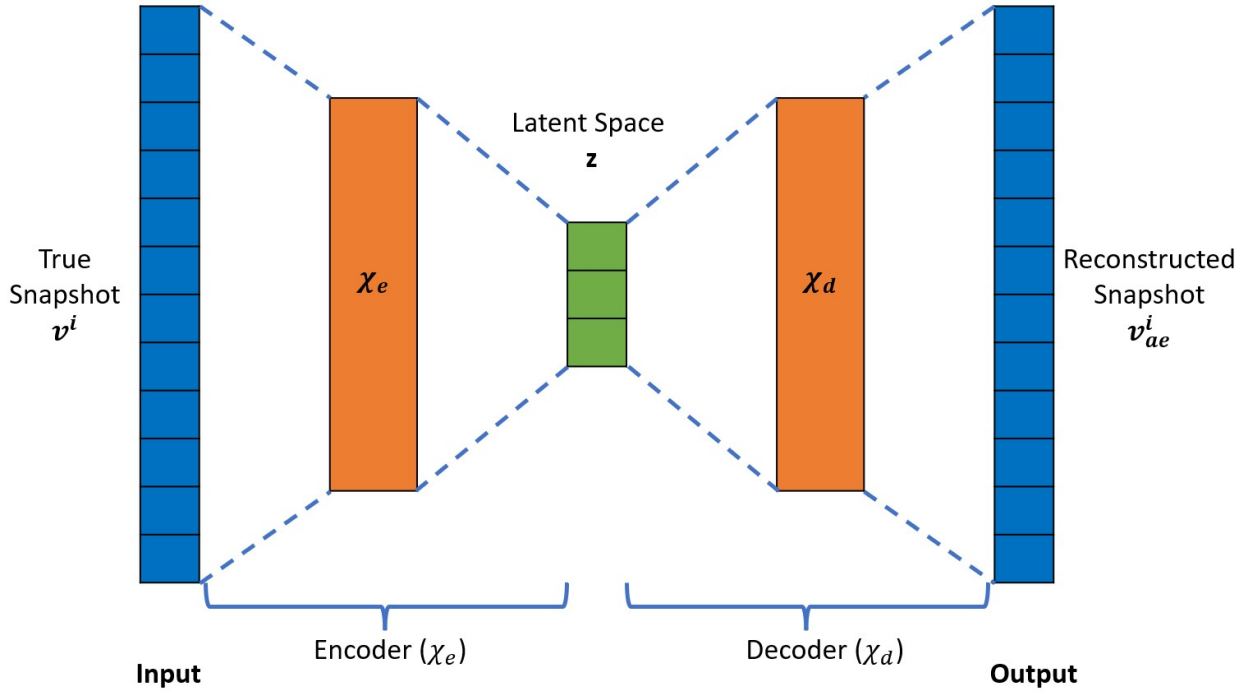



Figure 3: Autoencoder architecture

An autoencoder learns the approximation of the identity mapping, $\chi$: $v^i \rightarrow v^i_{ae}$ such that $v^i \approx v^i_{ae}$ and $\chi : \mathbb{R}^{n_s} \rightarrow \mathbb{R}^{n_s}$, where $n_s$ is the number of nodes in the solution vector $v^i$. This process is accomplished using a two-part architecture. The first part of the autoencoder network is the encoder $\chi_e$, which maps a high-dimensional input vector $v^i$ to a low-dimensional latent vector $z^i$: $z^i = \chi_e(v^i; \theta_e)$ and $z^i \in \mathbb{R}^m$ $(m \ll n)$. The second part is called a decoder, $\chi_d$,

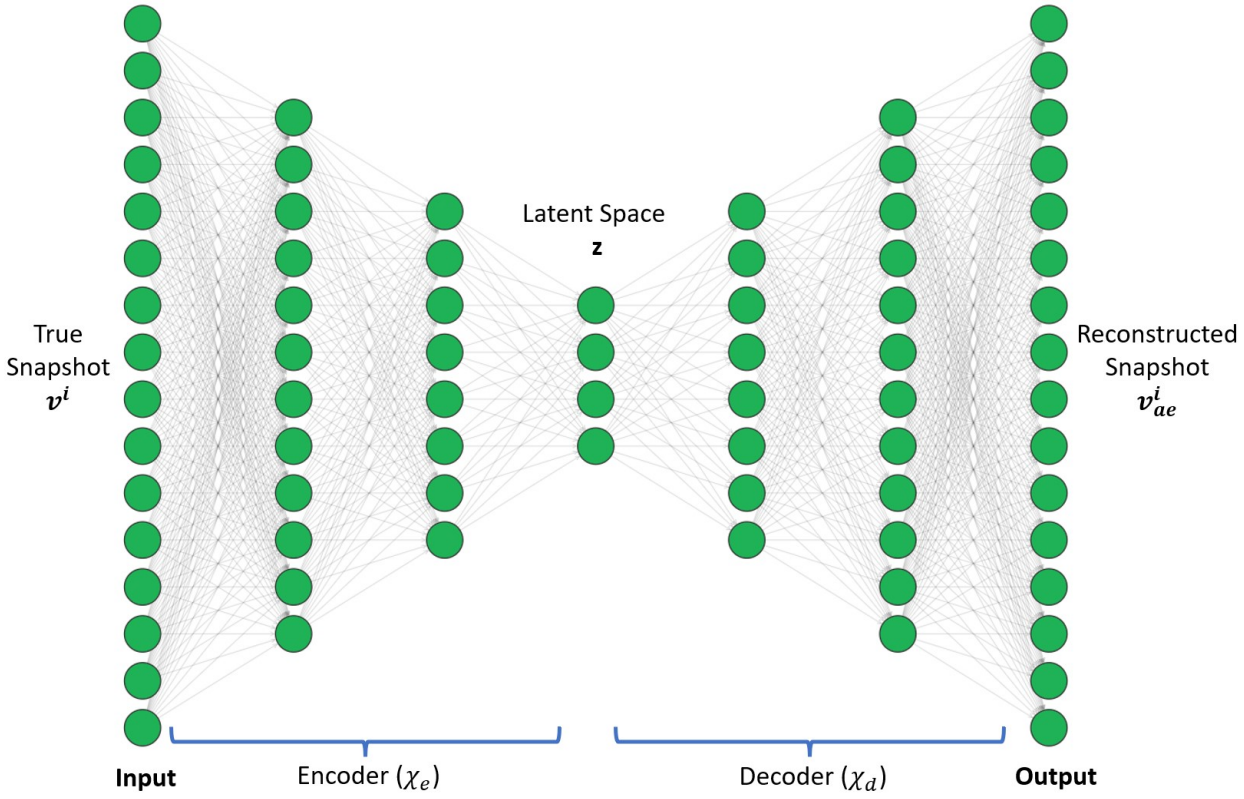



Figure 4: MLP Autoencoder architecture

which maps the latent vector $z^i$ to an approximation $v^i_{ae}$ of the high-dimensional input vector $v^i$: $v^i_{ae} = \chi_d(z^i; \theta_d)$. The combination of these two parts yields an autoencoder network (Figure 3) of the form $\chi : v^i \rightarrow \chi_d \circ \chi_e(v^i)$. The autoencoder model is trained by computing optimal values of the parameters ($\theta_e$, $\theta_d$) that minimize the reconstruction error over all the training data [18]:

$$\theta_e, \theta_d = argmin\mathcal{L}(v^i, v^i_{ae}) \tag{1}$$

where $\mathcal{L}(v^i, v^i_{ae})$ is a chosen measure of discrepancy between $v^i$ and its approximation $v^i_{ae}$. The restriction ($\dim(z^i) = m$) $\ll$ ($n = \dim(v^i)$) forces the autoencoder model to learn the salient features of the input data via compression into a low-dimensional space and to then reconstruct the input, instead of directly learning the identity function. Autoencoder architectures are generally comprised of MLPs (called AAs) [18], convolutional neural network autoencoders (called CAEs) [25, 23, 28], or a combination of both. While small-sized problems can be effectively modelled via an MLP architecture, problems involving data of high spatial complexity require CAE autoencoders for effective and accelerated spatial compression. The architecture of an MLP autoencoder, with two fully connected dense layers (hidden layers) in the encoder network and a mirrored decoder network, is shown in (Figure 4). The Convolution autoencoder consists of two convolution layers, each followed by batch normalization, swish activation, and an average pooling layer, as described in (Figure 5).

### 2.3 Forecasting Techniques

The dataset (Section 2.1) post compression by the encoder ($\chi_e$) produces $N$ samples of the form: $Z = [z^{i-n_t+1}, \ldots, z^{i-1}, z^i] \epsilon \mathbb{R}^{m \times n_t}$, $z^{i+1} \epsilon \mathbb{R}^m$, which are used to train the following forecasting models.

#### 2.3.1 Long Short-Term Memory (LSTM)

LSTM [32] is a special type of recurrent neural network (RNN) that is well-suited for performing regression tasks based on time series data. The main difference between the traditional RNN and the LSTM architecture is the capability of an LSTM memory cell to retain information over time and an internal gating mechanism that regulates the flow of

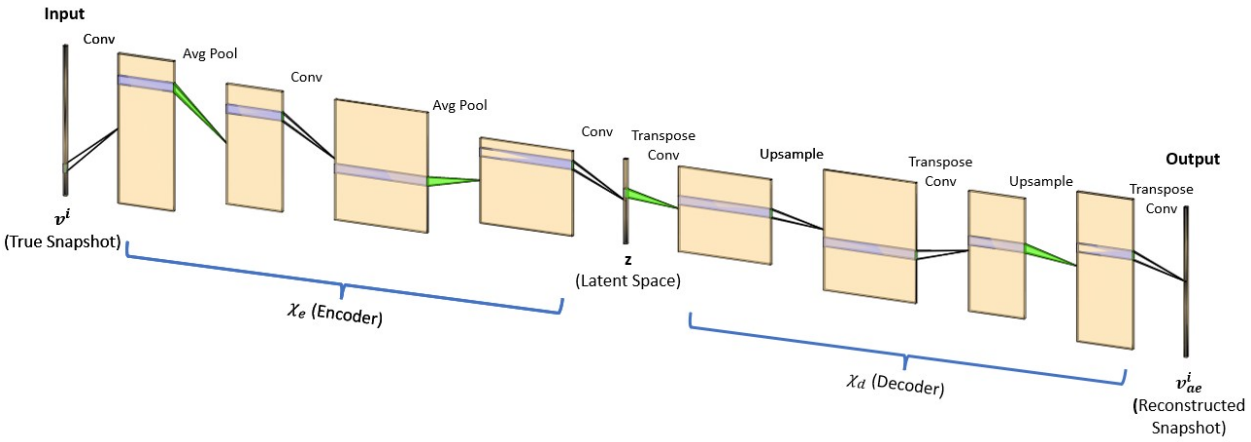



Figure 5: Convolutional Autoencoder architecture

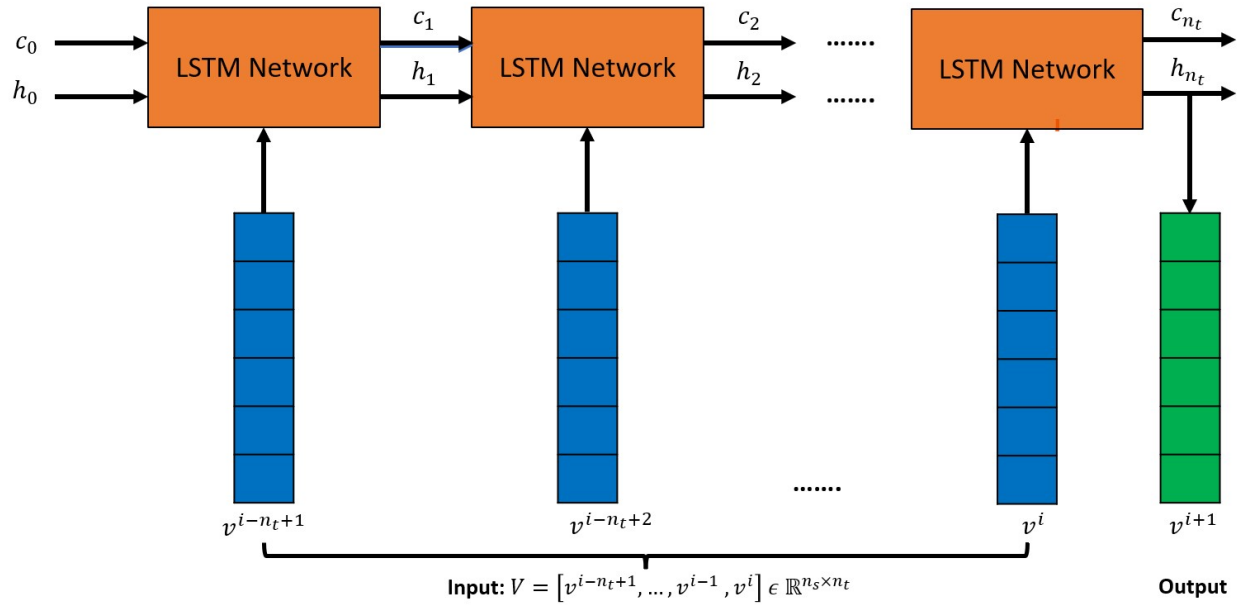



Figure 6: LSTM recurrence on vector sequence

information in and out of the memory cell [33]. The LSTM cell consists of three parts, also known as gates, that have specific functions. The first part, called the forget gate, chooses whether the information from the previous step in the sequence is to be remembered or can be forgotten. The second part, called the input gate, tries to learn new information from the current input to this cell. The third and final part, called the output gate, passes the updated information from the current step to the next step in the sequence. The basic LSTM equations for an input vector $v^i$ are:

$$input\,gate : \zeta_{in} = \alpha_s \circ F_{in}(v^i) \tag{2}$$

$$forget\,gate : \zeta_{for} = \alpha_s \circ F_{for}(v^i) \tag{3}$$

$$cell\,state : c_i = \zeta_{for} \odot c_{i-1} + \zeta_{in} \odot (\alpha_t \circ F_a(v^i)) \tag{4}$$

$$output\,gate : \zeta_{out} = \alpha_s \circ F_{out}(v^i) \tag{5}$$

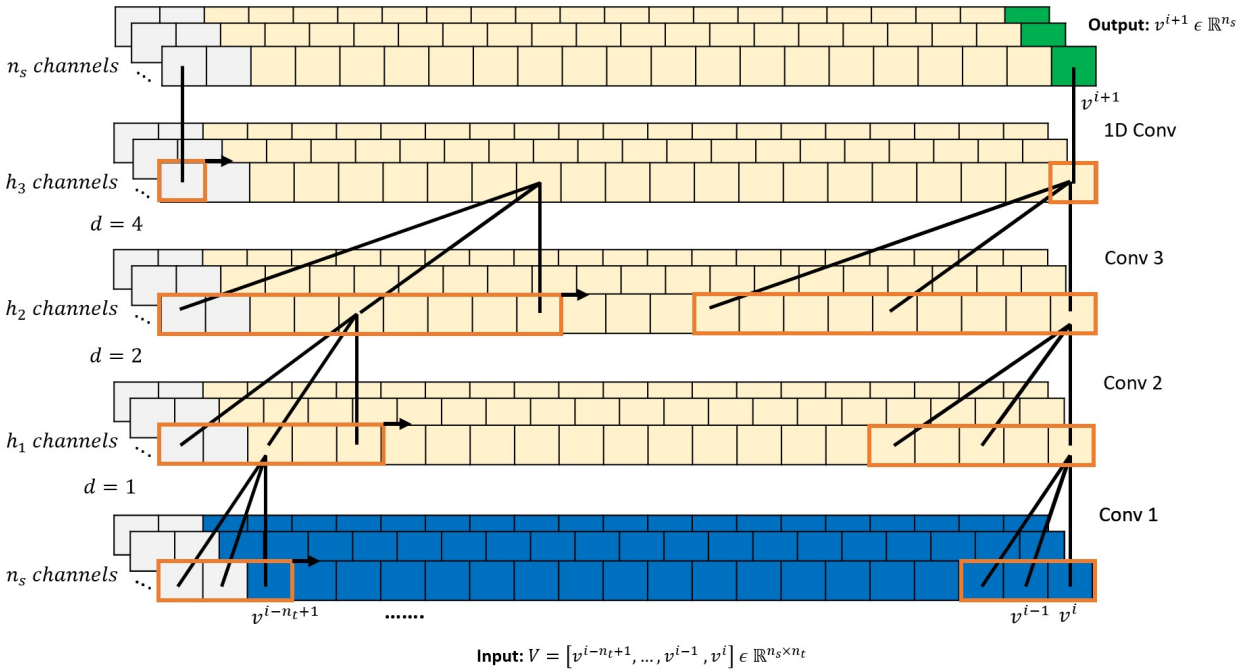



Figure 7: 1D dilated filters convolving on the temporal dimension of vectors

$$output : h_i = \zeta_{out} \circ \alpha_t(c_i) \tag{6}$$

Here, $F$ refers to a linear transformation defined by a matrix multiplication and bias addition, that is, $F(v^i) = Wv^i + b$, where W $\epsilon \mathbb{R}^{h \times n_s}$ is a matrix of layer weights ($h$ is number of neurons in the LSTM cell), b $\epsilon \mathbb{R}^h$ is a vector of bias values, and $v^i \epsilon \mathbb{R}^{n_s}$ is the input vector to the LSTM Cell. Also, $\alpha_s$ and $\alpha_t$ denote sigmoid and hyperbolic tangent activation functions, respectively, which are standard choices in an LSTM network, and $x \odot y$ denotes a Hadamard product of two vectors $x$ and $y$. The sequence of snapshot vectors of $n_t$ time-steps: $V = [v^{i-n_t+1}, \ldots, v^{i-1}, v^i]$, with $V \epsilon \mathbb{R}^{n_s \times n_t}$ trains the LSTM network, with recurrence over time (Figure 6), to predict the subsequent vector $v^{i+1}$. The core concept of an LSTM network is the cell state $c_i$, which behaves as the "memory" of the network. It can either allow greater preservation of past information, reducing the issues of short-term memory, or it can suppress the influence of the past, depending on the actions of the various gates during the training process.

### 2.3.2 Temporal Convolution Network (TCN)

The TCN is based on two principles [34]: The network produces an output of the same length as the input, and there can be no leakage from the future into the past. To verify that the first principle is respected, the TCN uses a 1D fully-convolutional network (FCN) where each hidden layer has the same length as the input layer, and zero padding of length $(k-1)$ is added to keep subsequent layers the same length as previous ones. To respect the second principle, the TCN uses causal convolutions (achieved by padding only on the starting side of input sequences), where the output at time $i$ is convolved only with elements from time $i$ and earlier in the previous layer (Figure 7). A TCN also makes use of dilated convolutions that enable an exponentially large receptive field. For an input sequence, $V \epsilon \mathbb{R}^{n_s \times n_t}$ and a kernel $K$ with learnable weights, $K \epsilon \mathbb{R}^k$ ($k$ is the kernel size), the element $O(s)$ with s $\epsilon \{0, 1, ..., n_t - k + 1\}$ produced by the dilated 1D convolution is:

$$O(s) = \sum_{j=0}^{k-1} V(s + j * d) \times K(j) \tag{7}$$

where $d$ is the dilation factor and $k$ is the kernel size. When using dilated convolutions, $d$ is increased exponentially with the depth of the network (eg., $d = 2^l$ at level $l$ of the network), ensuring that some filter hits each input within a large effective history.

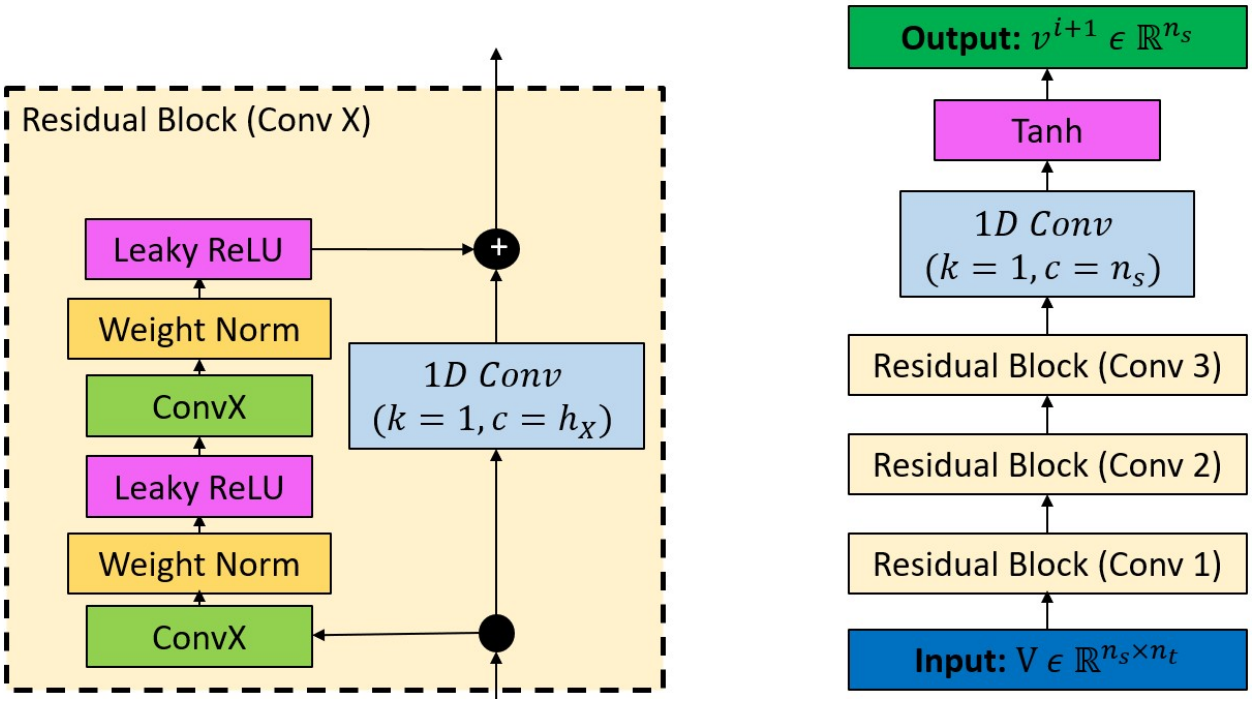



Figure 8: Residual Block (left) and TCN architecture (right)

In the TCN model employed here, a generic residual block is used in place of a convolutional layer. A residual block contains a branch leading to a series of transformations obtained by layers of TCNs, whose outputs are added to the input $V$ of the block to obtain $O_{rb}$:

$$O_{rb} = Activation(V + F(V)) \tag{8}$$

Within a residual block (Figure 8), the TCN has two layers of dilated causal convolution with weight normalization and non-linearity, with a leaky rectified linear unit (leaky ReLU). To account for different input-output widths during addition operations, a 1D convolution (kernel size = 1 and channels = $n_s$) is used to ensure the element-wise addition operator ($\oplus$) receives tensors of the same shape.

When convolving along the temporal axis, this (*standard*) TCN model uses information available from all the prior time-steps (due to the large receptive field) to evaluate the next time-step, as sketched in Figure 7. The model takes in a sequence of $n_t$ vectors corresponding to a look-back window of size $n_t$ : $V = [v^{i-n_t+1}, \ldots, v^{i-1}, v^i]$, with $V \, \epsilon \, \mathbb{R}^{n_s \times n_t}$. The filters convolve along the temporal axis for all the $n_s$ vector nodes, since the nodes are passed in as the channels. However, the results produced from this model (Section **??**) do not propagate beyond the training domain. Therefore, another model is proposed here, where the dilated convolutions of the TCN model convolve along the spatial axis and thus use the information available from the neighbouring nodes to determine the future time-step value of the node. This model takes in a sequence of $n_t$ vectors corresponding to a look-back window of size $n_t$ in a transposed manner, such that the $n_t$ solution vectors are on separate channels: $V^T \, \epsilon \, \mathbb{R}^{n_t \times n_s}$, where $V = [v^{i-n_t+1}, \ldots, v^{i-1}, v^i]$. This model produces significantly better results than the TCN on a temporal axis, but the causal padding and dilations employed are of no significance when the convolution filter operates along the spatial axis. Another architecture for modeling the system dynamics, with 1D convolutions and without any dilations or causal paddings, is therefore proposed in the following section.

### 2.3.3 A Convolution Neural Network (CNN) for time forecasting

A convolutional layer convolves filters with trainable weights on the input vector $v^i$ [31]. Such filters are commonly referred to as convolutional kernels. In a convolutional neural network, the inputs and outputs can have multiple channels. For a convolutional layer with $n_{in}$ input channels and $n_{out}$ output channels, the total number of convolutional kernels is $n_k = n_i \times n_o$. Each kernel slides along the spatial direction, and the products of kernel weights and vector

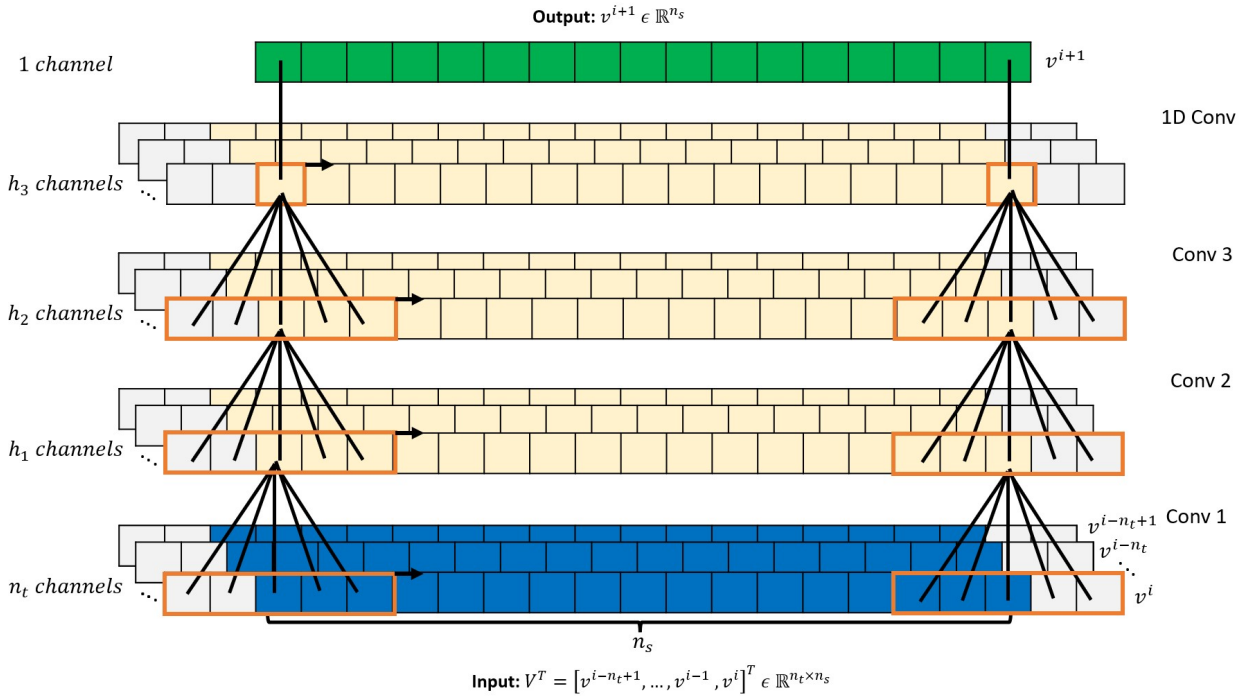



Figure 9: 1D CNN filters convolving on spatial dimension of vectors

nodes are computed at all sliding steps. For an input vector $v^i$ and a kernel $K$, the corresponding output feature map $O(s)$ with s $\epsilon\{0, 1, ..., n_s - k + 1\}$ (where $k$ is the kernel size) is given by:

$$O(s) = \sum_{j=0}^{k-1} v^i(s+j) \times K(j) \tag{9}$$

Zero padding of size $(k-1)/2$ is added to both sides of the output feature map to maintain the spatial dimension as $n_s$. The forecasting model of CNN takes in a sequence of vectors with $n_t$ time-steps in a transposed manner as its input: $V^T \epsilon \mathbb{R}^{n_t \times n_s}$, where $V = [v^{i-n_t+1}, \ldots, v^{i-1}, v^i]$, so that the filter convolves on the spatial dimension of size $n_s$, and the $n_t$ vectors lie on separate channels, as shown in Figure 9. The CNN architecture (Figure 10) consists of $X$ residual blocks ($X$ is a hyperparameter), in which the input to each block, after transformation (to make the channels equal) from a 1D Convolution layer (kernel = 1 and channels = 1) is added to the output from the block. A residual block consists of two convolution layers, each followed by a weight normalization and a leaky ReLU activation layer.

### 2.4 Uncertainty Quantification using Deep Ensembles

Deep neural networks, when applied in their traditional form, only predict the mean values of the output and do not provide any information regarding the uncertainty in the predicted output. Deep Ensembles address this issue by using an ensemble of variance-informed deep neural networks. Such neural networks possess a dual output [29, 35, 36]. In the context of the forecasting models employed with deep ensembles in this study, this means that the output size is twice the dimension $m$ of the predicted latent vectors ($z^{i+1}$) of solution steps, because the output contains both a mean value $\mu^z$ and a raw variance $\rho^z$. The raw variance is constrained to positiveness via a softplus activation, which produces the output variance ${\sigma^z}^2$.

$${\sigma^z}^2 = softplus(\rho^z) = \log(1 + \exp \rho^z) \tag{10}$$

The predicted variance accounts for the noise or spread in the data utilized by the forecasting models, thereby becoming a measure of the epistemic uncertainty [37].

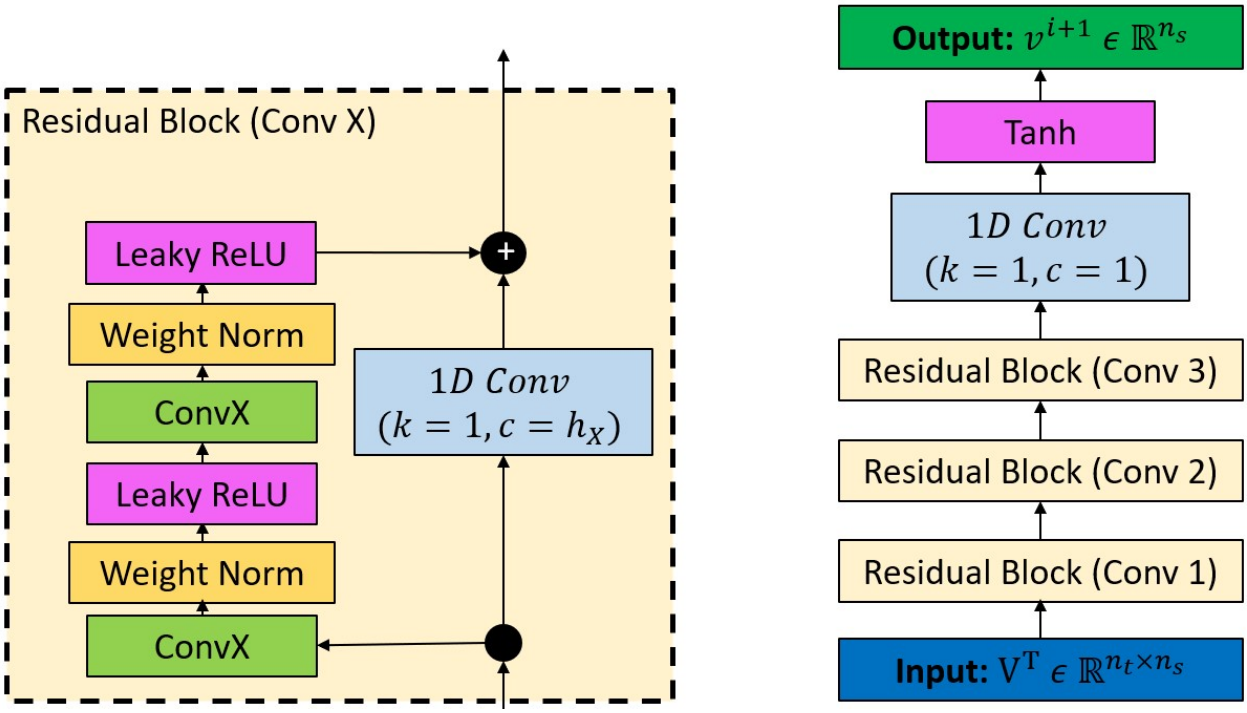



Figure 10: Residual Block (left) and CNN architecture (right)

### 2.4.1 Ensemble model training

The post-compression dataset (Section 2.1) from the encoder network produces $N$ samples of the form: $Z \epsilon \mathbb{R}^{m \times n_t}$, $z^{i+1} \epsilon \mathbb{R}^m$, which train the forecasting model with the objective of minimizing the negative log-likelihood (NLL) loss function, as given below:

$$J_{NLL}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \frac{\log {\sigma_i^z}^2}{2} + \frac{(z_i - \mu_i^z)^2}{2{\sigma_i^z}^2} \tag{11}$$

An optimizer (Adam algorithm [38]) is used to update the model parameters (weights and biases) - $\theta(\mathbf{w}, \mathbf{b})$. To account for the epistemic uncertainty, $M$ sets of $\theta(\mathbf{w}, \mathbf{b})$ are randomly initialized, thereby creating $M$ independent forecasting networks and introducing variability in the training step [36, 39]. The epistemic uncertainty is associated with the model variance and its data-fitting capabilities, which diminish with the increase in training data. This uncertainty is more dominant in our case, compared to aleatoric uncertainty, since the dataset is obtained by analytical solutions. Therefore, the variance behaves as an indicator of the likelihood that the model is making predictions that are out-of-distribution.

The predicted latent moments ($\mu_\theta^z$, ${\sigma_\theta^z}^2$) from each model create a probability mixture that can be approximated as a single Gaussian distribution, leading to a mean and variance obtained for the ensemble as:

$$\mu_*^z = \frac{1}{M} \sum_{i=1}^{M} \mu_{\theta_i^z} \tag{12}$$

and

$$ {\sigma_*^z}^2 = \frac{1}{M} \sum_{i=1}^{M} ({\sigma_{\theta_i}^z}^2 + \mu_{\theta_i}^z) - \mu_*^z \tag{13}$$

### 2.4.2 Predictions of the expanded solution vectors

During the online (testing) stage, the forecasting model predicts the latent vector $z^i$ along with the confidence interval at all time-steps after the input sequence. To obtain the full physical solution, the decoder ($\chi_d$) is used to reconstruct

or expand the obtained latent vector. Although the decoding operation can be directly applied to the predicted mean $\mu_*^z$, this does not hold true for the predicted variance $\sigma_*^{z2}$, because it does not transform linearly to the expanded space by using the decoding operation. Since the latent vector $z^i$ is represented as a normal distribution $\hat{z}^i \sim \mathbb{N}(\mu_*^z, \sigma_*^{z2})$, a sufficiently large sample from $\hat{z}^i$ can be drawn and reconstructed individually to produce the full solution vector $\hat{v}^i = \chi_d(\hat{z}^i)$.

The unscented transform, as proposed by Julier et. al [40], is used to calculate the predicted mean ($\mu_*^v$) and variance ($\sigma_*^{v2}$) of the full-order solution in expanded space ($v^i \, \epsilon \, \mathbb{R}^{n_s}$) from the predicted mean and variance of $z^i$, with the help of nonlinear transformation ($\chi_d$). A set of $2m + 1$ points (known as sigma points) with a sample mean and sample variance, $\mu_*^z$ and $\sigma_*^{z2}$, respectively, are chosen, and the nonlinear function is applied to each point to yield a set of transformed points, of $\mu_*^v$ and $\sigma_*^{v2}$ sample mean and variance, respectively. The primary difference of this approach from the Monte-Carlo method is that the samples are not drawn at random but instead according to a specific, deterministic algorithm. The points are obtained using:

$$\begin{aligned} \tilde{z_0} &= \mu_*^v, \;\; W_0 = k/(m+k) \\ \tilde{z_i} &= \mu_*^v + (\sqrt{(m+k)P_{zz}})_i, \;\; W_i = 1/2(m+k) \\ \tilde{z}_{i+n} &= \mu_*^v - (\sqrt{(m+k)P_{zz}})_i, \;\; W_i = 1/2(m+k) \end{aligned} \tag{14}$$

where $k \, \epsilon \, \mathbb{R}$ is a constant, $P_{zz} \, \epsilon \, \mathbb{R}_{m \times m}$ is the covariance matrix formed by placing the elements of $\sigma_*^{z2}$ as the diagonal, $(\sqrt{(m+k)P_{zz}})_i$ is $i^{th}$ row or column of the matrix square root of $(m+k)P_{zz}$ and $W_i$ is the weight associated with the $i^{th}$ point. The transformation is then achieved via the following procedure:

1. Transform each sigma point ($i \, \epsilon \{0, 1, ..., 2n\}$) by the non-linear decoder function $\chi_d$ to obtain:

$$y_i = \chi_d(\tilde{z_i}) \tag{15}$$

2. Predict the mean for the full-order solution using the weighted average of transformed points:

$$\mu_*^v = \sum_{i=0}^{2n} W_i y_i \tag{16}$$

3. Predict the variance for the full order solution from the diagonal elements of the covariance matrix $P_{vv}$, which are the weighted outer products of the transformed points:

$$P_{vv} = \sum_{i=0}^{2n} W_i (y_i - \mu_*^v)(y_i - \mu_*^v)^T \tag{17}$$

$$\sigma_*^{v2} = diag(P_{vv}) \tag{18}$$

### 2.4.3 Metrics

To evaluate the performance of the previous architectures, the following metrics are used:

*Mean Squared Error ($L_2$ Norm)*: The average of the square of the difference between the actual $v_i$ and predicted values $\hat{v}_i$ over $N$ samples:

$$MSE = \frac{\sum_{i=1}^{N} (v_i - \hat{v}_i)^2}{N} \tag{19}$$

*Mean Absolute Error ($L_1$ Norm)*: The average of the difference between the two vectors $v_i$ and $\hat{v}_i$ over $N$ samples:

$$MAE = \frac{\sum_{i=1}^{N} \|v_i - \hat{v}_i\|}{N} \tag{20}$$

*Relative $L_2$ Norm Error*: The relative $L_2$ norm error (referred as error) is calculated as:

$$RelativeError = \frac{\sqrt{\sum_{i=1}^{N} (v_i - \hat{v}_i)^2}}{\sqrt{\sum_{i=1}^{N} v_i^2}} \tag{21}$$

## 3 Results and Discussion

The capability of the autoencoders (MLP-AE and CAE) to efficiently transform high-dimensional vectors to a low dimensional space, and that of the forecasting models (LSTM, TCN, and CNN) to accurately model the system dynamics were tested using advection-dominated flow problems (1D Burgers' and Stoker's problems).

### 3.1 1D Burgers' problem

The test case involves the one-dimensional Burgers' equation, which is a non-linear advection-diffusion PDE. The equation along with the initial and Dirichlet boundary conditions are given by

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial t^2} \tag{22}$$

$$u(x,0) = u_0, x\epsilon[0,L], u(0,t) = u(L,t) = 0 \tag{23}$$

$$u(x,0) \equiv u_0 = \frac{x}{1+\sqrt{\frac{1}{t_0}}\exp(Re\frac{x^2}{4})} \tag{24}$$

where the length $L = 1m$ and the maximum time $T_{max} = 2s$. The solutions obtained from the above equations produce sharp gradients even with smooth initial conditions if the viscosity $\nu$ is sufficiently small, due to the advection-dominated behavior. The analytical solution of the problem is given by:

$$u(x,t) = \frac{\frac{x}{t+1}}{1+\sqrt{\frac{t+1}{t_0}}\exp(Re\frac{x^2}{4t+4})} \tag{25}$$

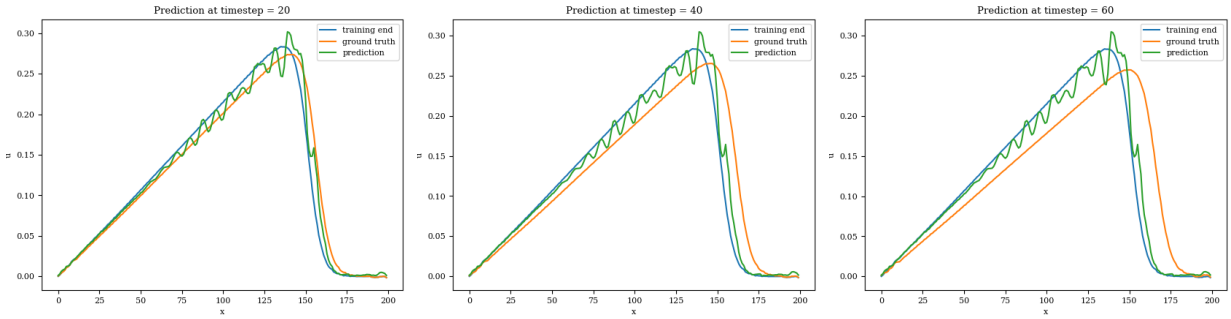
where $t_0 = \exp(\frac{Re}{8})$ and $Re = 1/\nu$. The high-fidelity solution vectors are generated by directly evaluating the analytical solution over a uniformly discretized spatial domain containing 200 grid points ($n_s = 200$) at 250 uniform time-steps ($T = 250$) for two different values of $Re$: 300 and 600. The solution vectors obtained are then used to train the autoencoder and forecasting models (Section 2.1). For the autoencoder training, 200 solution vectors are chosen at random time-steps, and the remaining 50 are used for validation. For the forecasting model, the training set is comprised of the first 150 compressed samples, each sample containing $n_t$ consecutive solution vectors (i.e. look back window = $n_t$), where $n_t$ is a hyperparameter. The validation set is composed of the remaining samples. For testing, $n_t$ latent vectors from the start of the dataset are fed to the forecasting model to predict the subsequent time-steps via auto-regression (Table 1).

| Dataset | Samples | Input | Output |
|---|---|---|---|
| Training | 1 | $[z^1, \ldots, z^{n_t-1}, z^{n_t}]$ | $z^{n_t+1}$ |
| | 2 | $[z^2, \ldots, z^{n_t}, z^{n_t+1}]$ | $z^{n_t+2}$ |
| | ... | ... | ... |
| | 150 | $[z^{150}, \ldots, z^{n_t+148}, z^{n_t+149}]$ | $z^{n_t+150}$ (training end) |
| Validation | 151 | $[z^{151}, \ldots, z^{n_t+149}, z^{n_t+150}]$ | $z^{n_t+151}$ |
| | ... | ... | ... |
| | 250-$n_t$ | $[z^{250-n_t}, \ldots, z^{248}, z^{249}]$ | $z^{250}$ |
| Testing | 1 | $[z^1, \ldots, z^{n_t-1}, z^{n_t}]$ | $[z^{n_t+1}, \ldots, z^{249}, z^{250}]$ |

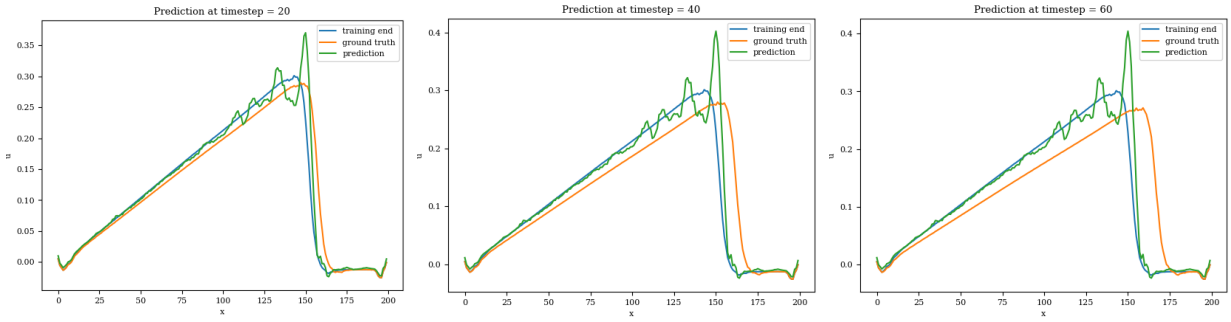
Table 1: Burgers' problem, Training, Validation and Testing dataset

#### 3.1.1 Autoencoders for spatial compression

Two types of autoencoder architectures (Section 2.2) - MLP (referred to as AE) and Convolutional (referred to as CAE) are proposed for the compression of solution vectors - $v^i \epsilon \mathbb{R}^{n_s}$ to latent vectors $z^i \epsilon \mathbb{R}^m$ by the encoder $\chi_e$. Sequences formed from these latent vectors ($[z^{i-n_t+1}, \ldots, z^{i-1}, z^i] \epsilon \mathbb{R}^{m\times n_t}$) are utilized to train the forecasting models - LSTM, TCN and CNN. The trained models are then used to forecast the latent vectors at subsequent time-steps to the sequence ($[z^{n_t+1}, z^{n_t+2}, \ldots, z^T] \epsilon \mathbb{R}^{m\times(T-n_t)}$) given as input to the forecasting model. The latent vectors are then reconstructed into solution vectors ($[v^{n_t+1}, v^{n_t+2}, \ldots, v^T] \epsilon \mathbb{R}^{n_s\times(T-n_t)}$) using the decoder $\chi_d$. The heat map plots obtained by




(a) Re = 300




(b) Re = 600

Figure 11: Burgers' problem, Extrapolative auto-regressive predictions by the LSTM modelfor time-steps = 180, 200 and 220. The training end time-step = 160; for Re=300 (a) and Re= 600 (b)

stacking these reconstructed solution vectors along the x-axis (spatial nodes-$n_s$ along y, time-steps-$n_t$ along x), are illustrated for both autoencoder models in Appendix A (Table 11).

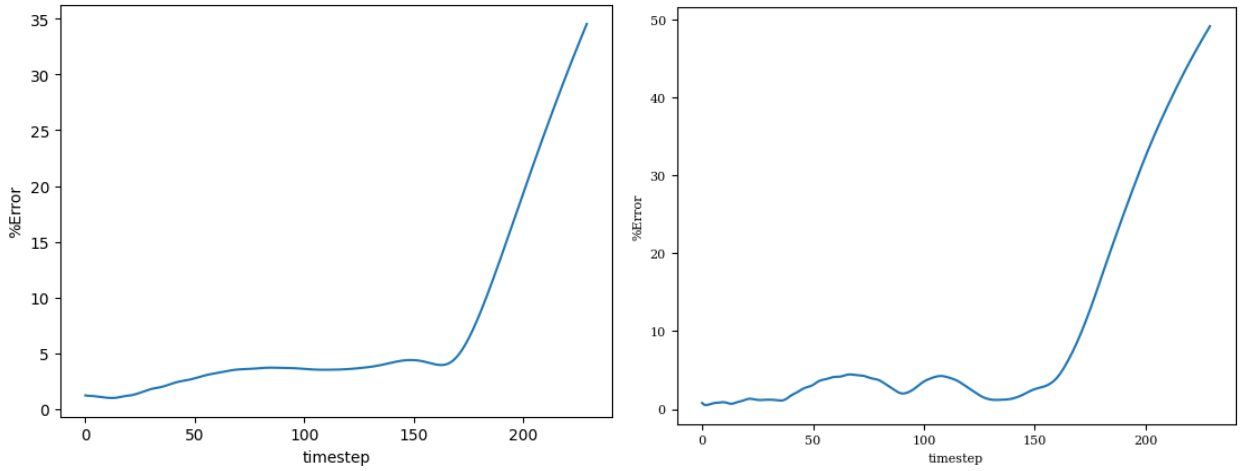
Both architectures, AE and CAE, are capable of efficiently compressing the solution vectors to latent vectors with few modes and fine reconstruction/decompression. However, only CAE compression followed by CNN autoregression produces accurate results on extrapolation. This is because the proposed CAE architecture is devoid of any dense layer (single layer of neurons), and therefore even during compression, the local spatial information in the vector remains preserved. This consistency facilitates the modeling of latent dynamics by the CNN model, as it convolves on the spatial axis of the input and utilizes information from the neighboring cells at the provided time-steps to predict nodal values at subsequent time-steps. The hyperparameters for the AE and CAE architectures are listed in Table 2, where encoder layers denote the number of neurons in the two dense encoder layers of AE, and the number of channels in the convolution layers of CAE. The decoders of both autoencoders are mirrored structures of their encoders.

| Hyperparameters | MLP AE | Convolutional CAE |
|---|---|---|
| Encoder layers | [100, 50] | [8, 32] |
| latent dimension ($m$) | 10, 25, 50 | 12, 25, 50 |
| Activation | relu, swish | relu, swish |
| Loss Function | MSE | MSE |
| Learning rate | $10^{-3}, 3 \times 10^{-4}$ | $10^{-3}, 3 \times 10^{-4}$ |

Table 2: Hyperparameters for the AE and CAE networks

### 3.1.2 LSTM model

When LSTM (Section 2.3.1)is used as the future step predictor, it takes in a sequence of $n_t$ (lookback window) latent vectors (spatial dimension = $m$) obtained by compression from the encoder network ($Z \epsilon \mathbb{R}^{m \times n_t}$) to produce the latent vector for the next time-step ($z^{i+1} \epsilon \mathbb{R}^m$). The LSTM model consists of multiple LSTM layers stacked together, each having a hidden dimension equal to the latent dimension of the solution vectors. Various sets of hyperparameters considered for the LSTM network, for both Re 300 and 600 are summarized in Table 3:




(a) Re = 300 (b) Re = 600

Figure 12: Burgers' problem, $L_2$ relative error of the autoregressive predictions with increasing time for Re = 300 (a) and Re = 600 (b) for the LSTM model

| Hyperparameters | Values |
|---|---|
| Sequence length ($n_t$) | 5, 10, 20 |
| LSTM layers | 1, 2, 3 |
| hidden/latent dimension ($m$) | 12,25,50 |
| Activation | tanh |
| Loss Function | MSE |
| Learning rate | $5 \times 10^{-4}$ |

Table 3: Hyperparameters for the LSTM network

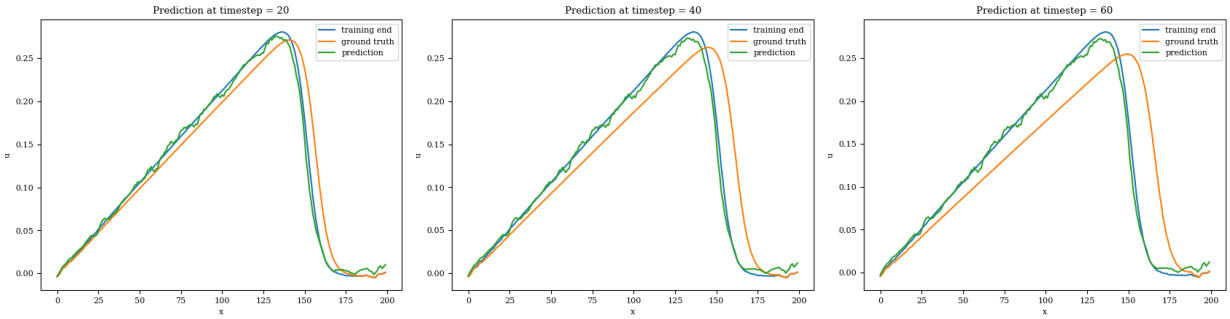
The models are trained in batches of size 15, and the loss values for both training and validation converge in 3000 epochs. The model with the least validation loss has a lookback window of size 10 and a single LSTM layer with hidden dimension 50 for both Re 300 and 600. The extrapolation (Figure 11) and error plots (Figure 12) obtained from these models show that the LSTM model accurately predicts the solution vectors for time-steps within the training domain ($i <= 150$), but the solution does not change for time-steps outside the training domain, and so the relative error increases drastically, reaching 35% for Re= 300 and 50% for Re=600.

### 3.1.3 TCN model

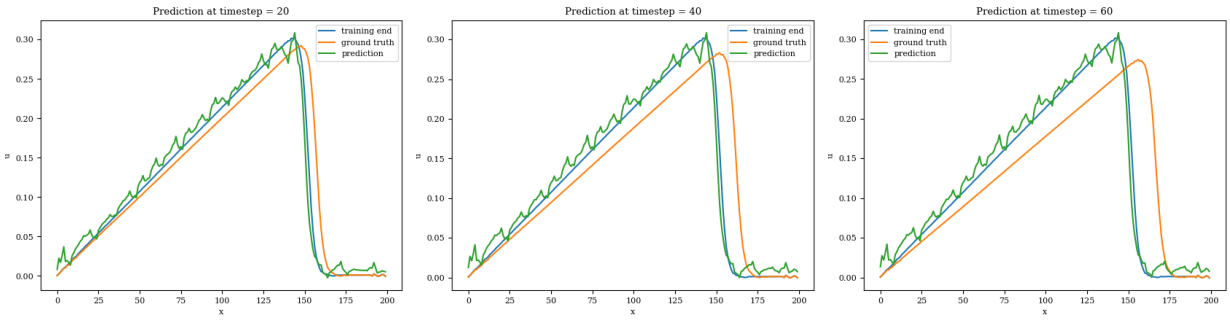
Similar to the LSTM, the TCN model (Section 2.3.2) takes in a sequence of $n_t$ latent vectors obtained by compression from the encoder network ($Z \, \epsilon \, \mathbb{R}^{m \times n_t}$) to produce the latent vector for the next time-step ($z^{i+1} \, \epsilon \, \mathbb{R}^m$), with the dilated convolutions operating on the temporal axis, and the latent dimension passed as a channel. The TCN model consists of either 2 or 3 TCN blocks, each with the same kernel size and number of channels, but with dilations increasing by a factor of 2 in subsequent blocks. The hyperparameters for the TCN network for both Re 300 and 600 are summarized in Table 4.

| Hyperparameters | Values |
|---|---|
| Sequence length ($n_t$) | 5, 10, 20 |
| TCN block channels | [32, 32], [64, 64], [32, 32, 32], [64,64,64] |
| latent dimension ($m$) | 12, 25, 50 |
| Kernel Size(k) | 3, 5, 7, 9 |
| Activation | tanh |
| Loss Function | MSE |
| Learning rate | $1 \times 10^{-4}$ |

Table 4: Hyperparameters for the TCN network

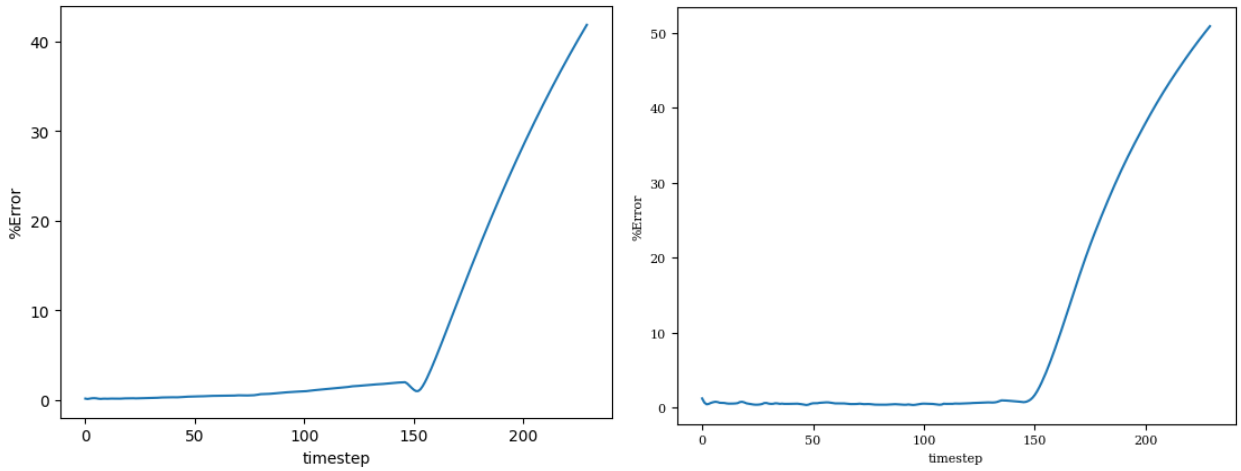



(a) Re = 300




(b) Re = 600

Figure 13: Burgers' problem, Extrapolative auto-regressive predictions using the TCN model (over time) for Re= 300 (a) and Re= 600 (b) and for time-steps = 180, 200 and 220; the training end time-step = 160

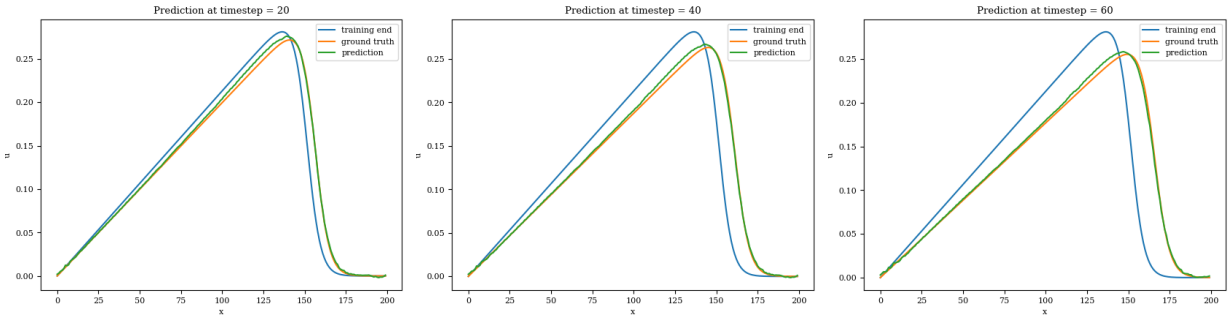



(a) Re = 300

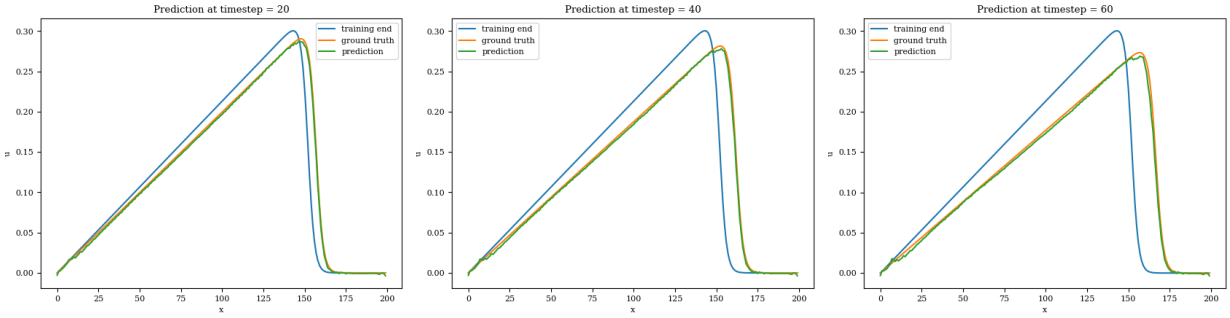
(b) Re = 600

Figure 14: Burgers' problem, $L_2$ relative error of the auto-regressive predictions with increasing time for Re = 300 (a) and Re = 600 (b) for the TCN model (over time)

When models are trained in batches of size 15, the training and validation losses reach their minimum values in 4000 epochs. The model with the least validation loss takes in sequence with lookback window 10 and latent dimension 50. For Re 300, the best model has 3 temporal blocks, each having 64 channels, whereas for Re 600, it has 2 TCN blocks with kernel size 3 and 64 channels each. The extrapolation (Figure 13) and error plots (Figure 14) obtained from these models indicate that the TCN model accurately predicts the solution vectors for time-steps within the training domain ($i <= 150$), but stops being accurate after the end of training, so that the error increases to 40% for Re= 300 and 50% for Re= 600. However, if the same model architecture operates on the input sequence, such that the dilated 1D convolutions propagate along the spatial axis, with each solution vector on a separate channel, then accurate forecasts




(a) Re = 300




(b) Re = 600

Figure 15: Burgers' problem: Extrapolative auto-regressive predictions using the proposed CNN model, for Re= 300 (a) and Re= 600 (b) and for time-steps = 180, 200 and 220; the training end time-step = 160

are produced, even outside the training domain. This encourages the development of a simpler predictive/forecasting model, devoid of dilations and causal padding since the exponentially increasing receptive field serves no purpose when operating along the spatial axis.

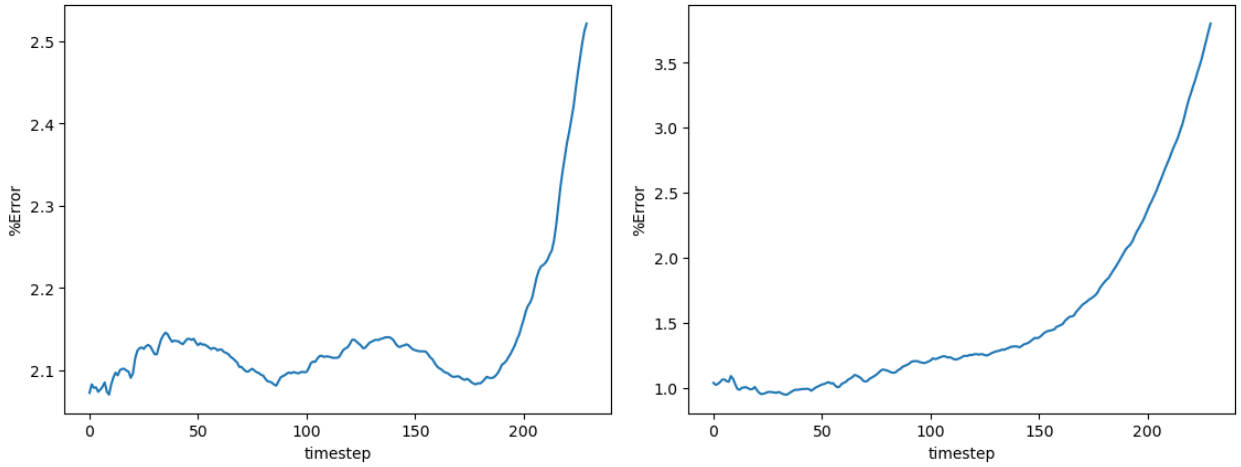
### 3.1.4 CNN model

The proposed CNN model (Section 2.3.3) takes in a sequence of $n_t$ latent vectors obtained by compression from the encoder network ($Z^T \epsilon \mathbb{R}^{n_t \times m}$) to produce the latent vector for the next time-step ($z^{i+1} \epsilon \mathbb{R}^m$), with 1D convolutions operating on the spatial axis (latent dimension) and $n_t$ latent vectors on separate channels. The CNN model consists of two residual blocks, each with the same kernel size and number of channels. The hyperparameters for the CNN network for both Re=300 and 600 are summarized in Table 5.

Training and validation loss converges in 3000 epochs for batch size 15. The model with the least validation loss has a lookback window of size 10, and each of its blocks has a kernel size of 3 and 50 channels for both Re= 300 and 600. It is clear from the extrapolation (Figure 15) and error plots (Figure 16) that the CNN model accurately models the latent dynamics, and predicts solution vectors accurately for time-steps beyond the training domain. The error values increase with time, due to the accumulation of errors, since each subsequent time-step is predicted auto-regressively; i.e., using

| Hyperparameters | Values |
|---|---|
| Sequence length ($n_t$) | 5, 10, 20 |
| CNN block channels | [50, 50], [100, 100], [200, 200] |
| latent dimension ($m$) | 12,25,50 |
| Kernel Size(k) | 3, 5, 7, 9 |
| Activation | tanh |
| Loss Function | MSE |
| Learning rate | $1 \times 10^{-4}$ |

Table 5: Hyperparameters for the CNN network




(a) Re = 300 (b) Re = 600

Figure 16: Burgers' problem: $L_2$ relative error of the auto-regressive predictions using the CNN model with increasing time for Re = 300 (a) and Re = 600 (b)

previously-predicted time-steps that contain slight errors. Still, the error reaches a mere 2.5% for Re= 300 and 3.5% for Re= 600, which is significantly less than that produced by other models.

### 3.2 1D Stoker's Equation

Stoker's solution describes the propagation and rarefaction wave resulting from a one-dimensional dam break over a wet, flat, frictionless bottom. Stoker's solution is considered among the most challenging benchmark test case due to its strong hyperbolic behavior and the discontinuity accompanying the propagation of the front wave resulting from the initial break. The dynamic is initiated by unequal water levels of both the upstream and downstream sides located in the middle of the studied domain of 100 m.

The upstream water level is considered as an input random variable whose values are uniformly sampled within its plausible variability range $h_{up}\,\epsilon\,\mu[8, 11]$, whereas the downstream water depth is kept constant at a deterministic value $h_{ds} = 1m$. The analytical solution for the water level is given as:

$$h(x,t) = \begin{cases} h_{up} & \text{if } x \leq x_A(t) \\ \frac{4}{9g}(\sqrt{gh_{up}} - \frac{x}{2t})^2 & \text{if } x_A(t) \leq x \leq x_B(t) \\ \frac{c_m^2}{g} & \text{if } x_B(t) \leq x \leq x_C(t) \\ h_{ds} & \text{if } x_C(t) \leq x \end{cases} \tag{26}$$

where $x$ = the axial position, $x_A(t) = x_0 - t\sqrt{gh_{up}}$, $x_B(t) = x_0 + t(\sqrt{gh_up} - 3c_m)$ and $x_C = x_0 + t\frac{2c_m^2(\sqrt{gh_up} - c_m)}{c_m^2 - gh_{ds}}$, in which $c_m = \sqrt{gh_m}$ [28]. For each selected value in the generated sample set of the upstream water level, the

| Dataset | Samples | Input | Output |
|---|---|---|---|
| Training | 1 | $[z^1, \ldots, z^{n_t-1}, z^{n_t}]$ | $z^{n_t+1}$ |
| | 2 | $[z^2, \ldots, z^{n_t}, z^{n_t+1}]$ | $z^{n_t+2}$ |
| | ... | ... | ... |
| | 250 | $[z^{150}, \ldots, z^{n_t+248}, z^{n_t+249}]$ | $z^{n_t+250}$ (training end) |
| Validation | 251 | $[z^{251}, \ldots, z^{n_t+249}, z^{n_t+250}]$ | $z^{n_t+251}$ |
| | ... | ... | ... |
| | 450-$n_t$ | $[z^{450-n_t}, \ldots, z^{448}, z^{449}]$ | $z^{450}$ |
| Testing | 1 | $[z^1, \ldots, z^{n_t-1}, z^{n_t}]$ | $[z^{n_t+1}, \ldots, z^{449}, z^{450}]$ |

Table 6: Stoker's problem: Training, Validation and Testing dataset

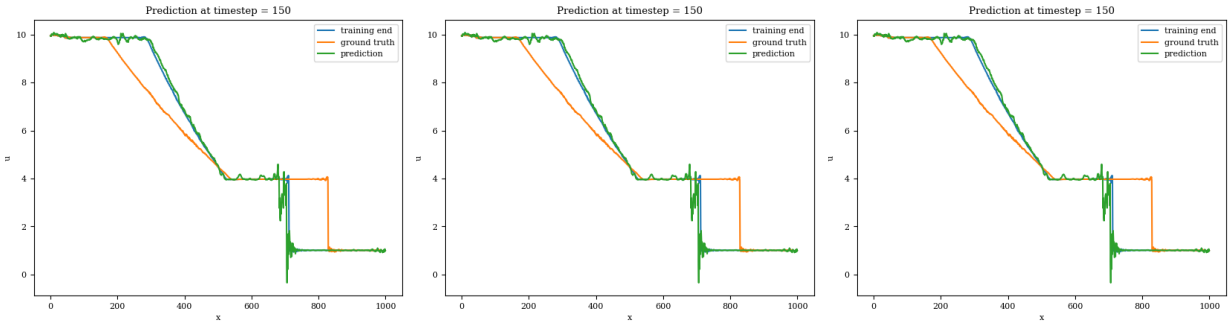



Figure 17: Stoker's problem: Extrapolative auto-regressive predictions of the LSTM model for time-steps = 310, 360 and 410; the end of training time-step = 260

analytical solution given above is evaluated over 1000 nodes ($n_s = 1000$) that contain the computational domain $x \epsilon [0, 100]$ m for all 450 time-steps ($T = 450$) of the temporal domain $t \epsilon [0, 3.6]s$. Four-hundred solution vectors, at random time-steps, train the autoencoder network, and the remaining 50 vectors are used for the validation. The forecasting models are trained by the first 250 compressed samples and validated using the remaining 200. During testing, the first $n_t$ latent vectors are utilized to predict vectors at subsequent time-steps via auto-regression (Table 6).

### 3.2.1 Autoencoder for spatial compression

A similar methodology to the Burgers' test case 3 is adopted for the training of autoencoder models, AE and CAE, and forecasting models, LSTM, TCN, and CNN, for the Stoker's problem. The heat map plots obtained by stacking the predicted solution vectors along the x-axis (spatial nodes-$n_s$ along y, time-steps-$n_t$ along x), are illustrated for both autoencoder models in appendix A (Table 12).

Both AE and CAE effectively transform the solution vectors to a reduced latent space, since they produce fine reconstruction for vectors within as well as outside of the training domain of the autoencoder model. But again, only the CAE-CNN model learns the latent dynamics accurately enough to predict solution vectors outside the training domain (extrapolation). The hyperparameters for the AE and CAE architectures are listed in Table 7.

| Hyperparameters | MLP AE | Convolutional CAE |
|---|---|---|
| Encoder layers | [500, 250] | [8, 32, 32] |
| latent dimension ($m$) | 25, 50, 125 | 25,50,125 |
| Activation | relu, swish | relu, swish |
| Loss Function | MSE | MSE |
| Learning rate | $10^{-3}, 3 \times 10^{-4}$ | $10^{-3}, 3 \times 10^{-4}$ |

Table 7: Hyperparameters for the AE and CAE networks

### 3.2.2 LSTM

The LSTM model receives input $Z \epsilon \mathbb{R}^{m \times n_t}$ and predicts $z^{i+1} \epsilon \mathbb{R}^m$ (Section 2.3.1). The LSTM model has an architecture similar to that of the Burgers' case, with multiple LSTM layers having a latent dimension as their hidden dimension. The hyperparameters of the LSTM network for the Stoker's problem are summarized in Table 8.

The models are trained in batches of size 15, and the loss values for both training and validation converge in 2400 epochs. The model with the least validation loss has a lookback window of size 10 and 2 LSTM layers with hidden dimensions of 125 each. The extrapolation (Figure 17) and error plots (Figure 18) obtained from these models show that the LSTM model accurately estimates the solution vectors for time-steps within the training domain ($i <= 250$), but fails outside the training domain, as the relative error reaches 25% for the 150th time-step post-training.

### 3.2.3 TCN model

The TCN model also receives a sequence $Z \epsilon \mathbb{R}^{m \times n_t}$ and forecasts $z^{i+1} \epsilon \mathbb{R}^m$, with the dilated convolutions operating on the temporal axis, and the latent dimension passed as a channel. The model contains three temporal blocks (Section

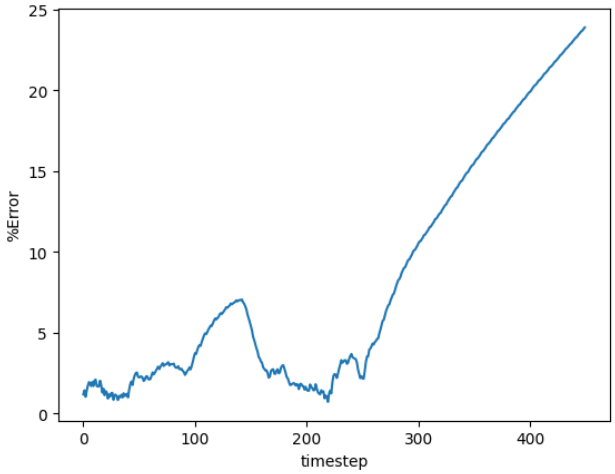



Figure 18: Stoker's problem" $L_2$ relative error of the auto-regressive predictions with increasing time for the LSTM model

| Hyperparameters | Values |
|---|---|
| Sequence length ($n_t$) | 5, 10, 20 |
| LSTM layers | 1, 2, 3 |
| hidden/latent dimension ($m$) | 25, 50, 125 |
| Activation | tanh |
| Loss Function | MSE |
| Learning rate | $5 \times 10^{-4}$ |

Table 8: Hyperparameters for the LSTM network

2.3.2), with the same kernels and channels in each, and dilations that increase in size by a factor of two in subsequent blocks. The hyperparameters for the TCN network are listed in Table 9.

| Hyperparameters | Values |
|---|---|
| Sequence length ($n_t$) | 5, 10, 20 |
| TCN block channels | [100, 100, 100], [200, 200, 200] |
| latent dimension ($m$) | 25,50,125 |
| Kernel Size(k) | 3, 5, 7, 9 |
| Activation | tanh |
| Loss Function | MSE |
| Learning rate | $3 \times 10^{-4}$ |

Table 9: Hyperparameters for the TCN network

Model training and validation loss reaches convergence by 1000 epochs, when training and validation is performed in batches of size 15. The best model (with the least validation loss) accommodates a sequence of vectors with a lookback window of size 20, and 125 latent modes. Each temporal block contains kernels of size 3 and has 200 channels. The extrapolation (Figure 19) and error plots (Figure 20) obtained from these models indicate that similar to the LSTM, the TCN also predicts the solution vectors with acceptable accuracy for time-steps within the training domain ($i <= 250$), but the solution stops propagating further in the extrapolative domain, and so the error increases to 23%.

### 3.2.4 CNN model

The proposed CNN model takes the transposed vector sequence $Z^T \epsilon \mathbb{R}^{n_t \times m}$ to produce $z^{i+1} \epsilon \mathbb{R}^m$, with 1D convolutions operating on the spatial axis or latent dimension, and the $n_t$ latent vectors present on separate channels. The CNN model consists of three residual blocks (Section 2.3.3), each with the same kernel size and number of channels. The hyperparameters for the CNN model are summarized in Table 10:

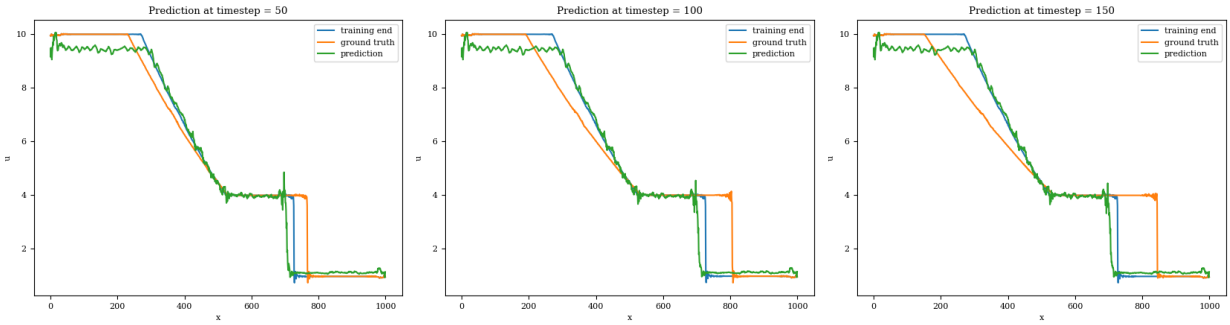



Figure 19: Stoker's problem: Extrapolative auto-regressive predictions using the TCN model for time-steps = 320, 370 and 420. The training end is at time-step = 270)

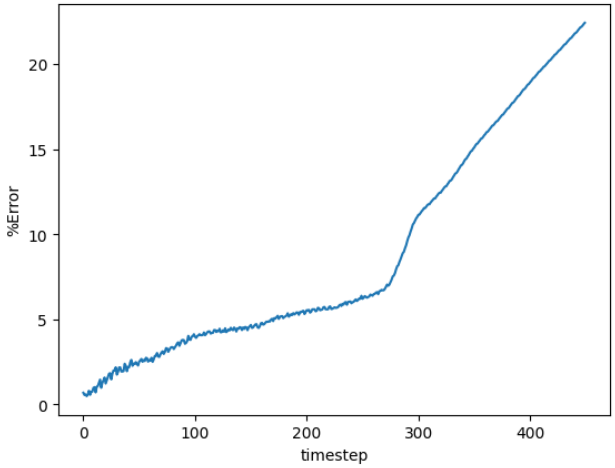



Figure 20: Stoker's problem: $L_2$ relative error of the auto-regressive predictions with increasing time for the TCN model

The training and validation loss converges at around 1000 epochs when batches of size 16 are used. The model with the highest accuracy (least validation loss and RA) has a lookback window of 20 steps, with 125 nodes in latent vectors at every step. Each of the three blocks possess a kernel of size 3 and 100 channels in the 1D convolution layers. The extrapolation (Figure 21) and error plots (Figure 22) indicate that the CNN model is capable of modelling the latent dynamics, since accurate forecasts are produced for time-steps beyond the training domain. The error values increase over time, reaching 5%, which is remarkably lower than that of earlier models.

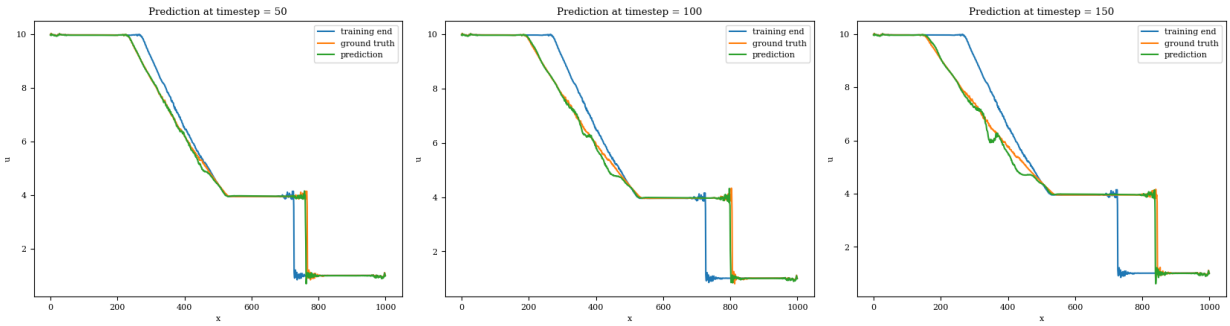



Figure 21: Stoker's problem: Extrapolative auto-regressive predictions using the CNN model for time-steps = 320, 370 and 420; the end of training is at time-step = 270

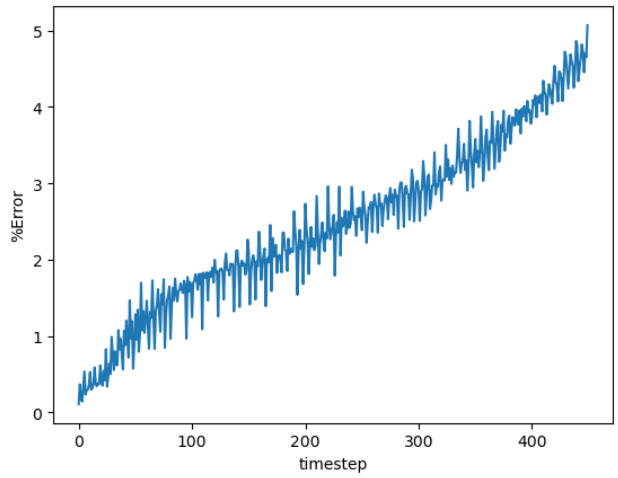



Figure 22: Stoker's problem, $L_2$ relative error of the auto-regressive predictions with increasing time for the CNN model

| Hyperparameters | Values |
|---|---|
| Sequence length ($n_t$) | 5, 10, 20 |
| TCN block channels | [50, 50, 50], [100, 100, 100], [200, 200, 200] |
| latent dimension ($m$) | 25, 50, 125 |
| Kernel Size(k) | 3, 5, 7, 9 |
| Activation | tanh |
| Loss Function | MSE |
| Learning rate | $3 \times 10^{-4}$ |

Table 10: Hyperparameters for the CNN model

## 4 Uncertainty Quantification using Deep Ensembles

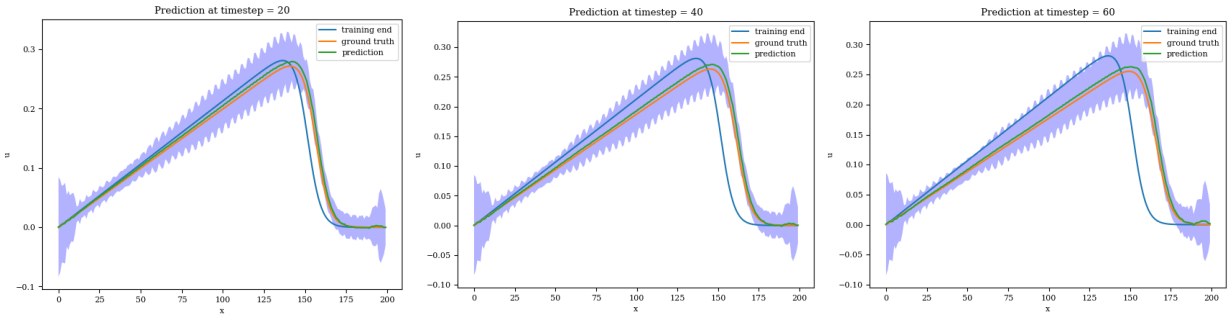
The training-testing methodology described previously2.4 for the evaluation of variance in the predicted output is adopted for the most stable (i.e., similar results for discrete initialization) and most accurate (i.e., least relative error on extrapolation) forecasting model, (i.e. the CNN future step predictor with a CAE autoencoder). After the model is trained to minimize the NLL, the mean and variance for the latent vector, $\mu_*^z$ and $\sigma_*^{z\,2}$ respectively, are calculated using predictions from $M = 10$ models. The mean and variance, $\mu_*^v$ and $\sigma_*^{v\,2}$ for predicted vectors $[v^{n_t+1}, v^{n_t+2}, \ldots, v^T]$ are then calculated using $2m + 1$ sigma points sampled (with $k = 0.2$) on $\hat{z}$ and decompressed in accordance with Equation 14 and 15.

The extrapolation results obtained by the ensemble are demonstrated in Figure 23 for the 1D Burgers' test case with Re = 300 and 600, and in Figure 25 for the 1D Stoker's test case.

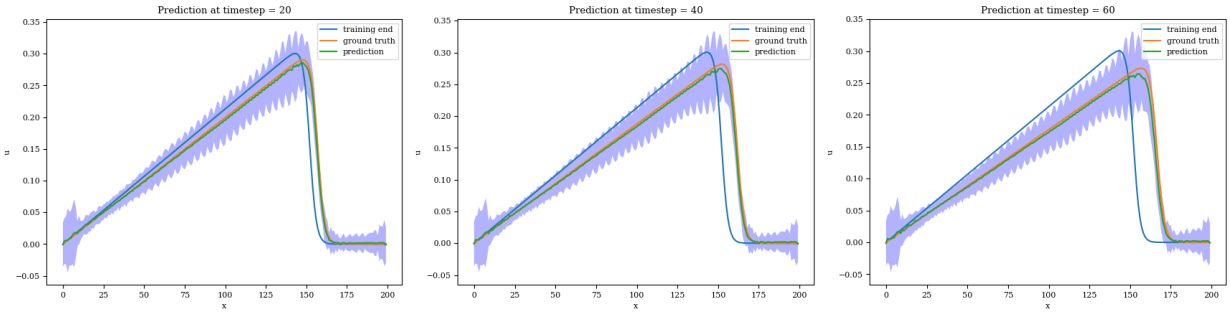
The predicted mean for all time-steps outside the training domain overlaps with the actual value, and the error reaches a maximum of 6% (Figure 24) for the Burgers' and 3.5% (Figure 26) for the Stoker's problem. The variance is maximum at the spatial terminal ends of the predicted vector solutions, as they lack neighbouring nodes on one side, which are essential for information extraction using 1D convolution filters for accurate future step estimations. The variance remains almost constant throughout the prediction time, indicating the robustness of the proposed model.

## 5 Conclusion

This study proposes a Convolutional Autoencoder(CAE) model for compression and a CNN future step predictor for forecasting vector solutions for subsequent time-steps to input vector sequences. The approximation accuracy and time extrapolation capabilities of the model are evaluated using two advection-dominated flow problems, a 1D Burgers' equation and a 1D Stoker's equation, which are characterized by sharp gradients and discontinuities, respectively. The models built especially for time-series forecasts, LSTM and TCN models with propagation over time, produce acceptable results within the training domain, but the solution stops changing during extrapolation. However, when

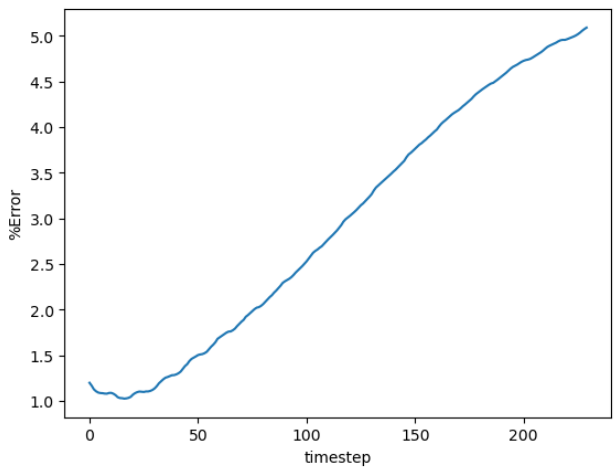



(a) Re = 300

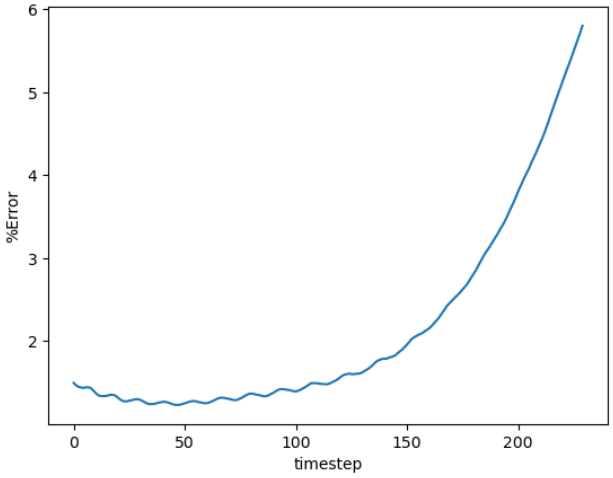



(b) Re = 600

Figure 23: Burgers' problem: Extrapolative auto-regressive predictions by the deep ensemble CAE-CNN model for time-steps = 180, 200 and 220 and for Re= 300 (a) and Re= 600 (b); the end of training is at time-step=160




(a) Re = 300




(b) Re = 600

Figure 24: Burgers' problem: $L_2$ relative error of the auto-regressive predictions with increasing time for Re = 300 (a) and Re = 600 (b) for a deep ensemble CAE-CNN model

the dilated convolutions propagate on the spatial axis, the models produce good predictions for extrapolation as well. The proposed CNN model for forecasting has an architecture (residual blocks with 1D convolutions propagating along space) similar to that of the TCN model, but without causal padding or dilation. These have been eliminated, as the increasing receptive field has no significance during convolution in space, and causal padding degrades the results by causing the information extracted from neighbouring cells to be shifted or scraped. The CNN model is capable of producing highly accurate predictions for both test cases, with less than 5% relative $L_2$ error in the extrapolation domain ($\approx$ 60% of the training time domain). However, the CNN model only produces accurate forecasts if it receives

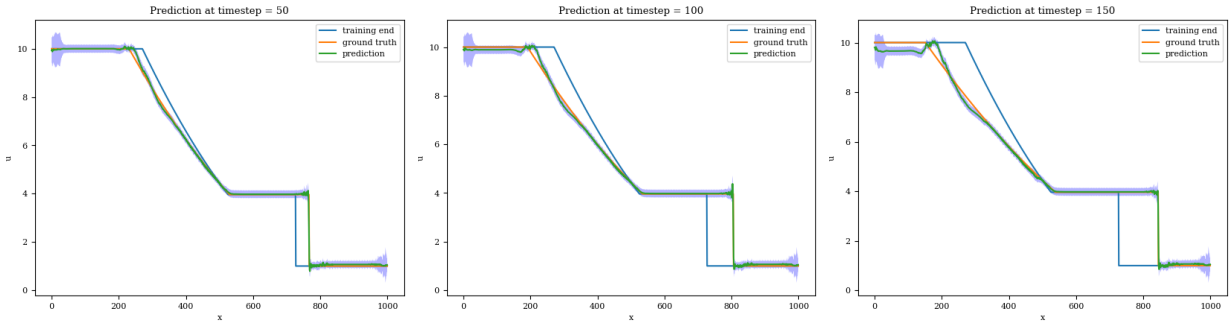



Figure 25: Stoker's problem: Extrapolative auto-regressive predictions by a deep ensemble CAE-CNN model for time-steps = 320, 370 and 420; the end of training is at time-step = 270

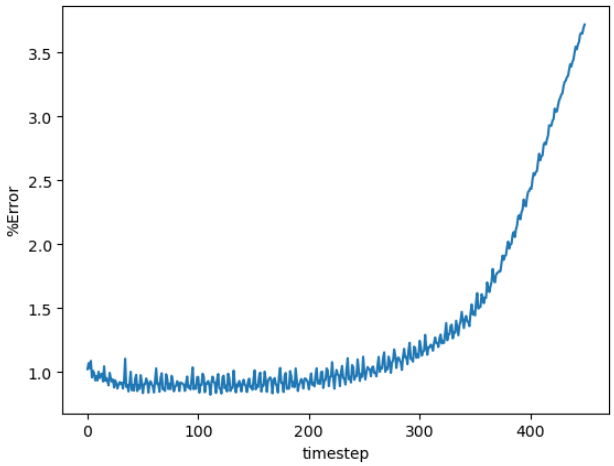



Figure 26: Stoker's problem: $L_2$ relative error of the auto-regressive predictions with increasing time for the deep ensemble CAE-CNN model

compressed latent vectors from the CAE model, since it preserves the local spatial information better than the MLP encoder during compression. Uncertainty quantification for the variance-informed CAE-CNN model is performed using deep ensembles, which produce accurate predictions, with low variance for long-term extrapolates as well (errors less than 6%). In addition, since the CNN model uses convolutions (like TCN [34]), the training and evaluation can be done in parallel for long input sequences, in contrast to the LSTM. Thus, a fast, accurate and robust framework is provided for order reduction. The model architecture is flexible and can be extended for two- and three-dimensional spaces as well, by increasing the dimension of the convolutional filters.

It would be interesting to test if the CAE-CNN architecture can adapt to problems where the solutions/data for the model are obtained using unstructured meshes. Future work could also focus on adapting the architecture to resolve real-life engineering problems.

## Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada and MITACS; the financial support is gratefully acknowledged. The computing resources were provided by the Digital Research Alliance of Canada, whose support is highly appreciated.

## References

[1] Clarence W. Rowley and Scott T.M. Dawson. Model reduction for flow analysis and control. *Annual Review of Fluid Mechanics*, 49(1):387–417, 2017.

[2] Kunihiko Taira, Steven L. Brunton, Scott T. Dawson, Clarence W. Rowley, Tim Colonius, Beverley J. McKeon, Oliver T. Schmidt, Stanislav Gordeyev, Vassilios Theofilis, Lawrence S. Ukeiley, and et al. Modal analysis of fluid flows: An overview. *AIAA Journal*, 55(12):4013–4041, 2017.

[3] G Berkooz, P Holmes, and J L Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Review of Fluid Mechanics*, 25(1):539–575, 1993.

[4] Alexander Lozovskiy, Matthew Farthing, Chris Kees, and Eduardo Gildin. Pod-based model reduction for stabilized finite element approximations of shallow water flows. *Journal of Computational and Applied Mathematics*, 302:50–70, 2016.

[5] Elnaz Rezaian, Rajarshi Biswas, and Karthik Duraisamy. Non-intrusive parametric reduced order models for the prediction of internal and external flow fields over automobile geometries. *Volume 4: Advances in Aerospace Technology*, 2021.

[6] Sourav Dutta, Peter Rivera-Casillas, Orie M. Cecil, and Matthew W. Farthing. Pynirom—a suite of python modules for non-intrusive reduced order modeling of time-dependent problems. *Software Impacts*, 10:100129, 2021.

[7] Alessandro Alla and J. Nathan Kutz. Nonlinear model order reduction via dynamic mode decomposition. *SIAM Journal on Scientific Computing*, 39(5), 2017.

[8] Danish Rafiq and Mohammad Abid Bazaz. Nonlinear model order reduction via nonlinear moment matching with dynamic mode decomposition. *International Journal of Non-Linear Mechanics*, 128:103625, 2021.

[9] Akshay S. Deshpande and Jonathan Poggie. dynamic mode decomposition of a highly confined shock-wave/boundary layer interaction. 2020.

[10] D. Xiao, F. Fang, C.C. Pain, and I.M. Navon. A parameterized non-intrusive reduced order model and error analysis for general time-dependent nonlinear partial differential equations and its applications. *Computer Methods in Applied Mechanics and Engineering*, 317:868–889, 2017.

[11] Sourav Dutta, Matthew W. Farthing, Emma Perracchione, Gaurav Savant, and Mario Putti. A greedy non-intrusive reduced order model for shallow water equations. *Journal of Computational Physics*, 439:110378, 2021.

[12] Zhengxiao Ma, Jian Yu, and Ruoye Xiao. Data-driven reduced order modeling for parametrized time-dependent flow problems. *Physics of Fluids*, 34(7):075109, 2022.

[13] D. Xiao. Error estimation of the parametric non-intrusive reduced order model using machine learning. *Computer Methods in Applied Mechanics and Engineering*, 355:513–534, 2019.

[14] George Karniadakis, Yannis Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. pages 1–19, 05 2021.

[15] J.S. Hesthaven and S. Ubbiali. Non-intrusive reduced order modeling of nonlinear problems using neural networks. *Journal of Computational Physics*, 363:55–78, 2018.

[16] Zhong Yi Wan, Pantelis Vlachas, Petros Koumoutsakos, and Themistoklis Sapsis. Data-assisted reduced-order modeling of extreme events in complex dynamical systems. *PLOS ONE*, 13(5), 2018.

[17] Romit Maulik, Arvind Mohan, Bethany Lusch, Sandeep Madireddy, Prasanna Balaprakash, and Daniel Livescu. Time-series learning of latent-space dynamics for reduced-order model closure. *Physica D: Nonlinear Phenomena*, 405:132368, 2020.

[18] Sourav Dutta, Peter Rivera-Casillas, Brent Styles, and Matthew W. Farthing. Reduced order modeling using advection-aware autoencoders. *Mathematical and Computational Applications*, 27(3):34, 2022.

[19] Federico Fatone, Stefania Fresca, and Andrea Manzoni. Long-time prediction of nonlinear parametrized dynamical systems by deep learning-based reduced order models, 2022.

[20] R.T.Q. Chen, Y Rubanova, J Bettencourt, and D Duvenau. Neural ordinary differential equations. *In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18), Montréal, QC, Canada,*, page 6572–6583, Dec 2018.

[21] Sourav Dutta, Peter Rivera-Casillas, and Matthew Farthing. Neural ordinary differential equations for data-driven reduced order modeling of environmental hydrodynamics. 03 2021.

[22] Pin Wu, Junwu Sun, Xuting Chang, Wenjie Zhang, Rossella Arcucci, Yike Guo, and Christopher C. Pain. Data-driven reduced order model with temporal convolutional neural network. *Computer Methods in Applied Mechanics and Engineering*, 360:112766, 2020.

[23] Jiayang Xu and Karthik Duraisamy. Multi-level convolutional autoencoder networks for parametric prediction of spatio-temporal dynamics. *Computer Methods in Applied Mechanics and Engineering*, 372:113379, 2020.

[24] C. Liu, R. Fu, D. Xiao, R. Stefanescu, P. Sharma, C. Zhu, S. Sun, and C. Wang. Enkf data-driven reduced order assimilation system. *Engineering Analysis with Boundary Elements*, 139:46–55, 2022.

[25] Romit Maulik, Bethany Lusch, and Prasanna Balaprakash. Reduced-order modeling of advection-dominated systems with recurrent neural networks and convolutional autoencoders. *Physics of Fluids*, 33(3):037106, 2021.

[26] Fangxiong Cheng, Hui Xu, and Xinlong Feng. Model order reduction method based on (r)pod-anns for parameterized time-dependent partial differential equations. *Computers & Fluids*, 241:105481, 2022.

[27] Claire E. Heaney, Zef Wolffs, Jón Atli Tómasson, Lyes Kahouadji, Pablo Salinas, André Nicolle, Ionel M. Navon, Omar K. Matar, Narakorn Srinil, Christopher C. Pain, and et al. An ai-based non-intrusive reduced-order model for extended domains applied to multiphase flow in pipes. *Physics of Fluids*, 34(5):055111, 2022.

[28] Azzedine Abdedou and Azzeddine Soulaïmani. Reduced-order modeling for stochastic large-scale and time-dependent problems using deep spatial and temporal convolutional autoencoders, 2022.

[29] Pierre Jacquier, Azzedine Abdedou, Vincent Delmas, and Azzeddine Soulaïmani. Non-intrusive reduced-order modeling using uncertainty-aware deep neural networks and proper orthogonal decomposition: Application to flood modeling. *Journal of Computational Physics*, 424:109854, 2021.

[30] Nicholas Geneva and Nicholas Zabaras. Modeling the dynamics of pde systems with physics-constrained deep auto-regressive networks. *Journal of Computational Physics*, 403:109056, 2020.

[31] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[32] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[33] Sergios Theodoridis. *Machine learning a Bayesian and Optimization Perspective*. Elsevier, Academic Press, 2020.

[34] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271*, 2018.

[35] D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, 1994.

[36] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, 2017.

[37] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5580–5590, Red Hook, NY, USA, 2017. Curran Associates Inc.

[38] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[39] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[40] Simon J. Julier and Jeffrey K. Uhlmann. New extension of the kalman filter to nonlinear systems. *SPIE Proceedings*, 1997.

# A Heat Map Plots for Burgers' and Stoker's problems

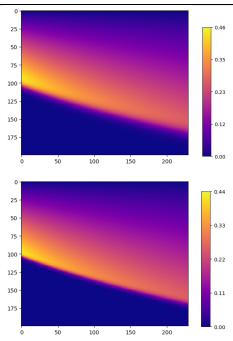
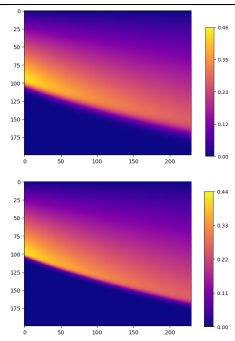
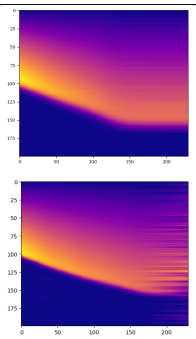
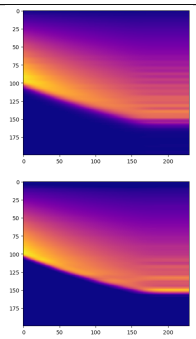
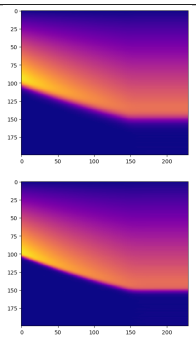
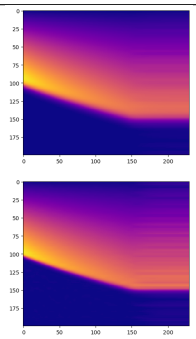
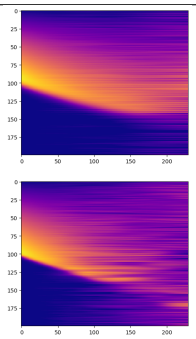
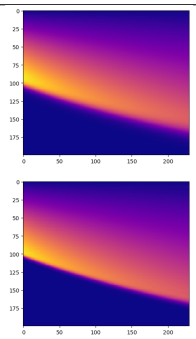
| Forecasting Models | MLP AE | Convolutional CAE |
|---|---|---|
| Ground Truth |  |  |
| LSTM |  |  |
| TCN |  |  |
| CNN |  |  |

Table 11: Burgers' problem, auto-regressive forecasts for Re= 300 (top) and Re= 600 (bottom) from the forecasting models when latent vectors are obtained by AE and CAE

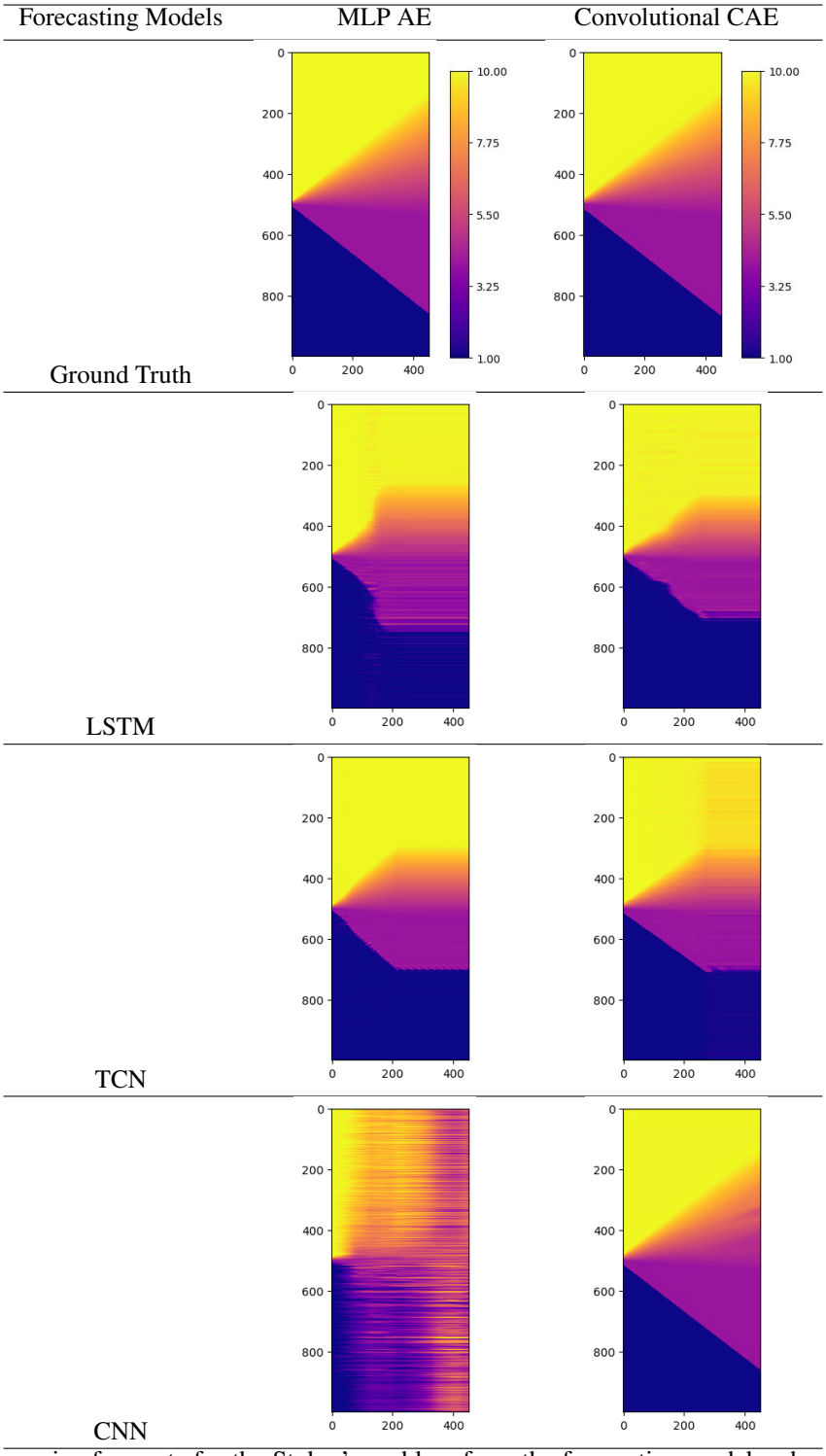



Table 12: Auto-regressive forecasts for the Stoker's problem from the forecasting models when latent vectors are obtained by AE and CAE

FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 429-437 | e-ISSN: 2347-470X

# Design and Development of a Solar Electric Delivery Pod

**Abhishek Arora[1] and Janardan Kesari[2]**

*[1,2]Department of Mechanical Engineering, Delhi Technological University, Delhi, India, [1]abhiarora4567@gmail.com, [2]profkesari@dtu.ac.in*

***Correspondence:** Abhishek Arora; Email: abhiarora4567@gmail.com

**ABSTRACT-** Use of autonomous solar electric vehicles for road delivery purposes in India remains highly unexplored area. The aim of these research paper is to cover major design aspects and develop a prototype solar electric delivery vehicle with autonomous drive for city logistic purposes. The vehicle design is constructed using the locally available raw materials and is aerodynamically tested. CAD modelling on SolidWorks'20 has been used to build a virtual physical model of the vehicle and aerodynamic testing is performed. MATLAB Simulink has been used to develop a mathematical simulation model with various subsystems like battery, motor, vehicle dynamics and PV charging module. Vehicle has been tested for 2 major drive cycles and results are plotted. The vehicle operates with a wheel hub BLDC motor by rechargeable Li-ion battery pack charged with the help of PV panel.

**General Terms:** Vehicle Development, Mathematical Modelling, Computer Aided Design, Solar Energy.

**Keywords:** Solar Vehicle, Electric Vehicle, Urban Freight, Industry 4.0, Autonomous vehicles, E-commerce.

ARTICLE INFORMATION

**Author(s)**: Abhishek Arora and Janardan Mishra
**Received**: 01/04/2022; **Accepted**: 17/07/2022; **Published**: 10/08/2022;
**e-ISSN**: 2347-470X;
**Paper Id**: IJEER100304;
**Citation:** 10.37391/IJEER.100304
**Webpage-link**:
https://ijeer.forexjournal.co.in/archive/volume-10/ijeer-100304.html

This article belongs to the Special Issue on **Recent Advancements in the Electrical & Electronics Engineering**

**Publisher's Note:** FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

## 1. INTRODUCTION

Exploring newer alternatives to conventional fossil-based energy is the prime focus of present generation amongst which solar energy seems to be the most promising one [13]. India is putting major emphasis into the subject since 1980s with setting up of the Ministry of New and Renewable Energy (MNRE) working to derive cleaner energy for the country. Last 2 decades witnessed a major boom in the development of renewable alternatives to fuels and with a net solar potential of 750GW laid out by MNRE by 2022 the country hopes to increase the share of solar energy to 175 GW with support from active government policies and solar parks [3].

The Indian transport sector uses around 94 million tons of oil equivalent which is 18% of total energy consumed by the country [10]. Along with that in the past decade, explosive growth of e commerce has generated massive amount of deliveries in urban areas with numbers growing at a rapid rate. The need for fast home parcel is putting a lot of pressure on the urban logistics network [5]. Implementation of EVs in local city logistics mainly in last mile deliveries seems like a viable option for freight transport industry [4]. European projects like BESTUFS 2006; SUGAR 2011, and BESTFACT working on developing best inventory for urban freight incorporate electric assistance in cargo cycles, delivery consolidation centers and using electric vans for last mile delivery. Such inventories are reflecting the growth of electric deliveries in modern cities [8]. On-Road Autonomous Delivery Robots (RADRs) equipped with sensors and navigation systems delivering groceries and parcels is the future [15]. The major barriers for EVs in real-life urban freight transport are manly their extensively high cost, charging duration, limited payload and range which are critical to the logistics companies. Though the national and many state governments are deriving out policies to provide incentives on purchase and use of EVs [6]. With NEMMP, FAME I and FAME II (by Department of Heavy Industry) the government of India is encouraging widespread adoption of EVs in major cities, further by GST slashing, exemption from permits requirements, road tax incentives and by building charging stations the government is working on developing a sustainable EV ecosystem in the country [10].

Immense amount of literature related to studies on EVs, autonomous drives, Solar energy is available but there is a lack of studies on compiling PV technology with electric autonomous drives and using it for urban logistics [2]. Developing a commercial solar electric vehicle presently seems unfeasible due to efficiency barrier of PV modules and Li-ion batteries along with cost constraints [16-17]. Also a major barrier is that solar energy is extremely variable over places and time [13]. We need a sustainable solution to these barriers to ensure growth and development in the sector along with least effect on the environment.

## 2. SYSTEM MODEL

### 2.1 Surveys and Reference Studies

Mamta L Prajapati et al studied and analyzed the success with focus on innovative developments of solar electric vehicles like

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 429-437 | e-ISSN: 2347-470X

cars, e-rickshaws, buses and in India as a large scale alternative for future [1]. Mohammad Waseem et al reviewed the solar mass scale integration of PV technology with 3 wheeled and 4 wheeled solar-electric drives, PV compatibility with robotics and self-driving modules and challenges in their adoption [2]. Muhammad Irfan et al analysed major features of solar based power like consumption, life, operational and maintenance cost, emissions and conditional requirements to derive out the potential for solar power in 4 major Indian cities [3]. S. Iwan et al. performed studied the electro mobility potential for city logistics and last mile deliveries using Nissan eNV200 in Szczecin, Poland [4]. R.A. de Mello Bandeira et al. performed similar study on using electric tricycles and LDVs as delivery vehicles covering economic environmental and social aspects to it [5]. S.M. Mirhedayatian, S. Yan created a frame network combining optimization techniques and economic analysis for delivery companies on response to EV support policies like taxes and subsidies by performing numerical experimentation and sensitivity analysis [6]. Muhammad Hosnee Mobarak et al modelled a solar electric vehicle and analysed the solar capture of 150 drivers on response to panel tilting and shading to calculate energy consumption, grid improvements, driver response, environmental and economic advantages of SEVs [7]. Christophe Rizet et al. analysed the effects of constraints of vehicle range, payload and emissions on transferring urban freight from ICEVs to EVs in France [8]. Tharsis Teoh et al reviewed the 4 opportunity charging (OC) strategies—during breaks and shift changes, during loading activity, during unloading activity, or while driving on highways using a parametric model to study effects on costs and emissions [9]. V.S. Patyal et al. analysed and modelled 13 major barriers to adoption and future prospects of EVs for different major states like Delhi, Karnataka, Tamil Nadu, UP, MP etc. in India to recognize pertinent barriers and insights into the EV economy [10]. Ashish Kumar and Lal Bahadur Prasad reviewed the issues and challenges to the mass adoption of Electric Vehicles with emphasis on policy recommendations, government schemes and ecosystem development [11]. Bambang Sri Kaloko et al. laid a foundation to the development of a solar electric vehicle using mathematical modelling of various vehicle dynamics and behaviors on MATLAB Simulink [12]. J. Ashok et al. developed a solar based electric vehicle with low eight and aerodynamic features for Indian roads. With solar different solar irradiation for different latitudes, optimized tilt angles were derived and an auto rickshaw was developed with appropriate weight, range and power [13]. Zahari Taha et al. constructed a solar car using reused and recycled parts after WSC off the shelf which was low in cost and at par in performance [14]. Jennings, Figliozzi studied the cost savings per dollar of using road autonomous delivery robots by comparing 3 major RADR developing technologies with conventional vehicle data and how it varied over size and time [15]. Paterson et al. designed and developed a record braking solar electric vehicle with arrange of 800km and speed over 130kmh [16].

## 2.2 Research Gap and Objectivity of Study

There was a lack of literature on development of solar based autonomous freight vehicles for Indian ecosystem. The rise of Industry 4.0 and smart manufacturing in India has paved the way for self-driven cars to be a reality in India. A major area of immediate impact is last mile freight and delivery logistics. Given the rise of E-commerce in India it is imperative to focus on cleaner ways of delivery and innovation. Major firms like Amazon (Amazon's Scout) are working to develop road autonomous delivery vehicles and bringing such models into their business line. Theoretical testing and research has supported the idea of replacing fossil based freight with electric self-driven systems. To take this one step further, integrating sustainable energy source for power like PV module seems highly appropriate. Recent large scale investment into the green energy sector and with government schemes like FAME I, FAME II and NEMMP India is moving towards electrification of its vehicle fleet at an astonishing rate. To aid such a revolutionary transformation it first needs to convert all its logistics to electric. The objective of this project was to design and develop such an electric vehicle to be used for logistics and delivery purposes mainly to be used on Indian roads. The vehicle would be autonomously driven and be powered by an onboard PV module making it one of a kind autonomous solar electric delivery pod. Based on findings and research conducted all over the world, we derived a combined methodology of preparing a mathematical model of all the vehicle subsystems on MATLAB Simulink and perform simulation testing on it. A CAD model was also developed for testing purposes and to work on design and ergonomics. As the vehicle was supposed to operate on Indian roads designed the vehicle for a top speed of around 60-80 kmph and a range of 60-80 km. The payload capacity was kept 25 kg. Commercially available parts with detailed specification sheets available were to be used and vehicle was supposed to be economical in manufacturing and operation.

# 3. COMPONENTS AND SPECIFICATIONS

The concept vehicle design consisted of (i) a chassis, (ii) a body, (iii) a traction motor with controller, (iv) a Li-Ion battery pack with a battery management system, (v) tires and (vi) a PV module with controller. The commercially available components used for the design and their major specifications used for designing purposes are listed in *Table 1 and Table 2* respectively. The complete design including all measurements and limitations were based on the specifications of these components so that the design is as close to the real world as possible. Subsystems have been developed individually. The vehicle subsystems have further been briefly described:

**Table 1: Commercial components used in vehicle**

| Vehicle Components | Commercial model used |
|---|---|
| BLDC in-hub motor | Tech Imperial 48V 2000W Hub Motor [21] |
| Cells used in Battery Pack | LG - Lithium Ion INR 21700 M50 18.20WH [17] |
| PV Panel | Luminous 330W 24V Polycrystalline Solar Panel, LUM 24330. [20] |
| Rear wheel | Ceat Zoom XL 80/100-17 46P Tubeless Bike Tyre, Rear [18] |
| Front Wheels | TVS Tyres Jaya 4.00-8 Tube-Type Auto Tyre [19] |
| Chassis Material | Carbon fibre - Hexcel AS4C (3000 filaments) |
| Body Material | Plastic - PP copolymer |

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 429-437 | e-ISSN: 2347-470X

## 3.1 Chassis and Body

The Chassis is the skeletal frame of the vehicle consisting of tires and suspension systems over which the vehicle body and accessories are attached. In the designed vehicle we used a carbon fibre chassis with a polypropylene copolymer vehicle body. The carbon fibre chassis was used because of its excellent weight to power characteristics and the copolymer plastic body for its low weight and durability. The chassis and body were mainly designed for low weight and optimal aerodynamic characteristics and were designed on Solidworks'20 using 3D sketching. The Body was based on teardrop design with equation: y = .25/.2*100*(.2969*sqr(x/75)-.126*(x/75)-.3516*(x/75) ^2+.2843*(x/75)^3-.1015*(x/75)^4) with x from 0 to 75. Similarly, iso pipes were used for chassis with 21.3 mm diameter and 2.3 mm thickness. The vehicle has been kept as spacious as possible so as to accommodate whatever it needs to deliver.

## 3.2 Wheels

The vehicle had a three wheeled configuration with 1 rear driven 17-inch diameter and 85 mm section width wheel and 2 front wheels with 8-inch rim diameter and 4-inch section width. Tubeless tires were used for the design with low rolling resistance and low weight. Both wheels had aspect ratio 1 and were attached to the vehicle chassis. CAD and mathematical model of wheels were designed with set specifications. The rear wheel had the BLDC hub motor.

## 3.3 Traction Motor

The traction motor is the part that replaces the engine in a conventional automobile to propel the electric vehicle. They derive electrical power from the battery pack and convert electricity to mechanical rotation at the wheels. They may be placed at the front or back according to design and connected to the wheels via driveline or some designs use a hub motor which is directly attached to the wheel. EV applications don't use conventional motors instead more efficient motors like BLDC, PMSM or SRM motors are put to use. In our design we used an in-hub BLDC motor which is attached to the rear wheel of our vehicle. The BLDC motor comes with a controller and works on direct current via phased pulses generation by the controller. BLDC motors are very efficient (our motor was 82% efficient), provide great power to weight ratio and they are very cost effective. A hub motor is used as it occupies lesser space (which is the purpose of a delivery vehicle), weighs less, is cheaper and provides excellent traction without any driveline losses. So for our design we used a 48V 2000W hub BLDC motor. The CAD design and Mathematical Simulink model were built as per specifications of the Tech Imperial motor.

**Table 2: Vehicle Component Specification**

| Specification | Value |
|---|---|
| | |
| **Chassis** | |
| Weight | 2.3 kg |
| | |
| **Body** | |
| Weight | 5.1 kg |
| | |
| **BLDC Motor** | |
| Operating Voltage | 48 V |
| Rated Power | 2000 W |
| Rated Torque | 12.5 Nm |
| Weight | 17.2 kg |
| Rated Current | 47 A |
| Rated RPM | 2850 |
| | |
| **Li-Ion Battery Pack** | |
| Nominal voltage of each cell | 3.63 V |
| Cell capacity | 18.2 Wh |
| Weight of each cell | 68 g |
| Cells in battery pack | 14S 14P |
| Battery pack voltage | 51 V |
| Battery pack current | 13.6 A |
| Battery pack capacity | 3.55 kWh |
| Weight of battery pack | 13.32 kg |
| | |
| **Rear Wheel** | |
| Weight | 4 kg |
| Rim diameter | 17 inch |
| Section width | 80 mm |
| Aspect ratio | 1 |
| | |
| **Front Wheel** | |
| Weight | 2 kg |
| Rim diameter | 8 inch |
| Section width | 4 inch |
| Aspect ratio | 1 |
| | |
| **PV Panel** | |
| Weight | 22.5 kg |
| Rated module voltage | 24 V |
| Peak power | 330 W |
| Open circuit voltage | 45.53 V |
| Short circuit current | 9.2 A |
| No. of cells | 72 |

## 3.4 Battery Pack

The battery pack is the most important and usually the most expensive part of an Electric Vehicle. It is analogous to the fuel tank of a conventional automobile. EV battery packs consist of modules connected in series. These modules themselves consist of cells connected in parallel. The standard cells used for EV battery packs are Li-Ion cells. They have significantly higher specific power and energy density than the lead acid or metal hydride cells also they can be discharged to deeper levels. Our design uses a Li-Ion cell which is collected into modules and packed into a battery pack with a configuration of 14S 14P to give a net battery pack voltage 51 V and capacity 3.5kWh. The battery pack is operated by a battery management system which monitors the state of health and state of charge of the pack to make sure pack operates on optimum level and safely. The Mathematical Simulink model and CAD model had been built and the pack was placed it the front of vehicle for better weight distribution.

## 3.5 PV Module

Photovoltaic panels convert heat energy from solar radiation to electricity. This generated emf is further boosted by a module and is used to charge the battery which turns the motor and

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 429-437 | e-ISSN: 2347-470X

propels the vehicle. The design uses a 72 cell PV panel with rated voltage 24V. The design vehicle has to operate in India, where there is abundance of solar irradiance in most places especially during summer months. The module design consisted pulse width modulation and a boost converter using a IGBT gate to charge the battery pack via a solar charging module. The specification sheet of Luminous LM 24330 was used to build the CAD model and Simulink model of the PV panel and it was placed over the top of the vehicle.

# 4. MODELING AND SIMULATION

## 4.1 CAD Modeling and Testing

All the components used in the pod design were recreated using Solidworks'20. The specifications were taken from the available data and were in accordance with the objectives of the project i.e. to maximize the range and payload capacity while reducing the weight and cost of the pod. We performed extensive research to find best off the shelf available products made from optimal material. Once all parts were designed and their material properties configured, they were assembled into the chassis and the copolymer body was added. The vehicle was 0.63 m wide ,1.62 m in length (without panel) and 0.74 m in height. The complete vehicle assembly was aerodynamically tested using flow simulation module of SW'20. The aerodynamic pressure drag at 60km/h (assumed as top speed) was around 33N. The calculated frontal area was around 0.41$m^2$ and the drag coefficient (Cd) was 0.46. The CAD models and flow simulation have been presented in *Figure 1-12*.

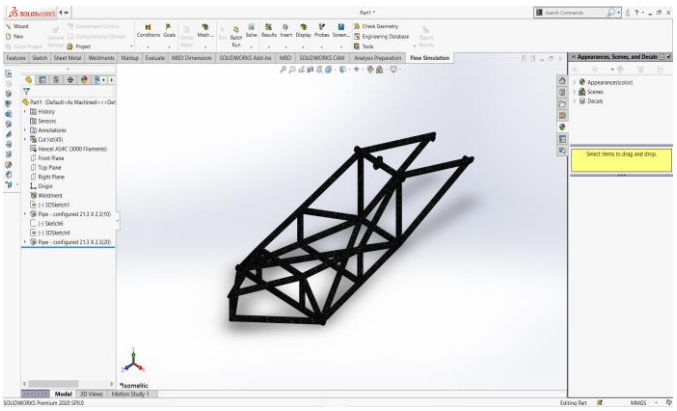


**Figure 1:** Chassis

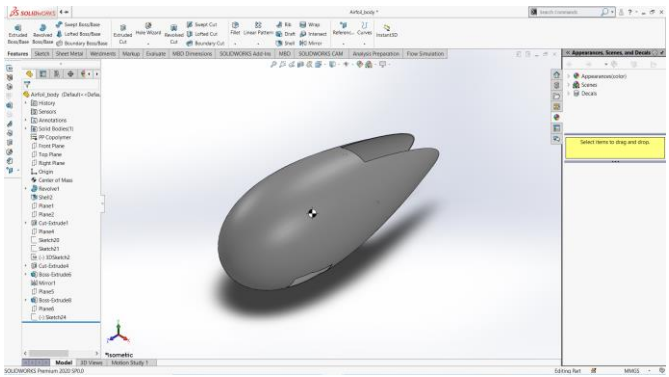


**Figure 2:** Body

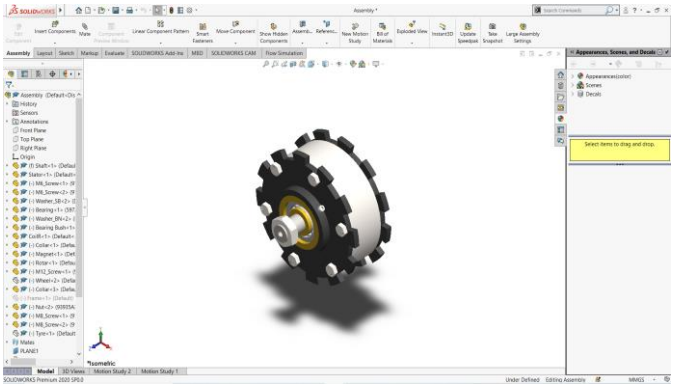


**Figure 3:** BLDC hub motor

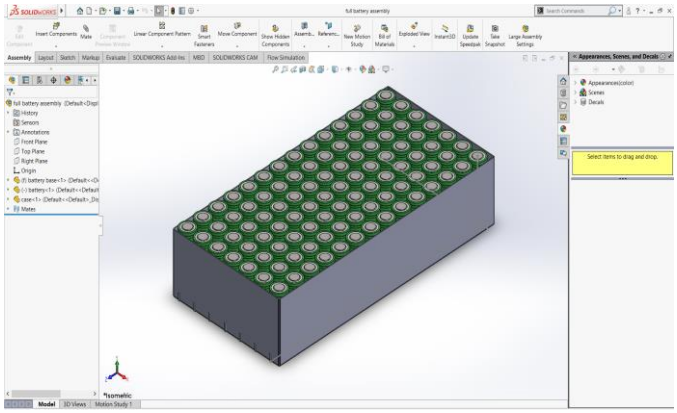


**Figure 4:** Battery pack

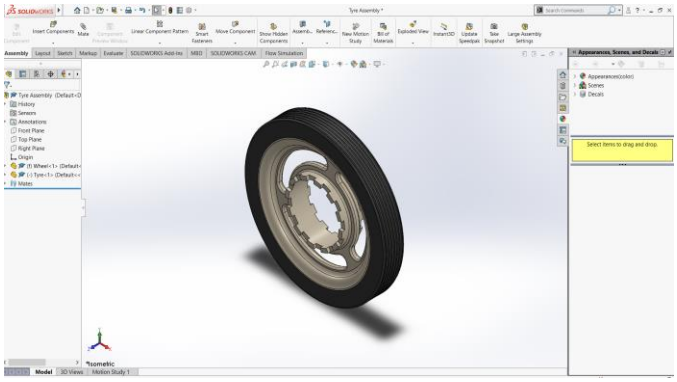


**Figure 5:** Rear wheel

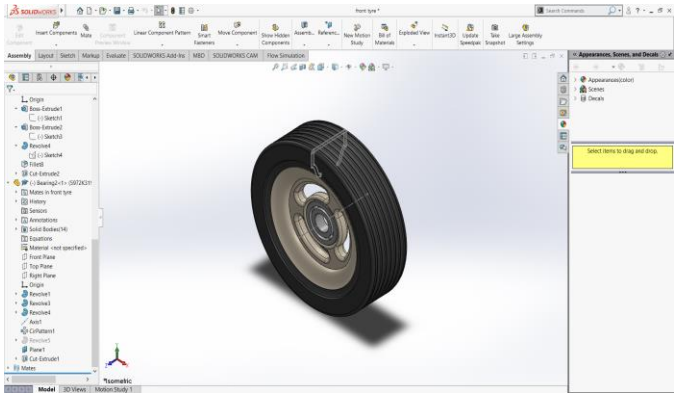


**Figure 6:** Front wheel

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 429-437 | e-ISSN: 2347-470X

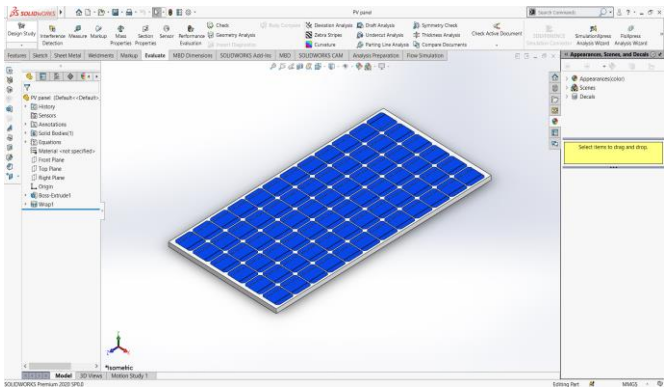


**Figure 7:** PV panel

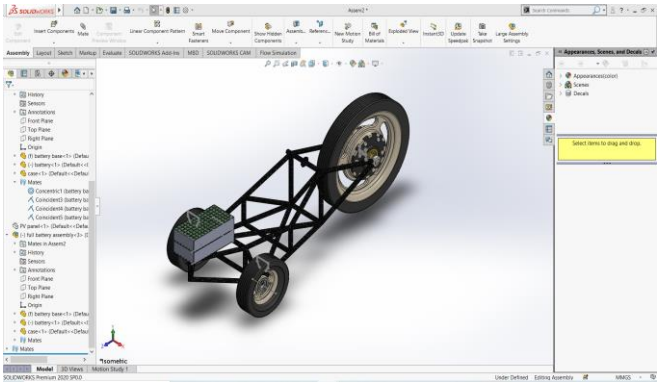


**Figure 8:** Chassis assembly

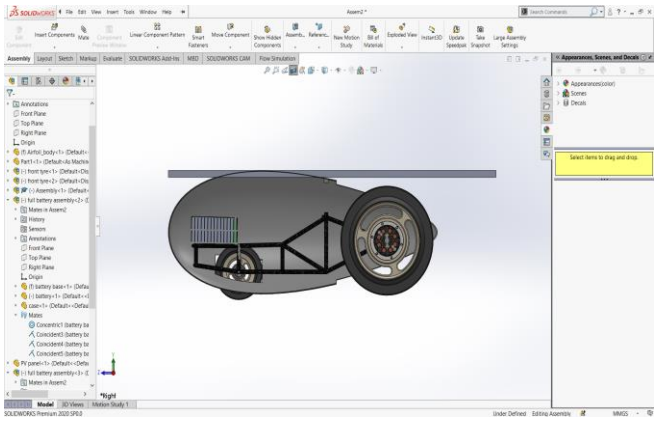


**Figure 9:** Vehicle Assembly

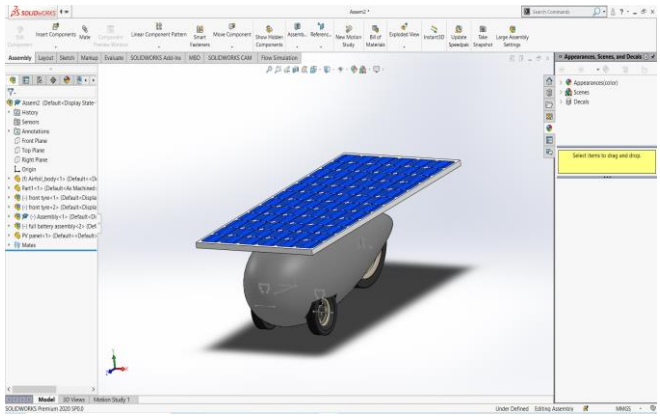


**Figure 10:** Complete vehicle

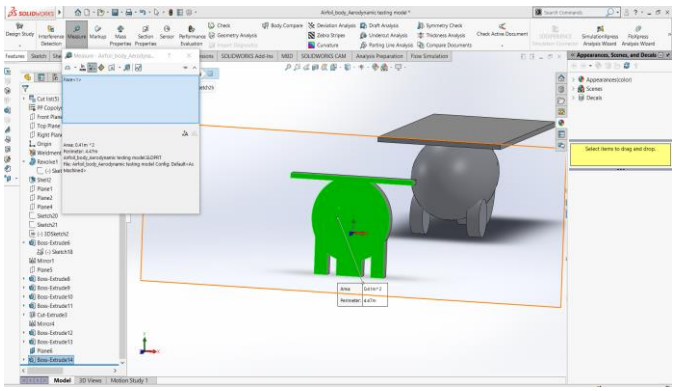


**Figure 11:** Frontal area calculation

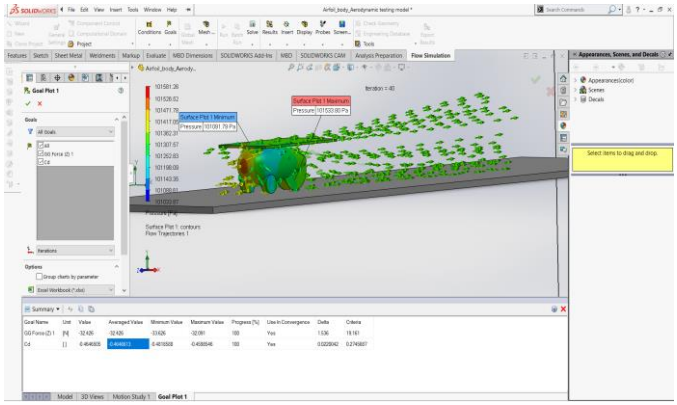


**Figure 12:** Air flow simulation

## 4.2 Mathematical Modeling

To perform various tests on the vehicle and check its operability it was imperative to mathematically model the different vehicle components and create a complete mathematical model of the project vehicle. We designed the equations-based subsystems on MATLAB Simulink using Simulink library and Simscape tools (for PV module). Major equation-based subsystems were designed and compiled to test the model for certified drive cycles. We tested the model for 2 drive cycles (WLTC_class1_1022 and FTP75) both having different characteristics to evaluate vehicle performance in different conditions. The complete Simulink model has been presented in *Figure 13* and the subsystems discussed briefly:

### 4.2.1 Drive Cycle Source

This provides the drive cycle input as velocity-time plot. In or project we test 2 drive cycles. (i) WLTC_class1_1022 cycle which is 1022 seconds 8103-meter-long cycle. It has a top speed of 17.9 m/s and frequent braking and accelerating maneuvers. It is suitable for inner city movement and heavy traffic conditions which is common on Indian roads. (ii) FTP75 cycle which is a 2474 second cycle 17667-meter cycle. It has a top speed of 25.3 m/s and is suitable representation of intense driving conditions with frequent acceleration and braking. Thus, both the drive cycles combined should provide an ideal indicator of project vehicle performance and represent majority of real-life cases. These drive cycles are fed to the leader vehicle and our project pod should follow this leader autonomously effectively.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 429-437 | e-ISSN: 2347-470X

#### 4.2.2 Control Module

This module consists of a leader vehicle, a mode selection module and a driver module. The leader gets the drive cycle input from the source module and initial speed and position input from the user to calculate position and speed. The mode selection module takes input leader reference position and relative speed to shift between 3 modes (i) free drive mode, (ii) vehicle following mode and (iii) emergency braking mode. During free drive mode when relative distance is greater than 50 m module commands free acceleration, when relative distance is less than 10 m module commands emergency braking. When relative position is between 10 m to 40 m module switches to a vehicle following mode which is based on a mpc controller and provides acceleration and braking commands on basis of reference leader's speeds and position and calculates our project vehicle's reference speed which it should maintain. The driver module inputs this reference speed and used a PID controller to output acceleration and braking command to motor and brake modules.

#### 4.2.3 Torque and Brake Distribution Module

This module receives acceleration and braking commands, present vehicle speed and motor rpm. It compares available motor's rated (& maximum) power which is 2000W and rated (& maximum) torque from the motor which is 12.5 N and calculates the net torque to be produced by the motor now. It also calculates and distributes regenerative and friction braking. The friction brake inputs act as input to the driveline module whereas the regenerative acts on the motor. The net motor torque is calculated by this difference and is fed to the Motor Energy and losses module and the driveline module.

#### 4.2.4 Motor Energy and Losses Module

This module inputs net motor torque, motor speed (in rpm) to calculate motor losses including power and energy losses, motor input energy and motor power. It is based on basic motor equations with values of constants taken from specification sheet of the motor. It mainly takes into account all the losses (copper and iron) and derives the net power input from the battery module.

#### 4.2.5 Battery Module

This module consists of the power storage system. It inputs net motor power requirements and accordingly supplies mathematically calculated voltage and battery current across its terminals. It takes into account the accessory loads, the open circuit voltage, the battery energy capacity, the discharge current and calculates the state of charge of the battery, energy and power at the terminals of battery during the drive cycle. It also accounts for the battery power and energy losses during battery operation. For the project vehicle the rated voltage is 51V, while the rated current is 13.6 A. The energy capacity of the battery pack is 3.54 kWh.

#### 4.2.6 Driveline Module

This module signifies the driveline of the vehicle. It inputs motor torque, friction brake command and present vehicle speed to calculate the driveline torque with driveline energy and power losses and eventually the net vehicle tractive force acting on the vehicle which is used for vehicle dynamics calculations. It takes into account the spin losses in the driveline and also calculates the motor speed (in rpm) which is required.

#### 4.2.7 Vehicle Dynamics Module

This module represents the base vehicle propulsion equations and vehicle dynamics. It inputs the net tractive force and calculates all the resistive forces including aerodynamic drag, gradient force, rolling resistance acting on the vehicle to derive final vehicle speed, tractive power, braking power and position.

#### 4.2.8 Solar Panel (PV) Module

This module consists of PV panels, a boost converter and the battery pack. The PV panels used have rated voltage 24V and OCV 45.53 V. There is a source input of solar irradiance and temperature over time. A boost converter with IGBT gate is provided with pulse width modulation of 10000Hz.The Li-ion battery pack is similar to the battery module and it gets charged over the cycle.

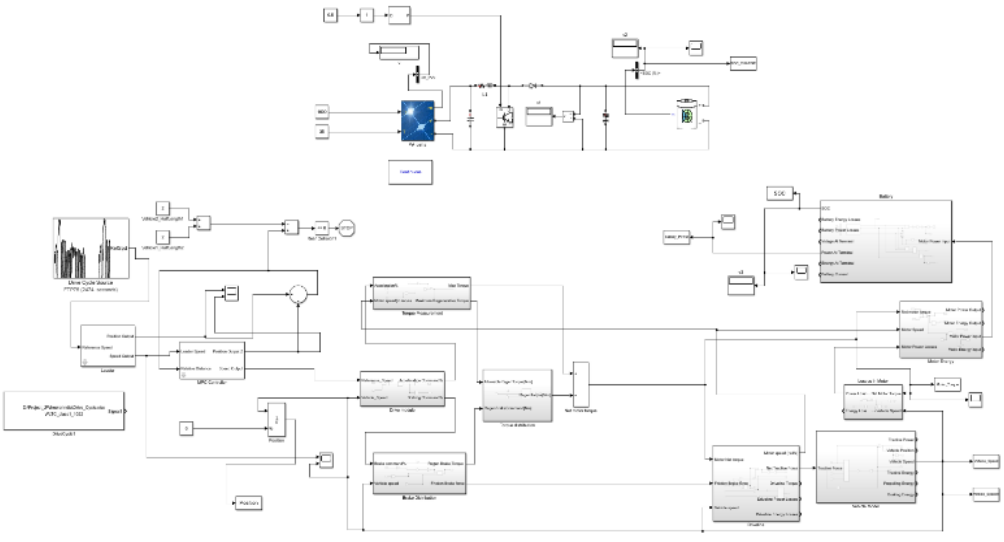


**Figure 13:** MATLAB Simulink Model

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 429-437 | e-ISSN: 2347-470X

## 4.3 Simulation Testing

The testing of the vehicle model was performed in 3 stages each for a different cycles and graphs were plotted for performance evaluation. For the first and the second stage the leader vehicle started 40 m ahead of the project vehicle with an initial velocity of 10m/s while the project vehicle started with an initial velocity of 5m/s. The initial battery SOC for 1st & 2nd stage was 95%. During the first stage the leader vehicle was subjected to the WLTC_class1_1022 Cycle. The simulation ran for 1022 seconds and the project vehicle's performed optimally for the whole cycle. During the second stage the leader vehicle was subjected to the FTP75 cycle. The simulation ran for 2474 seconds and the project vehicle's performed efficient for most of the cycle except for the top speed peaks where the vehicle was unable to cross the motor speed barrier. During the third stage with initial battery SOC at 50%, the PV module was subjected to constant solar irradiance of 1000W/m$^2$ and temperature 35$^o$C for 300 seconds. The module worked efficiently and the battery was charged. 5 major performance graphs of the vehicle were plotted over time for the WLTC_class1_1022 Cycle and the FTP75 cycle tests and compared. These were velocity plot, displacement plot, state of charge plot, battery power plot and motor torque plot. State of charge plot for the PV module test was derived.

### 4.3.1 Plot 1: Velocity Time Graphs

As shown in *Figure 14-15*, velocity plots clearly show the vehicle completes both the cycles efficiently. During WLTC cycle the vehicle attains a top speed of 17.9 m/s at around 770 seconds mark which is equivalent to the cycle. While for FTP75 cycle the vehicle attains a max speed of around 21.5m/s where as the top speed requirement for the cycle was 25.3m/s at the 241 seconds and 2211 seconds barriers. Other than this the project vehicle is able to follow the velocity profile of both the cycles while accelerating and braking. The majority time the velocity of vehicle is 8-12 m/s for WLTC cycle and average speed is higher during the second half while it is 10-15 m/s for FTP75 cycle and it seems constant throughout except for the stop at 1367 second till 1989 seconds.

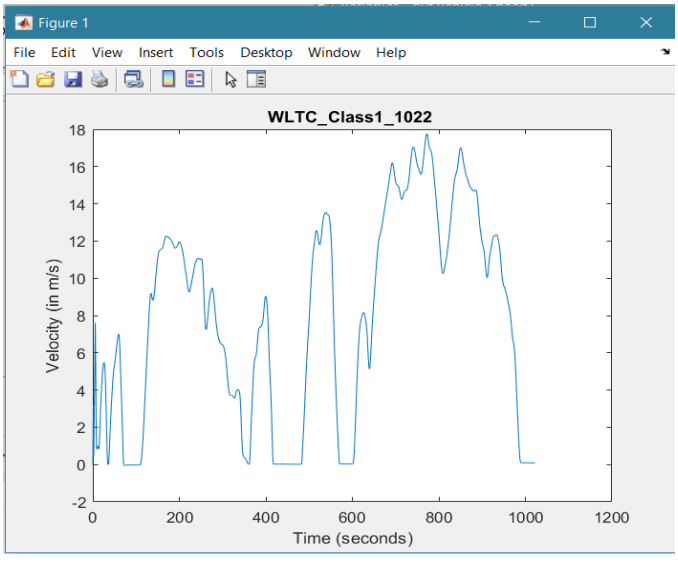



**Figure 14:** Velocity (WLTC 1022 Cycle)

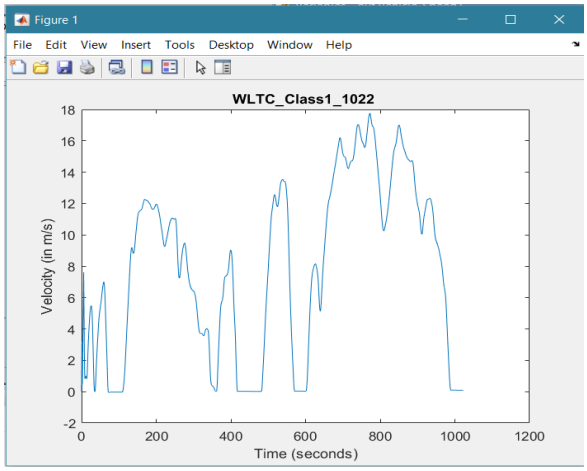



**Figure 15:** Velocity (FTP75 Cycle)

### 4.3.2 Plot 2: Displacement Time Graphs

As shown in *Figure 16-17* the vehicle completes the journey of both the cycles efficiently. The vehicle covers 8103 m following the WLTC cycle while it covers 17667 m following the FTP75 cycle. The slopes are analogous to vehicle speeds during these drive cycles. The vehicle constantly moves during the WLTC cycle while it stops at the 12km mark for about 10 minutes during the FTP75 cycle.

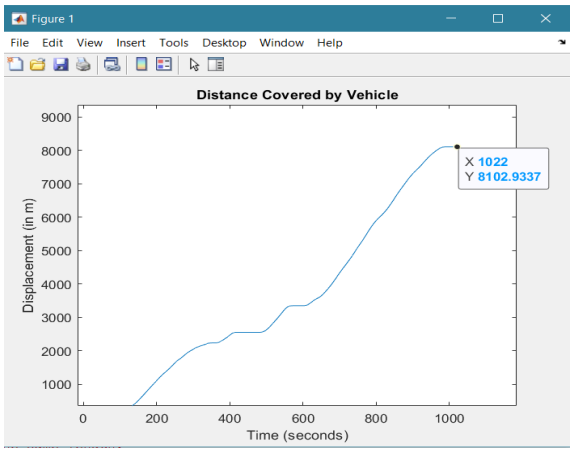



**Figure 16:** Displacement (WLTC 1022 Cycle)

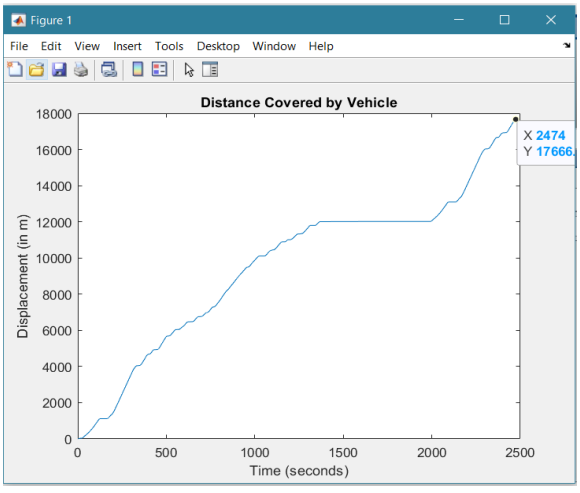



**Figure 17:** Displacement (FTP75 Cycle)

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 429-437 | e-ISSN: 2347-470X

### 4.3.3 State of Charge-Time Graphs

As shown in *Figure 18-19* the battery pack of the vehicle discharges during both the cycles. The WLTC cycle is shorter and there are lesser instances of braking and acceleration so battery discharges from 95% to 82.06% during the 1022 second cycle. The slope is relatively steeper after the 600 second mark due to higher speeds in the 2nd half of the cycle. The FTP75 cycle is relatively larger although the vehicle stops for a significant time after 1367 seconds for 622 seconds. The battery discharges from 95% to 62.6% during the 2474 second cycle. The slope is abruptly steep after the 2000 second mark after which vehicle attains higher speeds.

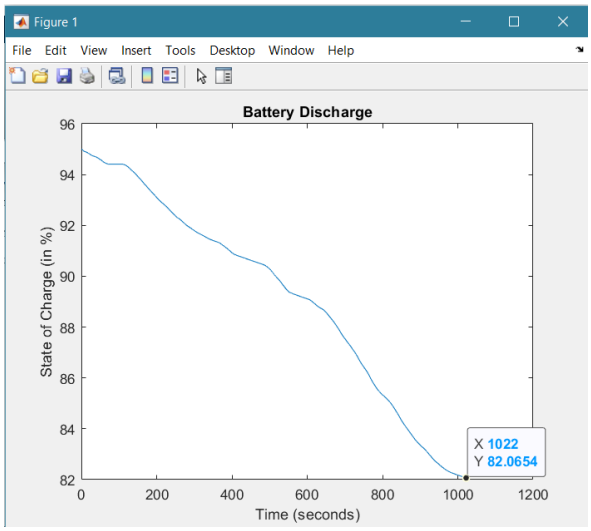



**Figure 18:** Battery SOC (WLTC 1022 Cycle)

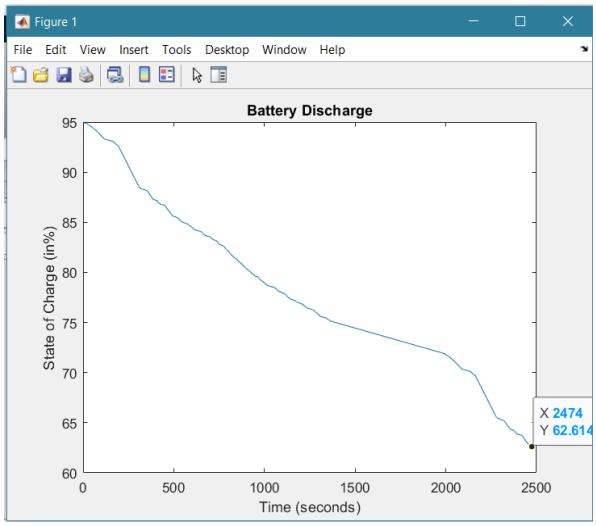



**Figure 19:** Battery SOC (FTP75 Cycle)

### 4.3.4 Plot 4: Battery Power and Time graphs

As shown in *Figure 20-21* the battery power follows a similar trend like the velocity profile. Same motor and battery pack have been subjected to both the cycles. During the WLTC cycle the power derived from the battery was mostly in the region 1800 – 2200 W and since the rated motor power is 2000 W it works efficiently. The power demand peaks at higher speeds to around 3000W but it is not for longer duration and can be sustained. During the FTP75 cycle the power derived from the battery is mostly in 2000 – 2300 W region which again proves that a 2000W rated power motor works efficiently for this cycle. The peak power demands more than 3700W although cannot be sustained due to limitations of motor. Overall the performance of the vehicle motor and battery pack is optimal.

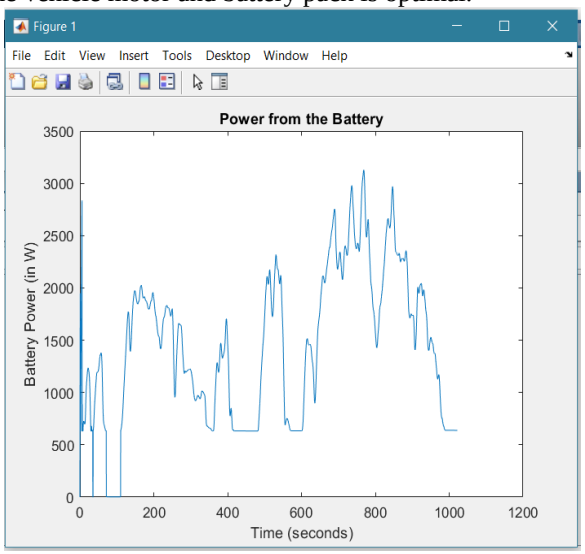



**Figure 20:** Battery Power (WLTC 1022 Cycle)

**Figure 21:** Battery Power (FTP75 Cycle)

### 4.3.5 Plot 5: Motor Torque and Time Graph

As shown in *Figure 22-23* the motor torque varies differently for both cycles. During the WLTC cycle the net motor torque mostly lies in the 8-11 Nm region. The motor works efficiently as the rated torque value is 12.5 Nm. There are less spikes in this cycle and the motor delivers the required torque without a load. During the FTP75 cycle however the required torque mostly lies in the 9-12 Nm region which is close to the rated value. There are spikes requiring larger torque of about 18-19Nm. This mainly due to intense acceleration and deceleration during the FTP75 cycle which is not the case of WLTC cycle. Overall the motor works close to its rated value and so performs optimally.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 429-437 | e-ISSN: 2347-470X

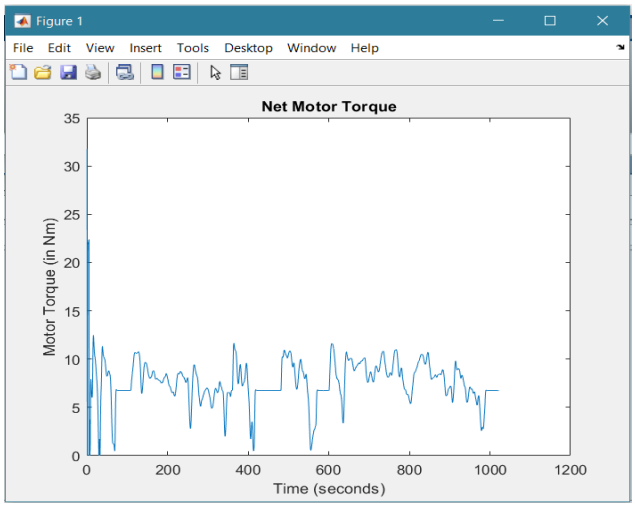



**Figure 22:** Motor Torque (WLTC 1022 Cycle)

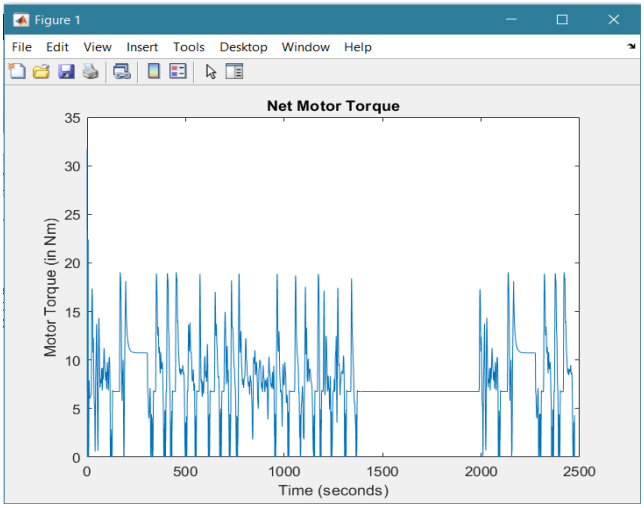



**Figure 23:** Motor Torque (FTP75 Cycle)

### 4.3.6 Plot 6: State of Charge and Time Graph

As shown in *Figure 24* during the third stage testing the PV module was put under solar irradiance 1000W/m$^2$ and temperature 35$^o$ for 300 seconds. The battery pack charged from 50% to 50.54%.

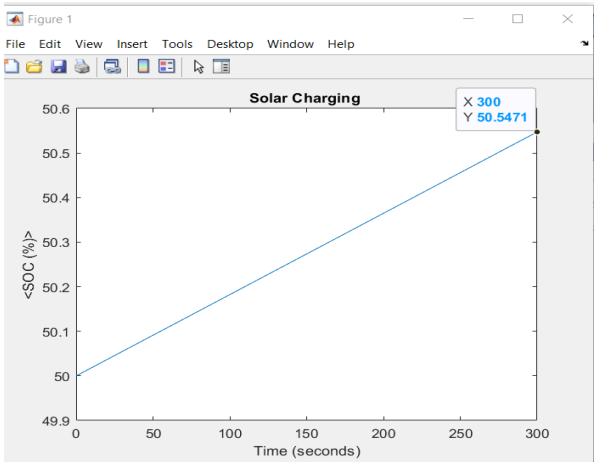



**Figure 24:** Solar Charging over 300 seconds

The net vehicle weight was around 80 kg so it was able to carry a load of 20kg effectively. The vehicle travels around 8km dropping 13% battery so the vehicle has a range of about 60-70 km on such extreme cycles. The vehicle ran smoothly with the PID driver model running efficiently. The attached solar panel was 1.96m$^2$ and it charged the battery from 50% to 50.54% in 300 seconds i.e. 5 minutes making it a viable source to charge the battery in areas with extreme sunny days. Major battery losses, driveline losses, motor losses were incorporated in the simulation making the vehicle fit for real world applications. The vehicle was inspired by vehicles designed for World Solar Challenge and used modern innovative subsystems like Li-ion cells, carbon fiber chassis, PV modules, regenerative braking etc.

## 5. CONCLUSION AND FUTURE SCOPE

Autonomous vehicles continue to be a distant reality. Although recent advances in AI and neural networks can turn this into a reality. Similarly, climate change is the biggest world issue and it's a global agreement that we need better energy alternatives and stop burning fossils. Solar energy is the cleanest and one of the most sought after sustainable alternatives to oil. India is presently going through an electric revolution in the automotive sector. Major reforms, incentives, schemes, startups and laws are forming up a new ecosystem for development of Indian automotive sector. Along with this the due to the rise of e-commerce, last mile freight deliveries have seen a massive explosion. The main objective of this particular project is to combine these all developments to provide a novel futuristic solution. The idea behind developing this model is to solve multiple problems together optimally. The approach of the project uses latest technologies like computer aided designing, mathematical modelling and simulation testing. The objectives of the project were based on providing one solution to four problems together. These problems were putting autonomous vehicles under safe conditions, using a renewable source for transforming mobility, providing the last mile deliver sector a sustainable alternative amid rising fuel prices and being a part of India's E-mobility revolution The project vehicle performed optimally and fulfilled all its objectives. Such solar electric vehicles have potential to replace small delivery vehicles for last mile deliveries. This project vehicle pod has been thoroughly tested for different cycles and has proven efficient in simulation testing. The vehicle has been constructed using off the shelf parts and it fulfilled the major objectives of the study like range, charging and speed criteria satisfactorily The vehicle is simple, lightweight and easy maintenance. It can be easily dismantled and parts like PV module and battery be used for other purposes. The vehicle is created using a mathematical model so different parts may be tested for the same model. Such model can be used to check the effectiveness of such vehicles on different drive cycles. The further steps include making the vehicle structurally stronger and performing major safety tests like impact fatigue and torsional tests. The present vehicle deals with only longitudinal dynamics so adding lateral motion and reforming the self-driving module is also due.

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 429-437 | e-ISSN: 2347-470X

## REFERENCES

[1] Mamta L Prajapati et al. 2020 IOP Conf. Ser.: Mater. Sci. Eng. 937 012033

[2] Mohammad Waseem et al. SN Applied Sciences (2019) 1:1459 | https://doi.org/10.1007/s42452-019-1458-4

[3] J. Renewable Sustainable Energy 12, 053701 (2020); https://doi.org/10.1063/1.5140236

[4] S. Iwan et al. Sustainable Cities and Society 74 (2021) 103167; https://doi.org/10.1016/j.scs.2021.103167

[5] R.A. de Mello Bandeira et al. Transportation Research Part D 67 (2019) 491–502; https://doi.org/10.1016/j.trd.2018.12.017

[6] S.M. Mirhedayatian, S. Yan Transportation Research Part D 58 (2018) 22–38; https://doi.org/10.1016/j.trd.2017.11.007

[7] Muhammad Hosnee Mobarak et al.; DOI 10.1109/TTE.2020.2996363, IEEE Transactions on Transportation Electrification

[8] Christophe Rizet et al. / Transportation Research Procedia 12 (2016) 500 – 507; doi: 10.1016/j.trpro.2016.02.005

[9] Tharsis Teoh et al. Sustainability 2018, 10, 3258; doi:10.3390/su10093258

[10] V.S. Patyal, R. Kumar and S. Kushwah Energy 237 (2021) 121554; https://doi.org/10.1016/j.energy.2021.121554

[11] Ashish Kumar, Lal Bahadur Prasad; 2018 International Conference on Computing, Power and Communication Technologies (GUCON) Galgotias University, Greater Noida, UP, India. Sep 28-29, 2018; 978-1-5386-4491-1/18/$31.00 ©2018 IEEE

[12] Bambang Sri Kaloko et al. International Journal of Computer Applications (0975 – 8887) Volume 24– No.6, June 2011

[13] J. Ashok, Pradnya P. Mirajkar & Archana L. Lakesar; International Journal of Mechanical and Production Engineering Research and Development (IJMPERD) ISSN (P): 2249-6890; ISSN (E): 2249-8001 Vol. 8, Issue 4, Aug 2018, 133-146 © TJPRC Pvt. Ltd.

[14] Zahari Taha et al. Proceedings of the IASTED International Conference Solar Energy (SOE 2009) March 16 - 18, 2009 Phuket, Thailand

[15] Jennings, D. and Figliozzi, M., 2020. Study of road autonomous delivery robots and their potential impacts on freight efficiency and travel. Forthcoming Transportation Research Record.

[16] Paterson et al. Proc IMechE Part D: J Automobile Engineering; DOI: 10.1177/0954407016630153

[17] Lg M50 Rechargeable Battery, For Power Saver, Size: 21700 - PDF Catalogue (indiamart.com)

[18] Ceat Zoom XL 80/100-17 46P Tubeless Bike Tyre, Rear : Amazon.in: Car & Motorbike

[19] TVS Tyres Jaya 4.00-8 Tube-Type Auto Tyre : Amazon.in: Car & Motorbike

[20] Buy Luminous 330W 24V Polycrystalline Solar Panel, LUM 24330 Online At Price ₹10599 (moglix.com)

[21] BLDC Motors – Tech Imperial (techimperial.in/bldcmotors)

© 2022 by the Abhishek Arora and Janardan Kesari. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Detection and Analysis of Flooding Attacks in Wireless Sensor Networks

**Gebrekiros Gebreyesus Gebremariam** ( kiros2004comp@gmail.com )
Delhi Technological University

**Jeepanda Panda**
Delhi Technological University

**S. Indu**
Delhi Technological University

**Research Article**

**Keywords:** WSN security attacks, Flooding attacks, Sensor nodes, malicious nodes, Performance Metrics, Message delivery

**Posted Date:** September 20th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-931986/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

# Abstract

Wireless sensor networks are distributed networks and randomly deployed in harsh environments. The scattered nature of the sensor nodes in an unattended environment makes them vulnerable to various types of attacks. Multiple attacks like flooding attacks degrade the network performance and lifetime by sending various requests to malicious nodes. This paper proposes the detection and analysis of flooding attacks in mesh wireless sensor networks using the Blue tooth communication package MATLAB 2020a toolbox. The flooding attack disabled the relay nodes during the routing path between sources and destination nodes for each simulation scenario. The simulations are conducted by the number of sensors variation for the random generator. After simulating the meshed nodes, the statistical results are the message delivery chance, average hope count, and critical relays. The opportunity of message delivery having a route between sources and destinations is 88.64%. The effectiveness of the proposed system for detection and analysis of flooding attacks is evaluated by random forest machine earing techniques with an average detection rate and accuracy of 99.83%, which is greater than by 3.8%.


# 1 Introduction

Wireless sensor networks (WSNs) are the network of sensor nodes with limited power processing, storing capacity and battery [1]. They are deployed with static or dynamic frame work with limited resources and mobility [2]. Wireless sensor nodes can be homogeneous or heterogeneous depending on power consumption, computational processing, communication range, hardware utility and cost. WSNs are used in various applications including environmental monitoring, military, healthcare, industrial, Vehicle cloud, smart grid, surveillance and home automations [3]. The sensor nodes collect and broadcast information to the base station from the surrounding. The open nature and changing topological architecture deployed in hostile environment leads the interrupts of the legitimate nodes. Wireless sensor networks are exposed to various attacks including flooding attacks, Denial of Services, blackhole attacks, etc. The flooding attack wastes the network resources such as the bandwidth and battery power by sending multiple routing request to the destination using multihop routes. The flooding attack targets routing protocols and exhausting the energy of the nodes [4].

WSNs are exposed to many types of attacks and threats that degrade the quality of services by injecting a malicious node to the network. The intruder in the network launch a false routing path, creates a false message, and compromises sensor node violating the network goals. The attacks are categorized into active and passive that damage the network. The active attacks exploit and project security protocol for packet modification and traffic analysis. While the passive attacks access the information and make disclosure the secure information. These type of attacks are difficult to detect like wormhole and gray hole attacks.

# 1.1 Flooding Attacks

Flooding attacks are network layer DoS attacks at the network layer. It has critical damaging effect on the network due the limited resources of the sensor nodes [3]. Flooding attack also affects the network performance and the routing protocols. These protocols use sequence numbers to detect the routing information and to ensure loop free routing. It possesses the following control routing messages including

request, reply and error. The flooding attack broadcasts multiple routing request packets in the presence of AODV routing protocol with malicious packets acting like normal traffic [5] as shown in Fig. 1. Flooding attacks disrupt the access of the network resources by flooding unnecessary network traffic in the network targeting the legitimate node. The attacker sends hello packets to the neighbor node making itself as authentic node which wrongly identified [6].

Wireless sensor networks are exposed to traffic flooding since they do not have packet filtering device firewall [7]. The packet could be originated from any source node containing the flooding attack. The flooding attack exploits the routing information and make the authorized node busy by sending multiple requests. The resource in the destination node are drained and bandwidth also reduced. The flooding attack can send the requests such as joining, data and leaving with multiple malicious attacks [8]. The resources in WSNs are depleted and functionally degraded by targeting the different layers of the network [9]. The flooding attack exhausts the memory, bandwidth, energy and computational processing of the legitimate node leading to denial of service. The DoS attacks drains the sensor node power supply and makes dead which affects the over performance and life time of the network. The DoS attacks send flooding requests making the server in accessible by using multiple hosts. These attacks include false node, message corruption, node replication and eavesdropping. These attacks deplete the resource in the network and exposed the normal node in all layers of the network [10]. Authorized users are restricted from using the network access.

The attacker broadcasts malicious hello packets to all the nodes in the network with enough power. It replicates and broadcasts itself as a root node the hello packet to all the nodes for communication [11]. The protocol authenticates and raises alarm for detection of the attacks in the network.

All the authorized nodes misidentify to detect the malicious node and cause all the nodes to send false packets to the destination. The flooding attack sends multiple fake packets consuming much resources to the destination [12].

The contribution of this work is to explore the analysis of flooding attacks in wireless sensor networks using a meshed sensor nodes. Some of the main contributions are bullet as follow:

- Creating and configuring wireless meshed network
- Analyzing and visualizing the flooding of routing information
- To drive the routing selected routing path source and destination pair of nodes.
- Enables to visualize the node statistics at the network layer for each node
- Analyze the probability of routing information at the destination node
- Explore the detection accuracy of the flooding attacks in meshed wireless sensor networks topology
- Investigate machine learning techniques for flooding attack detection and classification

# 2 Literature

Various related works were reviewed on flooding attacks in wireless sensor networks using routing protocols and detection schemes. They also highlighted the various techniques for detection and prevention of flooding attacks.

H. Moon et al. [1] examined analysis of flooding attacks using various simulation parameters and network performance. A. Sana et al. [3] presented the effect of flooding attack in wireless sensor networks using AODV protocol. They also discussed the performance of the network using simulation parameters. P. Mohammadi and A. Ghaffari [5] discussed detection and defending techniques of flooding attacks using dynamic source routing protocol on the network layer of the network. They also presented the average transmission of routing request packets for the network performance. S. Gurung and S. Chauhan [12] Examined mitigation of flooding attacks using a dynamic threshold value containing three phases. The scheme also has a node with flooding intrusion detection system for MANETs for prevention and detection of flooding attacks. This special node monitors the node status and improves the network performances and normalizes the routing load. V. Moorthy et al. [13] Presented the analysis of flooding attacks in software defined WSNs with communication protocols including in the topology TCP and UDP. They also studied the causal effect and probability of the DoS scenario over the time frame using communication protocols. The Bayesian network for reliable communication is studied using the trust analysis. S. E. Quincozes and J. F. Kazienko [14] examined the performance of machine learning classification techniques for detection of flooding, blackhole and gray hole attacks in wireless sensor networks. They also discussed the evaluation metrics of the detection accuracy using WSN dataset. M. Agarwal et al. [15] Proposed machine learning based intrusion detection and prevention system for detecting and analyzing DoS flooding attacks for Wi-Fi networks. They also discuss detect and localize flooding attack in network with evaluation metrics including detection rate and precision. J. David and C. Thomas [16] presented the analysis the network traffic features for detection of distributed denial of service attacks using statistical technique. They also discussed the extraction and characteristics of the attack using public dataset as a bench mark for performance evaluation. Table 1 shows the summary of the various research works for detection of flooding attacks. B. D. Patel and A. D. Patel [17] reviewed network layer and transport layer routing attacks in wireless sensor networks. They also discussed prevention and detection of wormhole and black hole attacks using trust base techniques. G. Kocher and G. Kumar [18] Conducted a systematic literature review on deep and machine learning techniques for attack detection using benchmark datasets. They also discuss the deep learning techniques for intrusion identification.

# 3 Network Model

The network model consists of meshed sensor nodes: relay nodes, malicious nodes, end nodes, low sensor nodes, aggregation nodes, and base stations [19]. All the sensor nodes are assumed to be heterogeneous with different processing capabilities in the proposed model as shown in figure 2. Failed node is regarded as suspected node and malicious node. The end nodes collect information from the physical environment. The end node are the sources nodes and the cluster head and base stations act as destination nodes. The relay nodes select the routing path and the flooding routing information between source destination pairs of nodes. The network model consists of Bluetooth stack protocols including attribute protocol, security manager protocol, generic attribute profile, adaptation protocol, logical link control, physical and link layer protocols [20]. This network model is applicable for security of home automation, internet of things and health care and fitness. They are also applied for industrial automations and asset tracking applications in wireless senor networks in security is vital.

# 4 Proposed Flooding Attacks Detection System

The proposed approach for detection and analysis of the flooding attacks in wireless sensor networks show in Fig. 3 with the main blocks. These include the dataset pre-processing, training samples, testing samples, detection and classification of the network traffic as normal and malicious data. The data processing step is very important step for the quality of data. The pre-processing consists of cleaning, integration, aggregation, transformation, normalization, reduction and discretization of the data for training and testing of the samples.

# 4.1 Dataset

The CIC-IDSS2017 dataset is used as a benchmark for evaluating the performance system for detection of the flooding attacks in WSNs. The dataset is available online on the Canadian Institute for cyber security research LAB. The dataset contains benign and some other attacks that include network traffic data [21]. It was generated for realistic background traffic for building the dataset. The dataset was created using 25 users with different protocols. The dataset contains 485881 Instances as shown in Tables 2 and 31 attributes with 80% of training and 20% of testing. There are five classes of denial of service (DoS) attacks including Slowhttptest, slowloris, Hulk, Heartbleed and GoldenEye [22]. The structure of the dataset is shown below in Table 1.

Table 1
Structure of the dataset with classes of attacks

| Class of attacks | Normal | slowloris | DoS Slowhttptest | DoS Hulk | DoS GoldenEye | Heartbleed |
|---|---|---|---|---|---|---|
| Training set | 252382.6 | 1504.8 | 1276.8 | 127302.4 | 6307.2 | 8 |
| Testing set | 63076.4 | 376.2 | 319.2 | 31825.6 | 1576.8 | 2 |
| Total | 315382 | 1881 | 1596 | 159128 | 7884 | 10 |

The data processing is used for rearrangements using the normalization, missing value imputation and aggregating for training and testing phases. The missing values are normalized with the average values of the present values [23].The minimum and maximum values are utilized for transforming the data into binary values as 0 and 1.

Table 2
The statistical distribution of selected features of the dataset

| Features | Number | Mode | Mean | Variance | Lower Limit | Upper Limit |
|---|---|---|---|---|---|---|
| DestinationPort | 485881 | 4855.64 | 4855.64 | 435.573 | 4814.73 | 4896.54 |
| FlowDuration | 485881 | 38856281 | 38856281 | 4.42E + 09 | 38725978 | 38986585 |
| TotalFwdPackets | 485881 | 12.6 | 12.6 | 1.638 | 10.09 | 15.11 |
| TotalBackwardPackets | 485881 | 14.39 | 14.39 | 2.842 | 11.09 | 17.7 |
| TotalLengthofFwdPackets | 485881 | 766.63 | 766.63 | 110.836 | 746 | 787.27 |
| FwdPacketLengthMax | 485881 | 316.83 | 316.83 | 1.003 | 314.87 | 318.8 |
| FlowBytess | 485771 | 387017.1 | 387017.1 | 50052397 | 373150.8 | 400883.4 |
| FlowPacketss | 485771 | 20427.24 | 20427.24 | 31564.76 | 20079.02 | 20775.45 |
| FlowIATMax | 485881 | 31827784 | 31827784 | 3.69E + 09 | 31708659 | 31946909 |
| FlowIATMin | 485881 | 49418.34 | 49418.34 | 9698378 | 43314.58 | 55522.11 |
| FwdIATTotal | 485881 | 38499251 | 38499251 | 4.44E + 09 | 38368713 | 38629788 |
| FwdIATMax | 485881 | 31702017 | 31702017 | 3.71E + 09 | 31582677 | 31821357 |
| BwdIATTotal | 485881 | 19764440 | 19764440 | 3.02E + 09 | 19656661 | 19872218 |
| BwdIATMax | 485881 | 13265776 | 13265776 | 1.9E + 09 | 13180325 | 13351227 |
| BwdIATMin | 485881 | 1321377 | 1321377 | 1.92E + 08 | 1294232 | 1348522 |
| FwdHeaderLength_A | 485881 | 318.01 | 318.01 | 717.945 | 265.49 | 370.52 |
| BwdHeaderLength | 485881 | 350.67 | 350.67 | 1137.461 | 284.56 | 416.77 |
| FwdPacketss | 485881 | 15533.1 | 15533.1 | 26639.68 | 15213.2 | 15853 |
| BwdPacketss | 485881 | 4910.09 | 4910.09 | 2117.417 | 4819.902 | 5000.279 |
| MaxPacketLength | 485881 | 2435.6 | 2435.6 | 16.862 | 2427.55 | 2443.65 |
| AveragePacketSize | 485881 | 415.2116 | 415.2116 | 0.367 | 414.024 | 416.3992 |
| AvgFwdSegmentSize | 485881 | 74.81795 | 74.81795 | 0.064 | 74.32138 | 75.31453 |
| FwdHeaderLength | 485881 | 318.01 | 318.01 | 717.945 | 265.49 | 370.52 |
| SubflowFwdPackets | 485881 | 12.6 | 12.6 | 1.638 | 10.09 | 15.11 |

| Features | Number | Mode | Mean | Variance | Lower Limit | Upper Limit |
|---|---|---|---|---|---|---|
| SubflowFwdBytes | 485881 | 766.63 | 766.63 | 110.836 | 746 | 787.27 |
| SubflowBwdPackets | 485881 | 14.39 | 14.39 | 2.842 | 11.09 | 17.7 |
| Init_Win_bytes_forward | 485881 | 6069.22 | 6069.22 | 310.508 | 6034.69 | 6103.76 |
| Init_Win_bytes_backward | 485881 | 1388.02 | 1388.02 | 98.666 | 1368.55 | 1407.48 |
| act_data_pkt_fwd | 485881 | 8.51 | 8.51 | 1.501 | 6.11 | 10.91 |
| min_seg_size_forward | 485881 | 26.26 | 26.26 | 0 | 26.24 | 26.27 |

# 4.2 Machine Learning Classifiers

The proposed system uses machine learning classification techniques for attack detection and classification as normal and attack using some evaluation metrics using the dataset CIC-IDSS2017 as a benchmark. Some of the machine learning algorithms for training and testing the dataset are Random Forest, Bagging, AdaBoost, J48, REPTree and multilayer Perception. They are used for data classification and performance analysis.

# 4.2.1 Random Forest

The random forest (RF) is a supervised machine learning technique which has a set of branch trees and each tree gives a classification in the forest [24]. RF builds a forest using several decision trees combining the average prediction trees which is appeared robust [18]. The random forest is the best function for building the better model due to its appropriate features for regression and classification tasks. It overcomes the over fitting problems in machine learning. The random forest has the advantages of low training time complexity, ranking features, rapid prediction and flexible for unbalanced datasets [25]. It has stages of branches for trees prediction and classification of attacks. It is effective for predicting missing values for large heterogeneous datasets. The RF classifier is best for classification of hyperspectral data in wireless sensor network. It creates random decision trees and combines predictors with randomly distributed vectors [26]. The predicted values are calculated for each trees. Random forests are bagged decision trees generated using the procedures shown below:

- CIC-IDSS2017,n, datasets F1…,$F_n$ form with replacement.
- Train a decision classifier $H_i()$ for each $D_i$ to the maximum k features.
- Compute the ensemble classifier the mean output as in Eq. (1) :

$$H(x) = \frac{1}{n}\sum_{k=1}^{n} H_i(x) \quad (1)$$

# 4.2.2 Multilayer Perception

Multilayer perception (MLP) is an artificial neural network structure used to detect and classify the various classes of attacks using the network traffic data characteristics as normal and abnormal [27]. It is effective for the estimation of classification and prediction techniques using backpropagation [28]. MLP consists of the input, hidden, and output layer arrangements for computing the elements. MLP is an ANN supervised machine learning technique used for predicting flooding attacks [18]. It is used for DoS attack traffic data classification and achieves highest accuracy for flooding attacks.

## 4.2.3 Bagging

Bagging is a machine learning ensemble classification technique effective for measuring the performance of the systems by voting the majority class when there is unfamiliar attributes in the dataset [29]. Bagging computed with decision tree classifier for reducing errors in the input variance. Bagging is a meta-data machine learning scheme used to improve regression and classification techniques [30]. It combines several machine learning techniques and generates accurate predictions. Bagging reduces and avoids over fitting variance for building robust model. Bagging is designed for multi classification systems.

## 4.2.4 Adaptive Boosting

Adaptive Boosting (AdaBoost) is a booting technique for classifying a set of samples into wrongly classified iterations. Whereas correctly classified samples are attenuated [29]. AdaBoost is a binary classification booting technique built by a classification model using the training samples in each iteration [30]. It is the combination of weak classifiers for creating a solid classifier and predicting the model. The AdaBoost scheme is used for feature selection and classification tasks to learn the trained positive and negative observations. It generates a single complex prediction in conjunction with other classifiers to develop rude logic using training data to improve performance.

## 4.2.5 J48 Classifier

This algorithm is a decision tree classifier technique that consists of an open-source java implementation constructing the trees from the training dataset [14]. It predicts the classes of unknown samples using the entropy concept. The J48 is a data mining tool [31] and a top-down strategy for making decision trees. The J48 technique is implemented by labeling from the training dataset. It follows the continuous divide and conquer method for making the decision tree [32]. The partitioning of the scheme for the decision tree begins with the training of the dataset at every node. It assigns an appropriate class label in the dataset and uses a normalized version of the entropy to measure the data impurity.

## 4.2.6 REP Tree

The reduced error pruning tree (REPTree) classifier works based on the principle of information gain and calculates the entropy for minimizing the error during the classification process. It is implemented using the fast decision tree scheme. The RepTree identifies the optimal decision in the decision tree and removes the noise during the construction and generation of the branches [25]. It checks the internal nodes duplication and pruned it for better decision tree accuracy. The RepTree acts as a classifier by pinpointing the attack for recognition. It is the fastest decision tree learning algorithm for building a model and prunes for reducing error pruning. It uses the concept of information for measuring the quality of information [33].

# 5 Simulation And Testing

In this section, the simulation environment of wireless sensor nodes is analyzed against the network constraints such as an end-to-end delay, packet loss, and throughput under the various scenario using MATLAB 2020a simulator. The simulations are conducted with varying number of sensors 50–55 under the area 160 m x 200 m, as shown below in Table 3. The blue tooth package is used for simulating the meshed sensor networks to visualize and analyze the flooding of messages.

Table 3
Simulation parameters and their values for experimental setup

| Parameters | Values |
|---|---|
| Topology | Mesh |
| Area (meter square) | 160x200 |
| Number of nodes | 50–55 |
| Protocol type | Security manager Protocol |
| Number of scenarios | 2 |
| Simulation time | 600 milliseconds |

The Bluetooth communication Toolbox library performs the Monte Carlo simulations on the meshed sensor [20] networks for obtaining the numerical results for the mesh wireless sensor networks for several iterations. Monte Carlo simulations [34] are examined on meshed Bluetooth networks with simulation parameters to analyze the network data traffic. The possibility of packet delivery is modeled by the node failure assuming the malicious activities of the attack on the network.

# 6 Experimental Result And Analysis

In this section, the two simulation scenarios and the data analysis will be discussed in detail. All the nodes are healthy and active without malicious nodes in the network. The plots are shown below in Figs. 4 and Fig. 5 for each source and destinations pairs of paths, respectively.

The experiment is conducted on DESKTOP-NDN9H5A environment, Intel(R) Xeon(R) Silver 4214 CPU @ 2.20GHz 2.19 GHz (2 processors), 128 GB (128 GB usable), and 64-bit operating system using MATLAB R2020a communication Toolbox for network planning and simulation. The data processing is done using machine learning techniques built-in WEKA Toolbox java programming language.

The simulated results of the paths are stored in the scenario one results are shown in Table 4 as raw of matrix. The number of hops for each pair is different and stored as an array of the matrix.

Table 4
Stored simulation results for scenario one with all normal modes

| Source | Destination | Path | Number Of Hops |
|---|---|---|---|
| 13 | 29 | [1x7 double] | 6 |
| 1 | 7 | [1x9 double] | 8 |

The network parameters are changed for testing the alternating paths among source-destination pairs of nodes. In this simulation scenario, the features of the relay node are disabled at node 41 to show the intruder is attacking this node in this work. The nodes are ignored from the network during the routing. Figure 5 shows the highlighting for each node path of source to destination pair. The simulated results of the routing information are stored in scenario as shown in Table 5 containing the number of hops as raw matrix of data.

Table 5
Stored simulation results for scenario two with all normal modes

| Source | Destination | Path | Number Of Hops |
|---|---|---|---|
| 13 | 29 | [1x8 double] | 7 |
| 1 | 7 | [1x11 double] | 10 |

The statistics of captured data in the network layer for each node contains the number of messages transmitted, received, application of Rx messages, relayed messages, and dropped messages in each node. The network information at each node contains the cumulative statistical data for all nodes to both scenarios. The numerical results are obtained by using the Monte Carlo scheme and modeled by the node failures as malicious nodes. The numerical results are also used for the analysis of message delivery of multiple simulations.

The relay nodes are deactivated arbitrarily for simulation schemes between the source and destination pairs. These simulations are conducted by changing the number of sensors for the random generator. The routing information is stored as raw data conforming to the path in the simulation scenario. The statistical simulation results are the rate of message transmission, average hope count, and critical relays. The probability of message delivery having a route between sources and destinations is 88.64%. The simulation results show the flooding activity of the two simulation scenarios in which the existence of the malicious node affects the network performance [1]. The flooding attack drops packets between the relay and end node that affect the routing path.

The energy profile of simulation for the wireless mesh sensor networks of the nodes end nodes, relay nodes, friend nodes, and low power nodes is configured in Fig. 6. All the nodes will receive the mesh message along the routing path and group. Friendship messages are degraded between low-power and configured friend

nodes. Each node computes the time consumed in various states to calculate the node's lifetime in the network.

The nodes calculate the time consumed in different conditions within the mesh wireless sensor networks. The results from the simulation show that transmission of a single message from the sources node to the destination node. The application traffic affects the transmission time at the destination node.

# 6.1 Performance Analysis of Classification Models

The assessment of the computational performance of the technique is based on normal and attacks samples using the dataset as a benchmark. The implementation of the proposed scheme is computed using a confusion matrix consisting of true negative rate, true positive rate, false-positive rate, and false negative. The confusion matrix contains the actual and predicted classifications showing the effectiveness of the classification models. There are also other robust performance metrics for the proposed system are bullet as f-measure, precision, recall, Matthews Correlation Coefficient (MCC) [35], receiver operating characteristics (ROC) Area, and Precision recall curve (PRC) Area for each class of attacks in the network. The mathematical equations for the metrics are depicted as follow to determine the proportion of attack detection and classification.

Where Tp is the true positive rate and Fp is the false negative rate. A Tp detects the attacks declared by the system as malicious nodes. Fp is the alarm that the system detects the normal traffic as malicious traffic.

The detection rate describes the number of attacks identified by the intrusion detection system out of the total malicious traffic data present in the dataset. Detection rate is the portion of malicious data classified as attacks and it is the true positive rate (TPR) [36]. It is determined by the confusion matrix and computed as Eq. (2):

$$\text{Recall} = \text{Detection rate} = \frac{T_p}{T_p + F_n} \quad (2)$$

Where Fn is the false positive. It is the IDS treatment for the attack behaviors as normal. The recall is the completeness of the results [37].
False positive rate (FPR) is proportion of the normal traffic which are misclassified as attacks and computed by Eq. (3):

$$\text{FPR} = \frac{F_p}{T_n + F_p} \quad (3)$$

Precision is the ratio of the normal class sample to the total number of traffic data in the dataset.

$$\text{Precision} = \frac{T_p}{T_p + F_p} \quad (4)$$

Accuracy (Ac) describes the number of attacks properly identified out of the total number of attacks illustrated by equation (5) [15]. It is the measure of the performance of the system with correctly detected

attacks. It is defined as the data correctly identified in percentage from the entire set of samples.

$$Ac = \frac{T_n + T_p}{T_n + F_n + F_p + T_p} \quad (5)$$

The accuracy is the measure of the performance of the system with correctly detected attacks. It is defined as the data correctly identified in percentage from the entire set of samples. Another performance F-measure is the combination of the recall and precision computed as using equation (6).it is the harmonic mean of recall and precision for estimating the accuracy of the proposed system.

$$F\text{-measure} = \frac{2xT_p}{2xT_p + F_p + F_n} \quad (6)$$

MCC is a balanced measure of the quality of binary classification in machine learning, taking the true positive and false positives [38]. It is computed using equation (7).

$$MCC = \frac{\frac{T_p}{N} S \times P}{\sqrt{PS(1-S)^2 + (1-P)^2}} \quad (7)$$

Where $N = T_n + T_p + F_n + F_p$, $P = \frac{T_p + F_p}{N}$, $S = \frac{T_p + F_n}{N}$

The CPU/GPU and memory consumption are taken into consideration for the effectiveness of the proposed IDS. They are secondary parameters and implementation issues that do not affect detection rate and accuracy.

The effectiveness of the proposed system is evaluated using machine learning techniques. Table 6 shows the overall performance of the six schemes using the CIC-IDSS2017 dataset as a benchmark with 10-cross validation folds. The performance of the system is evaluated by parameters Speed, time to build the model (TTB), Relative mean square error (RMSE), Relative root square error (RRSE), Relative absolute error (RAE), Mean absolute error (MAE), Correctly classified (CC) and incorrectly detected (ICC) and Detection accuracy. The predicted values on the test samples are assumed to be X1, X2, …, Xn; the actual values are Y1, Y2,…, Yn [39]. The Xi is the numeric test samples for prediction. The MA error, as shown in Table 6, is the mean of the individual errors ignoring the sign it is computed as:

$$MA\ Error = \frac{|X_1 - Y_1| + \ldots + |X_n - Y_n|}{n} \quad (8)$$

Where X denotes the predicted values, Y is the actual values, $\bar{Y}$ is the absolute mean values, and n is the number of variables. The root means the square error is computed by the Eq. (9) for reducing the dimensionality of the predicted quantity given below:

$$\text{RMSE} = \sqrt{\frac{(X_1 - Y_1)^2 + \ldots + (X_n - Y_n)^2}{n}} \quad (9)$$

The relative errors are essential for showing the misconception about the actual values. The average normalizes the errors predicted values, which are calculated by the absolute error given by Eq. (10):

$$\text{Relative absolute error} = \frac{|X_1 - Y_1| + \ldots + |X_n - Y_n|}{|Y_1 - Y| + \ldots + |Y_n - Y|} \quad (10)$$

The root relative squared error (RRSE) is computed by the equation (11) shown below.

$$\text{RRSE} = \sqrt{\frac{(X_1 - Y_1)^2 + \ldots + (X_n - Y_n)^2}{\left(X_1 - Y\right)^2 + \ldots + \left(X_n - Y\right)^2}} \quad (11)$$

Table 6 The comparison performance of the summary for each classification Models

| Evaluation of Instances | Random Forest | Multilayer Perception | Bagging | AdaBoostM1 | J48 | REPTree |
|---|---|---|---|---|---|---|
| CC | 99.83% | 98.184% | 99.82% | 94.61% | 99.82% | 99.80% |
| ICC | 0.171% | 1.82% | 0.1832% | 5.38% | 0.18% | 0.1961% |
| Kappa statistic | 0.9964 | 0.9609 | 0.9961 | 0.8833 | 0.9962 | 0.9958 |
| MAE | 0.0007 | 0.0081 | 0.0008 | 0.0667 | 0.0007 | 0.0008 |
| RMSE | 0.0193 | 0.0726 | 0.0197 | 0.137 | 0.0204 | 0.0211 |
| RAE | 0.4423% | 5.1817% | 0.4849% | 42.4763% | 0.4574% | 0.4972% |
| RRSE | 6.8817% | 25.9023% | 7.0443% | 48.9029% | 7.2724% | 7.5322% |
| TTBM (seconds) | 323.75 | 576.07 | 227.2 | 32.12 | 66.07 | 26.39 |

The random forest classifier has greater detection accuracy with the same cross-validation, 10 and 80% of training and 20% of testing the dataset to detect the class of attacks. The performance metrics of the random forest classifier are shown in Table 7.

Table 7
Detailed detection accuracy of the class of attacks using random forest

| | Performance Metrics (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Class of attacks | TP Rate | FP Rate | Precision | Recall | F-Measure | MCC | ROC | PRC |
| Normal | 0.999 | 0.002 | 0.999 | 0.999 | 0.999 | 0.996 | 1.000 | 1.000 |
| Slowloris | 0.795 | 0.001 | 0.772 | 0.795 | 0.783 | 0.782 | 0.999 | 0.889 |
| Slowhttptest | 0.996 | 0.000 | 0.997 | 0.996 | 0.997 | 0.997 | 0.999 | 0.999 |
| Hulk | 1.000 | 0.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| GoldenEye | 0.998 | 0.000 | 0.999 | 0.998 | 0.999 | 0.999 | 1.000 | 1.000 |
| Heartbleed | 0.909 | 0.000 | 1.000 | 0.909 | 0.952 | 0.953 | 1.000 | 0.986 |
| Weighted Avg | 0.998 | 0.001 | 0.998 | 0.998 | 0.998 | 0.996 | 0.998 | 0.997 |

The effectiveness of the random forest technique is compared with other machine learning approaches for its validation, as shown below in Table 8. The actual positive metrics are used for measuring the detection accuracy of the different classes of attacks. The average detection accuracy of the three machine learning techniques is 99.83%, 98.20%, 99.20%, and 94.61% for the random forest, multilayer perception, bagging, and AdaBoost.M1, respectively. It is also compared with J48 and REPTree machine learning schemes for flooding attacks classification.

Table 8
Performance comparison of the true positive rate for the different classes of attacks using machine learning schemes.

| | Machine learning techniques | | | | | |
|---|---|---|---|---|---|---|
| classifiers | Random Forest | MLP | Bagging | AdaBoost.M1 | J48 | REPTree |
| Normal | 0.999 | 0.998 | 0.998 | 0.972 | 0.998 | 0.998 |
| Slowloris | 0.795 | 0.096 | 0.828 | 0.000 | 0.890 | 0.839 |
| Slowhttptest | 0.996 | 0.000 | 0.991 | 0.000 | 0.990 | 0.991 |
| Hulk | 1.000 | 0.972 | 1.000 | 0.963 | 1.000 | 1.000 |
| GoldenEye | 0.998 | 0.952 | 0.996 | 0.000 | 0.996 | 0.995 |
| Heartbleed | 0.909 | 0.000 | 0.909 | 0.000 | 0.909 | 0.909 |

The performance of the proposed system is compared with M. Agarwal et al.[15] using machine learning schemes visualized, as shown in Fig. 7 below. The detection rate and accuracy of the proposed scheme random forest technique is effective flooding attack detection.

The proposed technique for detection and analysis of flooding is further compared with previous related works for its confirmation. The comparison with C. R. Panigrahi [28] suggests that random forest techniques achieve better performance for heterogeneous data in wireless sensor networks for flooding attacks, as shown in Fig. 8 for each metrics of the performance measures.

L. Mohammadpour et al. [40] Examine Convolutional neural network means convolution layer (CNN-MCL) deep learning technique for detection and classification DoS attacks. Table 9 shows the comparative performance of the random forest technique with CNN-MCL using the three evaluation metrics.

Table 9
Detailed detection accuracy of the class of attacks using random forest

| | **Evaluation measures** | | | | | |
|---|---|---|---|---|---|---|
| Class of attacks | Recall | | Precision | | F-Measure | |
| | CCN-MCL | RF | CCN-MCL | RF | CCN-MCL | RF |
| Slowloris | 97.96 | 77.20 | 57.30 | 79.50 | 72.30 | 78.30 |
| Slowhttptest | 91.50 | 99.70 | 56.43 | 99.60 | 69.81 | 99.70 |
| Hulk | 99.10 | 100.00 | 58.64 | 100.00 | 73.68 | 100.00 |
| GoldenEye | 98.71 | 99.90 | 58.19 | 99.80 | 73.22 | 99.90 |
| Weighted Avg | 96.82 | 94.20 | 57.64 | 94.72 | 72.25 | 94.48 |

The performance of the proposed machine learning technique for identification and classification of flooding attacks is proved by comparing it with related DoS detection techniques. J. Ramprasath and V. Seethalakshmi [41] suggested DoS attack identification using rule-based dynamic access control lists (RDACL) with the simulated virtual network in software-defined networks. Figure 9 shows the comparative performance of flooding attack detection using different evaluation metrics in percentage. The effectiveness of the random forest is better in terms of precision, accuracy, and F-measure. F. Wu et al. [42] Analyzed the four classes of attack detection using the dataset KDDCUP99 by the application improved random forest-support vector machine learning model. The average accuracy of this model is 97.04%, which is less than compared to the new techniques.

S. E. Quincozes and J. F. Kazienko [14] Assessed the detection and classification accuracy of flooding, black hole, and gray hole DoS WSNs attacks using machine learning approaches. The detection of flooding attacks is 98.07% which is less than the new approach, 99.83%. This shows that the random forest approach is more accurate for heterogeneous classes of DoS attacks in the network. The accuracy of the scheme is more accurate than S. E. Quincozes and J. F. Kazienko [16] using the dataset as benchmark DARPA 98, which is 99.5%. N. Dharini et al. [43] Achieved 96% accuracy for energy prediction of flooding attack detection using the lightweight energy prediction technique. This further comparison approves the random forest techniques for the detection of flooding attacks using heterogeneous datasets. M. F. Suleiman and B. Issac [44]

researched attack detection using machine learning classifiers with the NSL-KDD dataset, which achieved an average detection accuracy of 99.76% less than the new approach.

The analysis of the effectiveness of the various machine learning classification techniques depends on the type of class of attacks. The random forest technique is the most accurate for detecting flooding attacks, with an average detection accuracy of 99.83%. M. M. Ahmed et al. [45] proposed WiNiP architecture for detecting internal and external jamming DoS attacks with an accuracy of 95.95% in the presence of an adversary, proving that random forest is more effective for detecting flooding attacks.

# 7 Conclusion And Future Works

Wireless sensor networks are vulnerable to various malicious attacks, such as flooding attacks on the network layer. The flooding attack makes the legitimate node busy by sending multiple routing requests and spreading false routing information to the destination node, acting itself as authenticate node. The flooding attack degrades the performance of the system by spreading fraudulent routing requests. The flooding attack disables the relay nodes between the source and destination nodes for each path of the simulation scenarios. These simulations are run by varying the number of sensors for the random generator. The statistical simulation results are the probability of message delivery, average hope count, and critical relays. The likelihood of message delivery having a route between sources and destinations is 88.64%. The simulation results show that flooding attacks increase packet dropping and affect the routing path between the source and destination nodes. Machine learning techniques are used to evaluate the proposed flooding attack detection and analysis using the cyber security CIC-IDSS2017 dataset as a benchmark. The random forest machine learning technique achieves 99.83% detection rate accuracy.

We will investigate the other network layer attacks in WSNs using different evaluation techniques in future works. We will examine other classes of attacks using other datasets in different WSN layers using machine learning techniques.

# Declarations

### Conflict of interest

The authors declare that there is no computing interest in this work.

### Data availability

The cyber security CIC-IDSS2017 dataset is used as a benchmark for evaluation available online https://www.unb.ca/cic/datasets/ids-2017.html.

# References

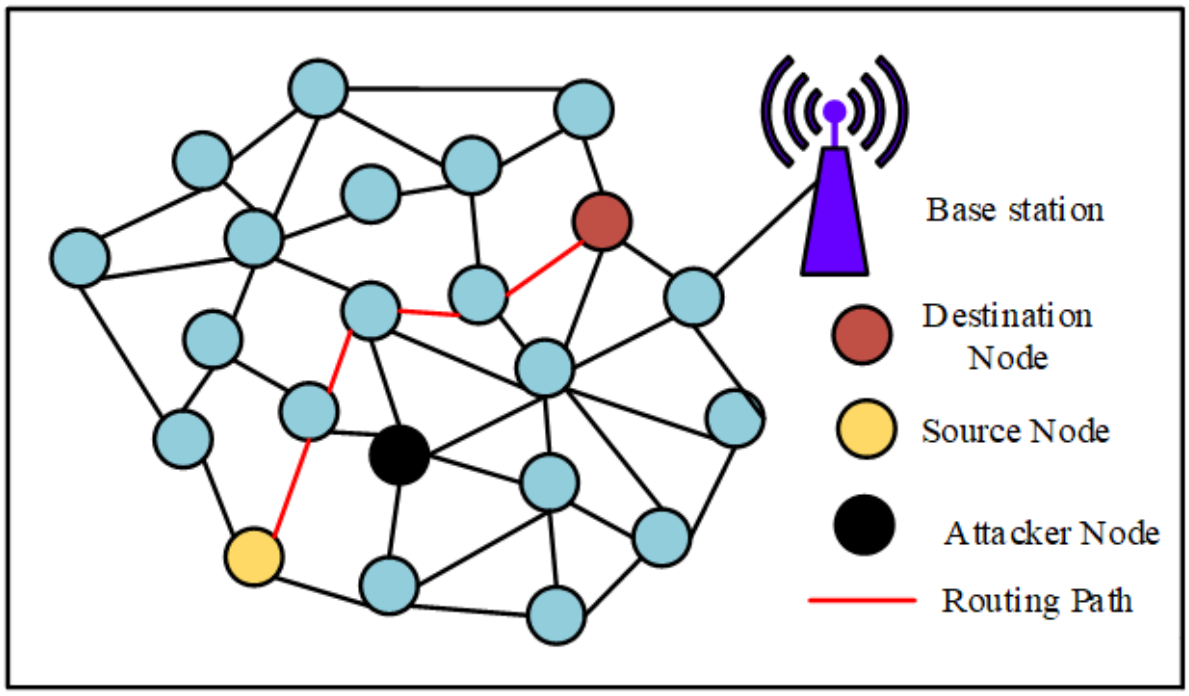
1. A. H. Moon, U. Iqbal, G. M. Bhat, and Z. Iqbal, “Simulating and analyzing RREQ flooding attack in Wireless Sensor Networks,” Int. Conf. Electr. Electron. Optim. Tech. ICEEOT 2016, pp. 3374–3377, 2016.

2. P. J. B. Pajila, E. G. Julie, and Y. H. Robinson, FBDR-Fuzzy Based DDoS Attack Detection and Recovery Mechanism for Wireless Sensor Networks, no. 0123456789. Springer US, 2021.
3. A. Sana, F. Youssef, and R. M. Driss, “Flooding attack on aodv in WSN,” 3rd Renew. Energies, Power Syst. Green Incl. Econ. REPS GIE 2018, pp. 1–5, 2018.
4. A. H. Farooqi, F. A. Khan, J. Wang, and S. Lee, “A novel intrusion detection framework for wireless sensor networks,” Pers. Ubiquitous Comput., vol. 17, no. 5, pp. 907–919, 2013.
5. P. Mohammadi and A. Ghaffari, “Defending Against Flooding Attacks in Mobile Ad-Hoc Networks Based on Statistical Analysis,” Wirel. Pers. Commun., vol. 106, no. 2, pp. 365–376, 2019.
6. M. N. U. Islam, A. Fahmin, M. S. Hossain, and M. Atiquzzaman, “Denial-of-Service Attacks on Wireless Sensor Network and Defense Techniques,” Wirel. Pers. Commun., vol. 116, no. 3, pp. 1993–2021, 2021.
7. K. N. Shivaramu, P. S. Prasobh, and N. Poti, “A Survey on Security Vulnerabilities in Wireless Ad Hoc High Performance Clusters,” Procedia Technol., vol. 25, no. Raerest, pp. 489–496, 2016.
8. F. Sadikin, T. van Deursen, and S. Kumar, “A ZigBee Intrusion Detection System for IoT using Secure and Efficient Data Collection,” Internet of Things, vol. 12, p. 100306, 2020.
9. O. A. Osanaiye, A. S. Alfa, and G. P. Hancke, “Denial of Service Defence for Resource Availability in Wireless Sensor Networks,” IEEE Access, vol. 6, pp. 6975–7004, 2018.
10. M. Premkumar and T. V. P. Sundararajan, “Defense Countermeasures for DoS Attacks in WSNs Using Deep Radial Basis Networks,” Wirel. Pers. Commun., no. 0123456789, 2021.
11. K. M and S. D, “Wireless Sensor Networks: Active and Passive attacks Vulnerabilities and Countermeasures,” Glob. Transitions Proc., 2021.
12. S. Gurung and S. Chauhan, “A novel approach for mitigating route request flooding attack in MANET,” Wirel. Networks, vol. 24, no. 8, pp. 2899–2914, 2018.
13. V. Moorthy, R. Venkataraman, and R. Gururajan, “Bayesian trust analysis of flooding attacks in distributed software defined networking nodes,” J. Ambient Intell. Humaniz. Comput., vol. 12, no. 7, pp. 7489–7498, 2021.
14. S. E. Quincozes and J. F. Kazienko, “Machine Learning Methods Assessment for Denial of Service Detection in Wireless Sensor Networks,” IEEE World Forum Internet Things, WF-IoT 2020 - Symp. Proc., pp. 1–6, 2020.
15. M. Agarwal, D. Pasumarthi, S. Biswas, and S. Nandi, “Machine learning approach for detection of flooding DoS attacks in 802.11 networks and attacker localization,” Int. J. Mach. Learn. Cybern., vol. 7, no. 6, pp. 1035–1051, 2016.
16. J. David and C. Thomas, “Efficient DDoS flood attack detection using dynamic thresholding on flow-based network traffic,” Comput. Secur., vol. 82, pp. 284–295, 2019.
17. B. D. Patel and A. D. Patel, “A trust based solution for detection of network layer attacks in sensor networks,” Proc. - 2016 Int. Conf. Micro-Electronics Telecommun. Eng. ICMETE 2016, pp. 121–126, 2016.
18. G. Kocher and G. Kumar, “Machine learning and deep learning methods for intrusion detection systems: recent developments and challenges,” Soft Comput., vol. 25, no. 15, pp. 9731–9763, 2021.
19. A. A. Jasim et al., “Secure and Energy-Efficient Data Aggregation Method Based on an Access Control Model,” IEEE Access, vol. 7, pp. 164327–164343, 2019.

20. “Bluetooth Mesh Flooding in Wireless Sensor Networks - MATLAB & Simulink - MathWorks Australia.” [Online]. Available: https://au.mathworks.com/help/comm/ug/bluetooth-mesh-flooding-in-wireless-sensor-networks.html. [Accessed: 26-Aug-2021].
21. “IDS 2017 | Datasets | Research | Canadian Institute for Cybersecurity | UNB.” [Online]. Available: https://www.unb.ca/cic/datasets/ids-2017.html. [Accessed: 30-Aug-2021].
22. A. Bansal and S. Kaur, Extreme gradient boosting based tuning for classification in intrusion detection systems, vol. 905. Springer Singapore, 2018.
23. M. Anbarasan et al., “Detection of flood disaster system based on IoT, big data and convolutional deep neural network,” Comput. Commun., vol. 150, no. November 2019, pp. 150–157, 2020.
24. D. Praveen Kumar, T. Amgoth, and C. S. R. Annavarapu, “Machine learning algorithms for wireless sensor networks: A survey,” Inf. Fusion, vol. 49, no. September 2018, pp. 1–25, 2019.
25. E. Hmouda and W. Li, “Validity Based Approach for Feature Selection in Intrusion Detection Systems,” Conf. Proc. - IEEE SOUTHEASTCON, vol. 2020-March, 2020.
26. S. Otoum, B. Kantarci, and H. T. Mouftah, “A Novel Ensemble Method for Advanced Intrusion Detection in Wireless Sensor Networks,” IEEE Int. Conf. Commun., vol. 2020-June, pp. 14–19, 2020.
27. N. A. Awad, “Enhancing network intrusion detection model using machine learning algorithms,” Comput. Mater. Contin., vol. 67, no. 1, pp. 979–990, 2021.
28. C. R. Panigrahi, Advanced Computing and Intelligent Engineering, vol. 1. 2018.
29. P. Paliwal and D. Kumar, ABC based neural network approach for churn prediction in telecommunication sector, vol. 84, no. Ictis. 2018.
30. P. Roy and C. Chowdhury, “A Survey of Machine Learning Techniques for Indoor Localization and Navigation Systems,” J. Intell. Robot. Syst. Theory Appl., vol. 101, no. 3, 2021.
31. P. Chandrasekar, K. Qian, H. Shahriar, and P. Bhattacharya, “Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing,” Proc. - Int. Comput. Softw. Appl. Conf., vol. 2, pp. 481–484, 2017.
32. I. A. Jimoh, I. Ismaila, and M. Olalere, “Enhanced Decision Tree-J48 with SMOTE Machine Learning Algorithm for Effective Botnet Detection in Imbalance Dataset,” 2019 15th Int. Conf. Electron. Comput. Comput. ICECCO 2019, Dec. 2019.
33. H. Chauhan, V. Kumar, S. Pundir, and E. S. Pilli, “A comparative study of classification techniques for intrusion detection,” Proc. - 2013 Int. Symp. Comput. Bus. Intell. ISCBI 2013, pp. 40–43, 2013.
34. N. Metropolis and S. Ulam, “The Monte Carlo Method,” J. Am. Stat. Assoc., vol. 44, no. 247, pp. 335–341, 1949.
35. H. Rahman et al., Advances in Intelligent Systems and Computing 1199 Progress in Advanced Computing and Intelligent Engineering, vol. 34, no. 5208. 2021.
36. M. Rabbani, Y. L. Wang, R. Khoshkangini, H. Jelodar, R. Zhao, and P. Hu, “A hybrid machine learning approach for malicious behaviour detection and recognition in cloud computing,” J. Netw. Comput. Appl., vol. 151, no. May 2019, p. 102507, 2020.
37. S. Razdan, H. Gupta, and A. Seth, “Performance Analysis of Network Intrusion Detection Systems using J48 and Naive Bayes Algorithms,” 2021 6th Int. Conf. Converg. Technol. I2CT 2021, pp. 1–7, 2021.

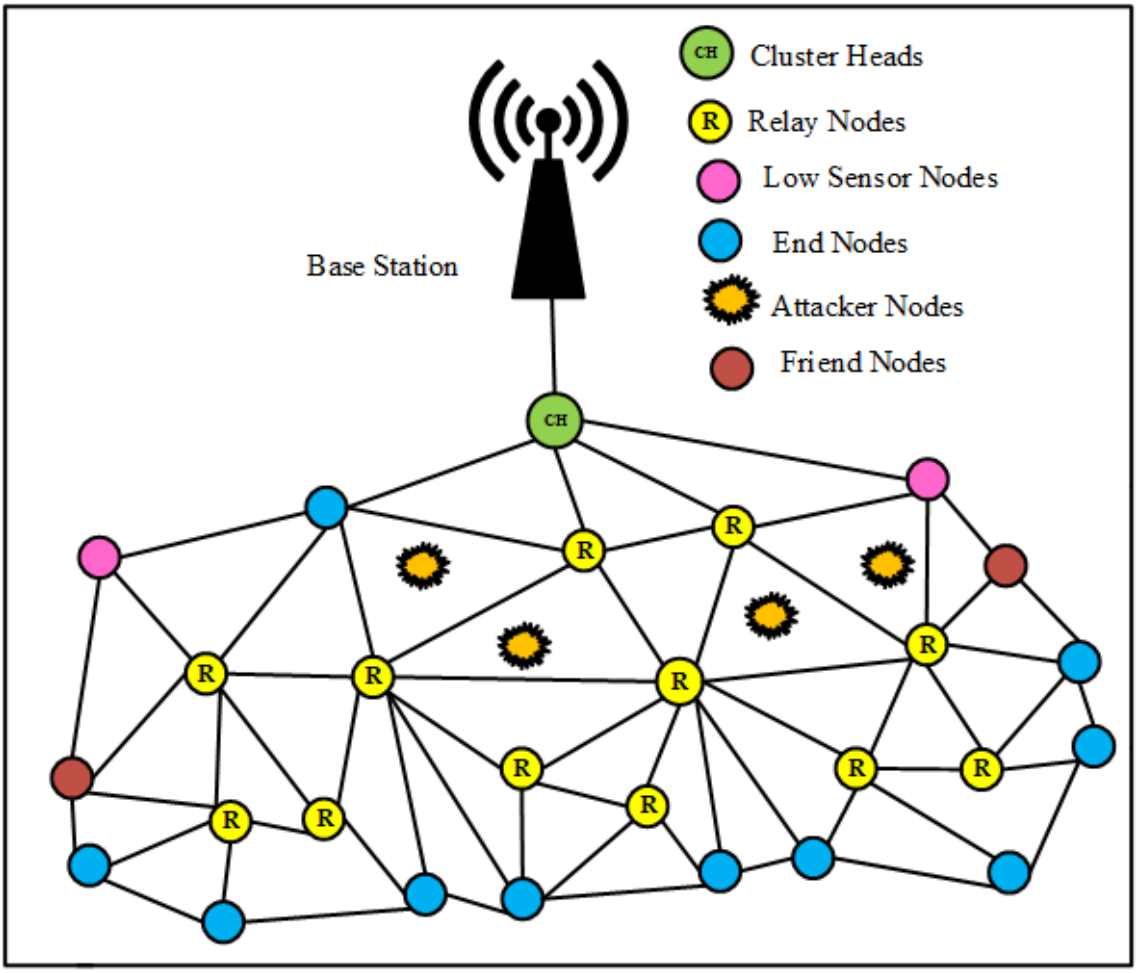
38. “What are TP rate, FP rate, precision, recall, F measure, MCC, ROC area and PRC areas in the Weka tool? - Quora.” [Online]. Available: https://www.quora.com/What-are-TP-rate-FP-rate-precision-recall-F-measure-MCC-ROC-area-and-PRC-areas-in-the-Weka-tool. [Accessed: 22-Oct-2020].
39. Y. Alsultanny, “Machine Learning by Data Mining REPTree and M5P for Predicating Novel Information for PM10,” Cloud Comput. Data Sci., pp. 40–48, 2020.
40. L. Mohammadpour, T. C. Ling, C. S. Liew, and A. Aryanfar, “A Mean Convolutional Layer for Intrusion Detection System,” Secur. Commun. Networks, vol. 2020, no. Ml, 2020.
41. J. Ramprasath and V. Seethalakshmi, “Mitigation of Malicious Flooding in Software Defined Networks Using Dynamic Access Control List,” Wirel. Pers. Commun., no. 0123456789, 2021.
42. F. Wu, T. Li, Z. Wu, S. L. Wu, and C. Q. Xiao, “Research on Network Intrusion Detection Technology Based on Machine Learning,” Int. J. Wirel. Inf. Networks, vol. 28, no. 3, pp. 262–275, 2021.
43. N. Dharini, R. Balakrishnan, and A. P. Renold, “Distributed detection of flooding and gray hole attacks in Wireless Sensor Network,” 2015 Int. Conf. Smart Technol. Manag. Comput. Commun. Control. Energy Mater. ICSTM 2015 - Proc., no. May, pp. 178–184, 2015.
44. M. F. Suleiman and B. Issac, “Performance Comparison of Intrusion Detection Machine Learning Classifiers on Benchmark and New Datasets,” pp. 19–23, 2021.
45. M. M. Ahmed, A. Ganguly, A. Vashist, and S. M. Pudukotai Dinakarrao, “AWARe-Wi: A jamming-aware reconfigurable wireless interconnection using adversarial learning for multichip systems,” Sustain. Comput. Informatics Syst., vol. 29, p. 100470, 2021.

## Figures



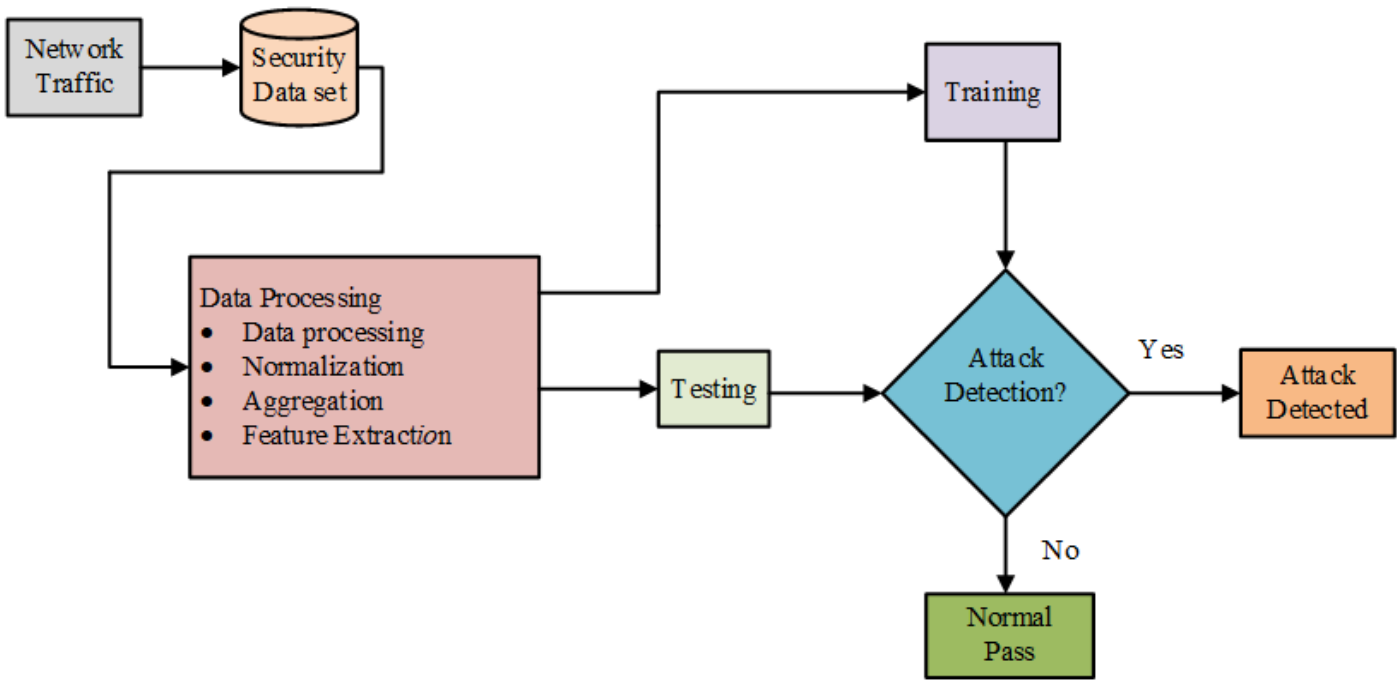
## Figure 1

Flooding attacks in wireless sensor networks



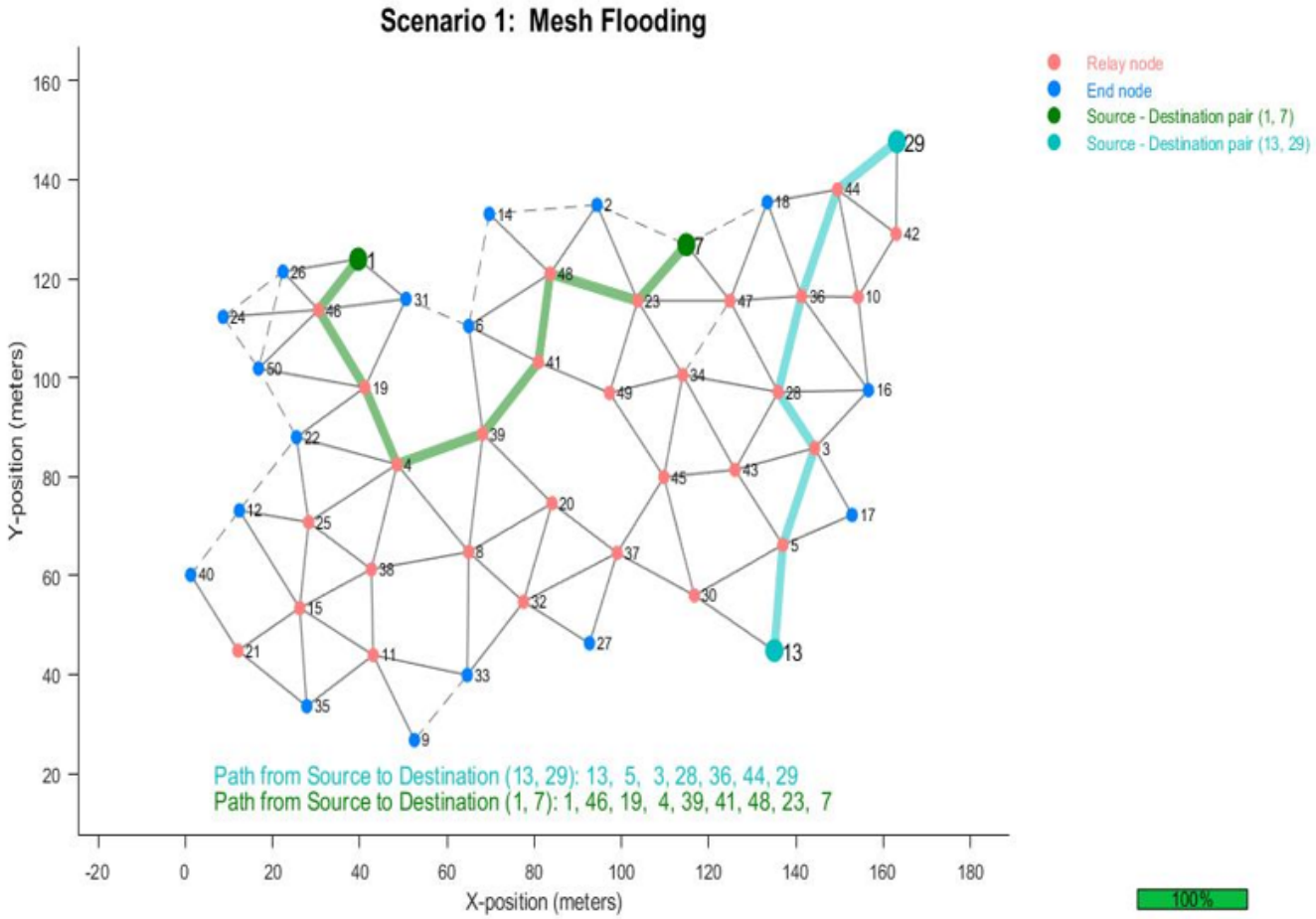

## Figure 2

Proposed meshed wireless sensor networks model



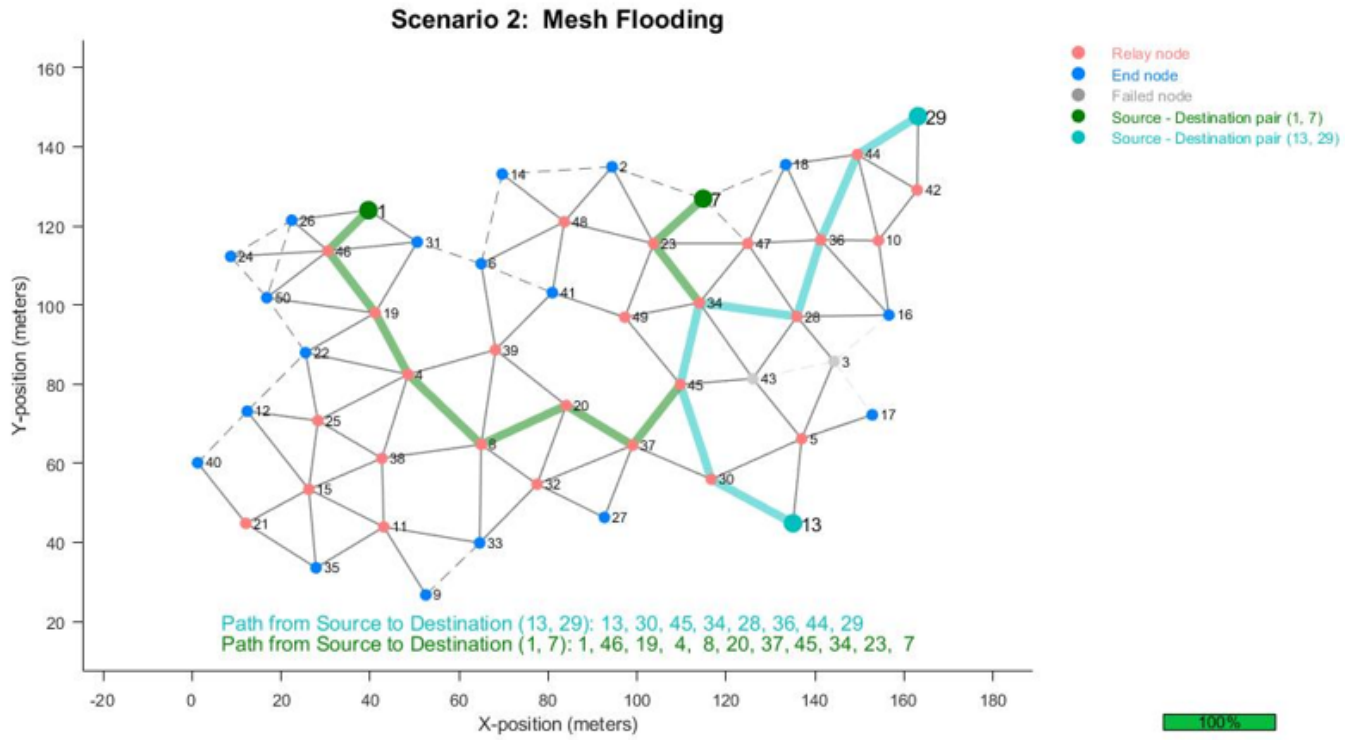

## Figure 3

Proposed block diagram for detection of flooding attacks in WSNs.



## Figure 4

Simulation scenario of wireless mesh sensor networks with all active nodes




## Figure 5

Simulation scenario of meshed wireless sensor nodes under failure nodes

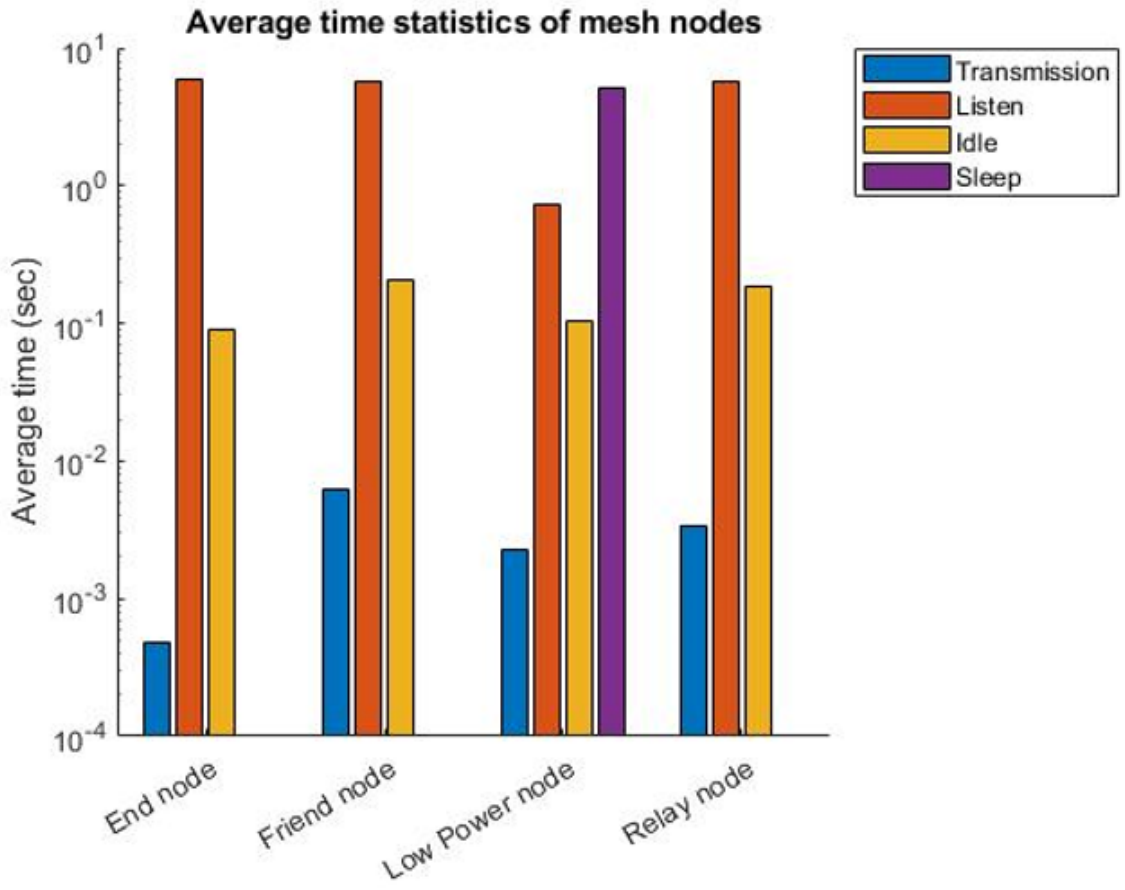
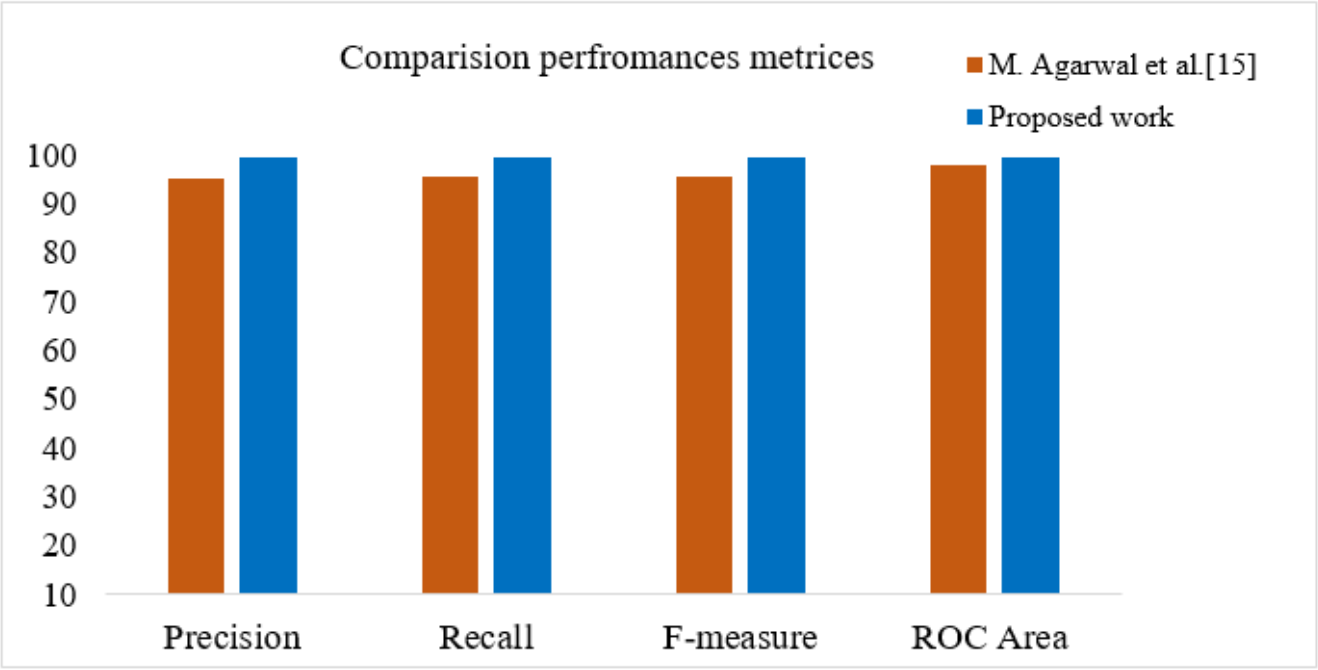


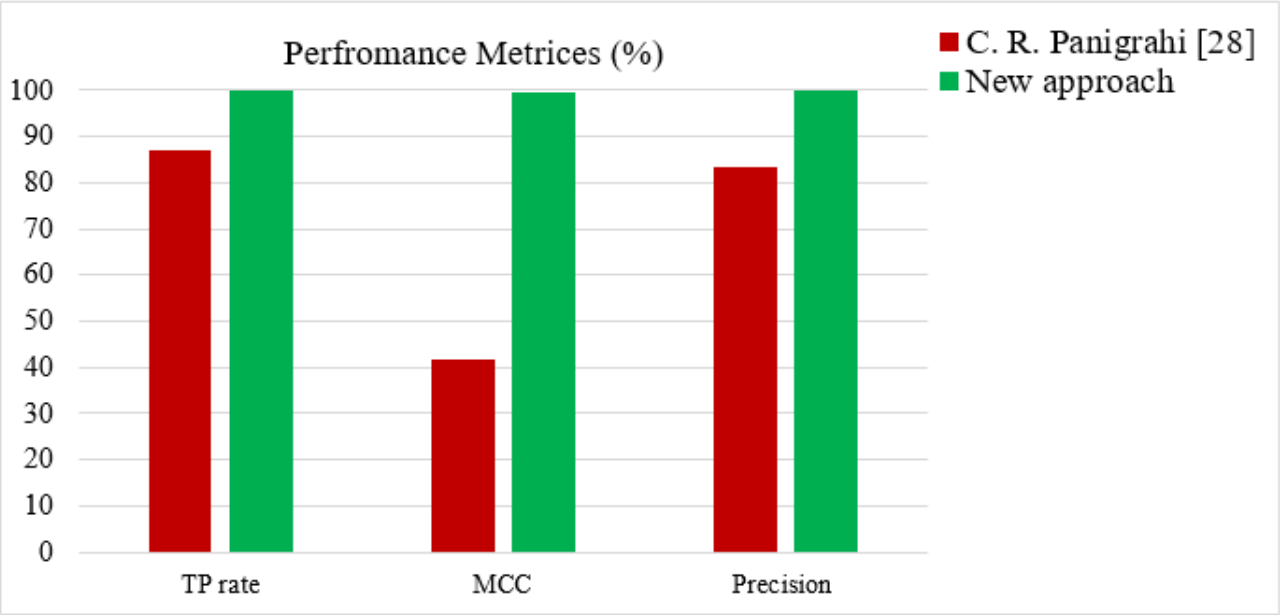

**Figure 6**

The timing characteristics of the various sensor nodes at different states



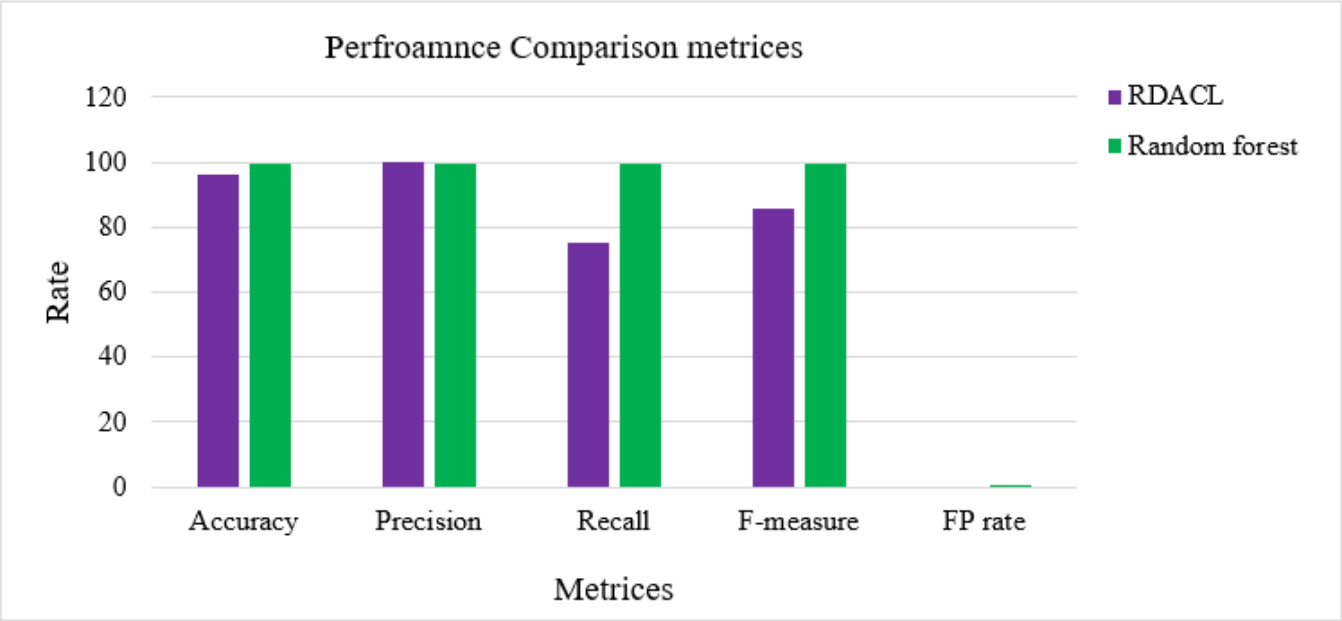
## Figure 7

Comparison of flooding attack classification techniques




## Figure 8

Comparative performance with related work C. R. Panigrahi




## Figure 9

comparison performance of different attack detection metrics

Proceedings of the Third International Conference on Electronics and Sustainable Communication Systems (ICESC 2022)
IEEE Xplore Part Number: CFP22V66-ART; ISBN: 978-1-6654-7971-4

2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC) | 978-1-6654-7971-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICESC54411.2022.9885448

# ECG Based Early Heart Attack Prediction Using Neural Networks

Ashish Kumar
*Department of Information Technology*
*Delhi Technological University*
Delhi, India
ksndelash@gmail.com

Ketan Rathor
*Solution Architect*
*Birla Carbon*
Mumbai, India
ketanrathor@gmail.com

Snehit Vaddi
*Dept of Computer Science and Engineering*
*Gitam University*
Vishakhapatnam , India
v.snehith999@gmail.com

Devanshi Patel
*Department of Information Technology*
*K. J. Somaiya College of Engineering*
Mumbai, India
devanshipatel885@gmail.com

Preethi Vanjarapu
*Department of Computer Science Engineering*
*CMR engineering college*
Telengana, India
preethivanjarapu94@gmail.com

Manichandra Maddi
*Department of Electronics and Communication engineering*
*Vaagdevi College of Engineering*
Telengana, India
Manichandramaddi000@gmail.com

***Abstract*—In today's world, a large amount of data is being collected every minute. This data came from different sources: health care, social media, business applications, manufacturing industries, and many more. The healthcare industry gives a large amount of data. Especially heart disease, which stops the flow of blood from vessels, resulting in chest pain and heart attack. Heart diagnosis at starting stage of the disease can lead to a successful cure of the disease. The classifications that are used to prevent the diagnosis are: - Naive Byes classification, the Support Vector Machine classification, and the k-NN classification. Hence, some deep learning techniques need to be tested on the dataset which has been taken and needs to be tested on it. The primary goal of this paper is to identify novel deep learning accuracy and a better neural network between convolutional neural networks and artificial neural networks..**

***Index Terms*—Heart Attack Detection, Deep Learning, Artficial Neural Network,Convolution Neural Network, Neural Networks**

## I. Introduction

The research about ECG gives results about declined heart function. An irregular ECG signal helps us recognise various circumstances as spotting changes in geometric variation. As per the statistics that have been shown, the heart rate generally ranges from around 60 beats per minute to about 100 beats per minute, which is in the case of a normal person. The main features which are used in the identification of the ECG are in very wide use these days, which can be referenced from [2].But the problems stay in their place where they are generally not resolved due to many factors which need to be taken into consideration. We use CNN to better classify the detection with better accuracy and in less time with better precision. The major variables in recognising any abnormality in an ECG are the main aspects which need to be catered to or taken in mind during the process of detection. There are still a lot of difficulties to be resolved before useful data can be obtained. From patient to patient, the accuracy with which diseases are recognised varies. We use CNN to better categorise the detection with more precision and accuracy in less time [3].

## II. Related Work

The general term "heart" is a main aspect of the research as the bodies are operated through it and there are many other diseases which can be caused due to the following heart problems. The different problems have been classified as a kind of the different names, but the thing that is related to all of them, despite the name, is the part or the organ heart. The heart rate is found with the help of many devices that are available in the market and the ECG, which is used to see if a patient is having a cardiac arrest or not. While some of the diseases are not at all serious, whereas at the same time other diseases may prove fatal [4]. Hence, we need to classify different types of heart diseases. To classify different heart disease, we need to use and take advantage of CNN algorithm. According to a study on ECG [7], an irregular ECG signal aids us in recognising various scenarios, such as noticing changes in geometric variation. As per the statistics that has been shown, the heart rate generally ranges from around 60 beats to about 100 beats per minute, which is in



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2022 at 04:27:44 UTC from IEEE Xplore. Restrictions apply.

the case of a normal person. The main features which are used in the identification of the ECG are in very use these days, which can be referenced from [2]. Therefore, having 1 minute having larger sample data means training of CNN would yield better results and the accuracy of classification would also be higher. Using this system not only reduces the need of using manual measures but also can check the status of heart with less cost. The accuracy is 99.46 percent, and distinct patients can be treated, all thanks to CNN [8]. The factors which are the main reason for the electrical impulse are also the main factors in recognising any issue in ECG. There are still a number of issues that need to be addressed in order to obtain meaningful data[6]. The precision with which illnesses are detected varies from patient to patient. We employ CNN to better categorize the detection with more accuracy and precision in less time.Health care industry plays a major role in database. But the main thing is to improve health care services at an affordable value. In heart disease, deformity of heart takes place, which causes chest pain. There is lots of data which health care industry has to manage. That's why it's the most difficult task to do. So, classifiers came into play which are used to detect and avert heart disease. These classifiers are different kind of machine learning classification algorithms which are used for the case [10]. This paper focuses on different type of classifiers used for storing data as there is a large amount of data that has to be stored. These classifiers work really well and are able to detect disease[11]. These classifiers use High tech technologies, which also increase the accuracy and increase the growth of the healthcare industry [5].

## III. Methodology

### A. Data preparation

The dataset was collected from the UCI ML Dataset. The dataset has 383 data points (or rows) and 14 features (or columns). Features present in the dataset are:Age, Sex, excang, caa, cp, trtbps, oldpeak, chol, fbs, restecg, thalachh, target, thall, slp. The information can be seen in figure 1.

| | age | sex | cp | trtbps | chol | fbs | restecg | thalachh | exng | oldpeak | slp | caa | thall | output |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 63 | 1 | 3 | 145 | 233 | 1 | 0 | 150 | 0 | 2.3 | 0 | 0 | 1 | 1 |
| 1 | 37 | 1 | 2 | 130 | 250 | 0 | 1 | 187 | 0 | 3.5 | 0 | 0 | 2 | 1 |
| 2 | 41 | 0 | 1 | 130 | 204 | 0 | 0 | 172 | 0 | 1.4 | 2 | 0 | 2 | 1 |
| 3 | 56 | 1 | 1 | 120 | 236 | 0 | 1 | 178 | 0 | 0.8 | 2 | 0 | 2 | 1 |
| 4 | 57 | 0 | 0 | 120 | 354 | 0 | 1 | 163 | 1 | 0.6 | 2 | 0 | 2 | 1 |

Fig. 1. Dataset overview

### B. Data Preprocessing

The dataset is analysed very carefully before submitting the data to the model to train from and give predictions. The dataset has 8 categorical features and 6 are numerical features. The dataset has no null values

### C. Feature Visualisation

Figure 2 represents the number of people having heart attacks in the given dataset. Figure 3 depicts effects of different features present in the dataset on the heart attack.Figure 4 shows the correlation matrix between the features of the dataset which is a very important part to know which feature correlates the most with which other feature and vice versa.Figure 5 shows the distribution of the features (age, trtbps, chol, thalachh, oldpeak) in the dataset according to the target vaiable. Figure 6 shows the distribution of the features (chest pain, number of major vessels, cholestrol, sex, thall, excng and age) in the dataset according to the target vaiable.

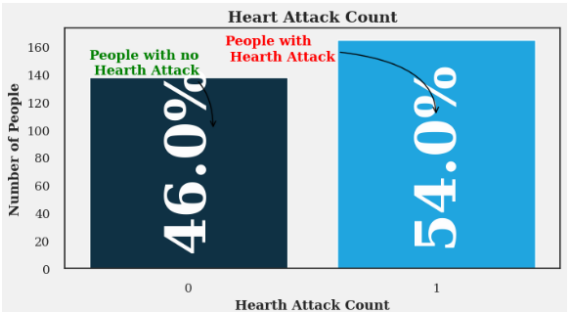



Fig. 2. Number Of People Having Attacks

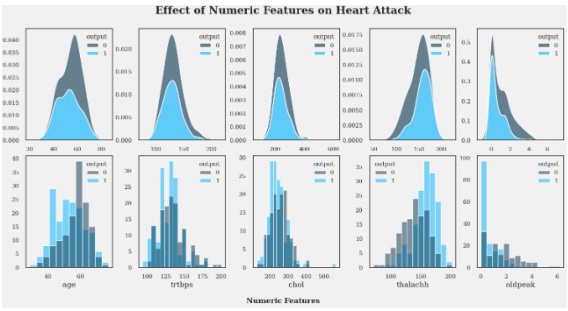



Fig. 3. Features Present In The Dataset

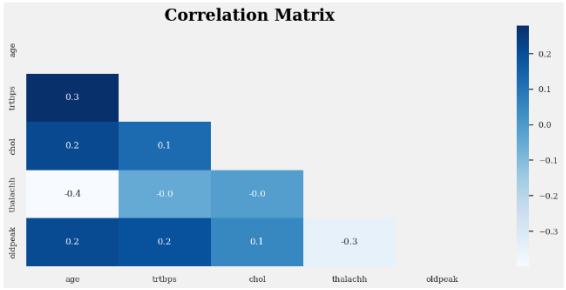



Fig. 4. Correlation Matrix

### D. Model Architecture

Convolutional neural networks are known for their high performance, their ability to identify and distinguish objects and images of similar or dissimilar types. A Convolutional Neural Network consists of 3 main layers: Convolutional layers, pooling layers, fully connected layers.It is the main



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2022 at 04:27:44 UTC from IEEE Xplore. Restrictions apply.

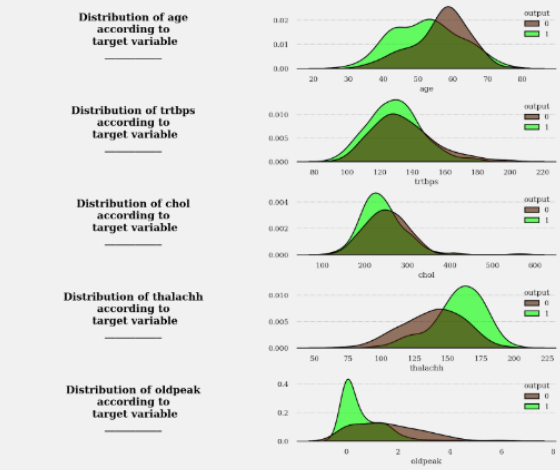



Fig. 5. Distribution of features-I

Fig. 6. Distribution of features-II

convolutional and computational layer; in this layer, the operation includes multiplication of set of weights with input images represented in matrices. Here, the array of weights is known as kernels. The input images become smaller of size to perform better and more performance centric operations. The image becomes gray-scale and is downgraded or down sampled. It is done to reduce the number of parameters in the input. We know the values denoted by each pixel is not connected directly to the output of the image. Therefore, in fully connected layer, each node of an input pixel is directly connected to the output pixel value. Similarly, a process needs to be followed in the case of the artificial neural network .

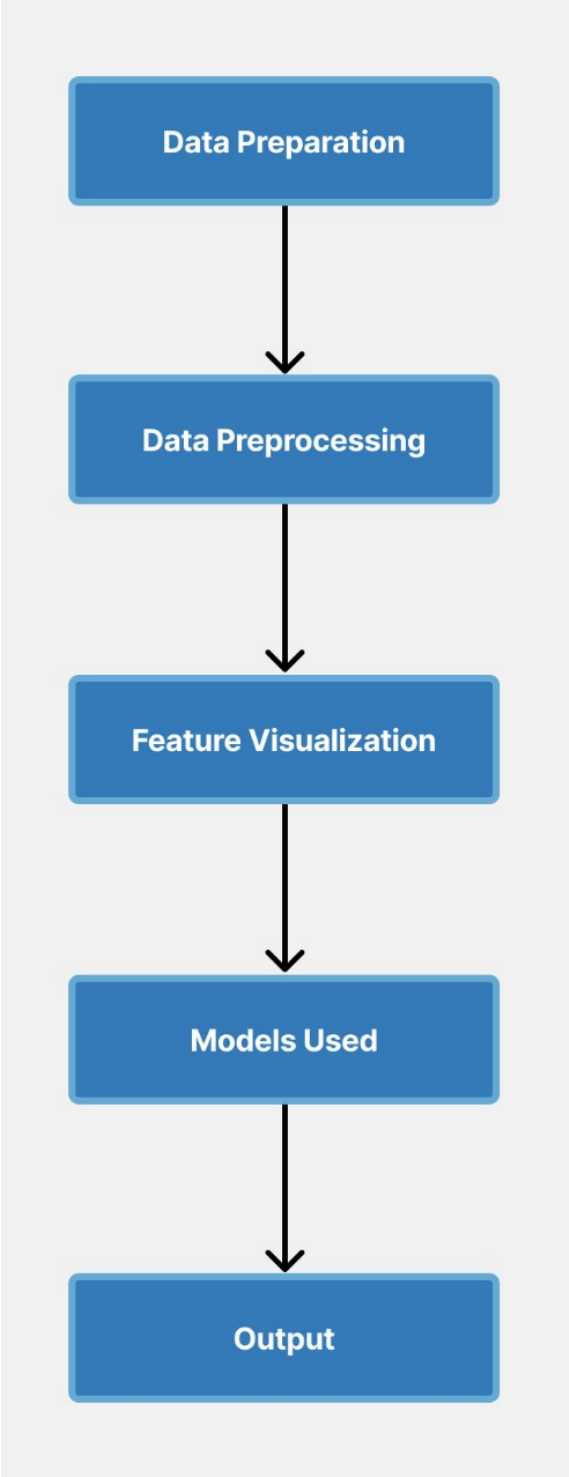



Fig. 7. Flowchart

## IV. Experimental Results

The feature visualisation that has been performed by the authors really helped in getting to know the features and how the parameters can be kept for the models which have been



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2022 at 04:27:44 UTC from IEEE Xplore. Restrictions apply.

TABLE I
ACCURACY TABLE

| Model Name | Author's Accuracy | Precision | Average F-score |
|---|---|---|---|
| Convolution Neural Network | 98% | 97% | 98% |
| Artificial Neural Network | 94% | 95% | 94% |

there.The artificial neural network and the convolution neural network were parameterised according to the observations which had been taken.The accuracies can be seen in table 1

## V. CONCLUSION

Due to the increasing factors which can cause a problem to the heart are increasing day by day hence an approach to solve it needs to be incorporated at a very early stage so that the attack can be stopped. Hence the authors devised an approach using the deep learning in which the convolution neural network outperformed the artificial neural network by 4 percent.

## FUTURE WORK

More and more data can be collected in the following problem statement and as this is a rising problem hence more people might face it up which can cause to give more and more data of them and a more precise one. The following models compared can also be done using comparison between the machine learning models.

## REFERENCES

[1] A. M. Zair, A. Bouzouad Cherfa, Y. Cherfa and N. Belkhamsa, "Heart Outer Surface Segmentation from Computed Tomography Images," 2022 7th International Conference on Image and Signal Processing and their Applications (ISPA), 2022, pp. 1-7, doi: 10.1109/ISPA54004.2022.9786309.

[2] P. Sasidharan, T. Rajalakshmi and U. Snekhalatha, "Wearable Cardiorespiratory Monitoring Device for Heart Attack Prediction," 2019 International Conference on Communication and Signal Processing (ICCSP), 2019, pp. 0054-0057, doi: 10.1109/ICCSP.2019.8698059.

[3] M. Makram, N. Ali and A. Mohammed, "Machine Learning Approach for Diagnosis of Heart Diseases," 2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), 2022, pp. 69-74, doi: 10.1109/MIUCC55081.2022.9781735.

[4] P. Yuan, "Research on predicting heart attack through active learning," 2021 2nd International Conference on Big Data Artificial Intelligence Software Engineering (ICBASE), 2021, pp. 12-18, doi: 10.1109/ICBASE53849.2021.00010.

[5] S. Manikandan, "Heart attack prediction system," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 817-820, doi: 10.1109/ICECDS.2017.8389552.

[6] N. Ehtiati, G. Attarodi, N. J. Dabanloo, J. F. Sedehi and A. M. Nasrabadi, "Prediction of ventricular tachycardia using nonlinear features of heart rate variability signal such as poincare plot, approximate and sample entropy, recurrence plot," 2017 Computing in Cardiology (CinC), 2017, pp. 1-4, doi: 10.22489/CinC.2017.099-274.

[7] R. Deepika, P. B. Srikaanth and R. Pitchai, "Early Detection of Heart Disease Using Deep Learning Model," 2022 8th International Conference on Smart Structures and Systems (ICSSS), 2022, pp. 1-4, doi: 10.1109/ICSSS54381.2022.9782295.

[8] R. Ramadoss and C. Vimala, "Classification of Pulmonary Emphysema using Deep Learning," 2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC), 2022, pp. 63-68, doi: 10.1109/ICESIC53714.2022.9783583.

[9] M. S. Raja, M. Anurag, C. P. Reddy and N. R. Sirisala, "Machine Learning Based Heart Disease Prediction System," 2021 International Conference on Computer Communication and Informatics (ICCCI), 2021, pp. 1-5, doi: 10.1109/ICCCI50826.2021.9402653.
B. Balakrishnan, K. Kumaran, A. Sankar and K. Rao, "Coronary Artery Disease Prediction using Machine Learning Algorithms," 2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC), 2022, pp. 95-100, doi: 10.1109/ICESIC53714.2022.9783592. [10]

[11] K. Mutijarsa, M. Ichwan and D. B. Utami, "Heart rate prediction based on cycling cadence using feedforward neural network," 2016 International Conference on Computer, Control, Informatics and its Applications (IC3INA), 2016, pp. 72-76, doi: 10.1109/IC3INA.2016.7863026.

[12] R. Kahankova, M. Mikolasova, R. Jaros, K. Barnova, M. Ladrova and R. Martinek, "A Review of Recent Advances and Future Developments in Fetal Phonocardiography," in IEEE Reviews in Biomedical Engineering, doi: 10.1109/RBME.2022.3179633.

[13] A. Pandiaraj, S. L. Prakash and P. R. Kanna, "Effective Heart Disease Prediction Using Hybridmachine Learning," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 731-738, doi: 10.1109/ICICV50876.2021.9388635.

[14] D. Hawashin, R. Jayaraman, K. Salah, I. Yaqoob, M. C. E. Simsekler and S. Ellahham, "Blockchain-based Management for Organ Donation and Transplantation," in IEEE Access, doi: 10.1109/ACCESS.2022.3180008.

[15] P. Moschovitis and A. Denisova, "Keep Calm and Aim for the Head: Biofeedback-Controlled Dynamic Difficulty Adjustment in a Horror Game," in IEEE Transactions on Games, doi: 10.1109/TG.2022.3179842.

[16] N. M. Lutimath, N. Sharma and B. K. Byregowda, "Prediction of Heart Disease using Random Forest," 2021 Emerging Trends in Industry 4.0 (ETI 4.0), 2021, pp. 1-4, doi: 10.1109/ETI4.051663.2021.9619208.

[17] P. Anuradha and V. K. David, "Feature Selection and Prediction of Heart diseases using Gradient Boosting Algorithms," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 711-717, doi: 10.1109/ICAIS50930.2021.9395819.

[18] Sathesh, A. "Performance analysis of granular computing model in soft computing paradigm for monitoring of fetal echocardiography." Journal of Soft Computing Paradigm (JSCP) 1, no. 01 (2019): 14-23

[19] Shakya, Subarna, and P. P. Joby. "Heart Disease Prediction using Fog Computing based Wireless Body Sensor Networks (WSNs)." IRO Journal on Sustainable Wireless Systems 3, no. 1 (2021): 49-58



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2022 at 04:27:44 UTC from IEEE Xplore. Restrictions apply.

# ECG Heartbeat Classification of Myocardial Infarction and Arrhythmia using CNN

Shrey Agarwal
Electronics and Communication Engineering
Delhi Technological University
New Delhi, India
shreyagarwal_2k18ec161@dtu.ac.in

Sarthak Khattar
Electronics and Communication Engineering
Delhi Technological University
New Delhi, India
sarthakkhattar_2k18ec155@dtu.ac.in

Saiyam Soni
Electronics and Communication Engineering
Delhi Technological University
New Delhi, India
saiyamsoni_2k18ec147@dtu.ac.in

Dr. Anukul Pandey
Electronics and Communication Engineering
Delhi Technological University
New Delhi, India
anukul.pandey@dtu.ac.in, anukul66@gmail.com

2022 IEEE India Council International Subsections Conference (INDISCON) | 978-1-6654-6601-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/INDISCON54605.2022.9862885

***Abstract*— The most common cause of death all over the world is cardiovascular disease (CVD), with Coronary Artery Disease (CAD) being the most common. To lower the risk of CAD consequences such as myocardial infarction, it is critical to diagnose and treat CAD early. The most popular diagnostic screening tool for detecting CAD through its symptoms like arrhythmia, is the electrocardiogram (ECG) readings. Current procedures for diagnosing CVDs, such as heart catheterization or blood tests, are invasive. Other non-invasive cardiac methods of testing have drawbacks, such as uncertainty about the best order, type, and frequency of cardiac imaging tests for various medical settings performance. In this paper we propose an approach to this problem using ECG values to train our respective models and focus on vertical improvement i.e. higher accuracy and better results for existing techniques rather than horizontal improvement i.e. building new techniques. We tested our techniques on PhysioNet's MIT-BIH and PTB Diagnostics Datasets, resampling certain data wherever necessary. As per our results, our models are able to predict the categories with average accuracies of 99% and 98% on different sub-categories of Myocardial Infarction and Arrhythmia ECG signals, respectively.**

***Keywords—ECG, Myocardial Infarction, Arrhythmia, CNN, Deep Learning***

## I. INTRODUCTION

Heart arrhythmias arise when the electronic signals that synchronize the beats of the heart don't work in the general manner. This faulty signaling causes the heart to beat either much faster causing tachycardia or much slower causing bradycardia or imperfectly. Atherosclerotic lesions with calcification can cause ischemia over time as the plaque volume expands and encroaches on the diameter of the coronary lumen, which can interfere and disrupt the overall myocardial blood flow [1]. This can lead to acute myocardial infarction, where the muscles of the heart die owing to an extended scarcity of oxygen [19, 12].

Electrocardiograms represent the electrical functioning of the heart is in the form of mapped signals. Irregular and disease-prone heart rhythms are hence, shown as low amplitude alterations in the signal. Even though these can be manually observed by a medical professional in a non-invasive way, confirmation requires invasive procedures like heart catheterization or blood tests. Doctors commonly use ECGs to monitor cardiac health. However, the primary issue with this practice involves the detecting and categorizing of distinct waveforms in the ECG signal. This is due to the fact that humans are error prone and take a lot of time in performing this task. Millions of people experience CVDs involving abnormal heartbeats which are even fatal in many cases. Therefore, a fast, accurate and low-cost diagnosis of such abnormalities is extremely important [2].

While there has been an increase in research and attention towards using computerized techniques involving ECG signals, many studies use ECG images as data for models, applying Computer Vision techniques to derive results [3]. Other studies focus on using ECG values themselves in CSV format and building a general transferable approach to the problem [4]. In such type of disease classification problems, Deep Learning models usually have many advantages over the conventional machine learning techniques. Feature selection is fully automated in deep learning models with a higher affinity to be trained with large datasets, unlike Machine Learning models that prefer smaller ones [20].

In this paper, we put forward an automated non-invasive solution to classify various Myocardial Infarction and Arrhythmia ECG waveforms using Deep Learning techniques. We resampled the unevenly large labelled data uniformly amongst most categories of MI and Arrhythmia signals, which allowed us to significantly improve the accuracy of the model across the board. We also considered conventional Machine Learning techniques and compared the results with our Deep Learning model to showcase the improvements.

The rest of this paper is organized as follows. Section II explains the related work done in the past. Section III presents the datasets used in this study. Section IV describes the details of the methods used. Section V presents the results achieved and Section VI concludes this study.

## II. RELATED WORK

Since this paper focuses on vertical improvement – pushing existing techniques to get the best results possible, it's important to compare our results with the ones achieved previously by others. Our model and approach is inspired by the work done in Kachuee *et al*.[4] with some key modifications that lead us to achieve the higher accuracy. [4] uses a deep residual CNN model to classify the heartbeat signals and detect MI and Arrhythmia. However, this model

978 -1-6654 -6601 -1/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 12,2022 at 06:44:54 UTC from IEEE Xplore. Restrictions apply.

requires high computational power. In this paper, we propose a solution which requires lower computational power and gives a higher average accuracy for prediction of labels. Most of the previous results have also been concentrated on MI classification [8, 9, 10] whereas, we showcase our results for both MI and Arrhythmia.

## III. Datasets

The dataset chosen in this paper is ECG heartbeat categorization dataset from Kaggle [14]. The dataset contains labelled ECG signal records which have been derived from PhysioNet MIT-BIH Arrhythmia [15, 16] and PTB Diagnostic ECG databases [15, 17]. These ECG signals correspond to heartbeats of normal cases and cases affected by Arrhythmia and Myocardial Infarction and have been sampled to 125Hz [14, 4].

The MIT-BIH database contains a total of 109446 samples which have been derived from ECG recordings of 47 distinct subjects. The samples are classified into 5 distinct categories:

- 0.0 – Normal beat
- 1.0 – Supraventricular premature beat
- 2.0 – Premature Ventricular Contraction
- 3.0 – Fusion of Ventricular and Normal beat
- 4.0 – Unclassifiable beat

The PTB database contains a total of 14552 samples which have been derived from ECG recordings of a total 290 distinct subjects: 52 healthy subjects, 148 subjects which have been diagnosed with MI and the remaining 7 subjects have been diagnosed with different diseases. The samples are classified into 2 different categories:

- 0.0 – Normal beat
- 2.0 – Abnormal beat (indicating MI)

## IV. Methodology

### A. *Preprocessing data*

The dataset for Arrhythmia and MI, which we have selected from Kaggle contains 188 columns in which the columns 0 to 187 describe the ECG signal while the 188$^{th}$ contains the category of signal. For both the datasets we normalized the values in order to change the values to a common scale and ease computation. Further we split the data into train, test and validation data.

It is noteworthy to mention that in MIT-BIH dataset the number of samples for the five categories differed significantly. While the number of samples for *Normal beat* were 72471, the number of samples for other categories were less than 7000. Thus, we split the data into 5 different data frames based on their category and resampled the data in each category up to 10000 samples.

We did not resample the data frame for *Normal beat* since there were considerable number of samples available. This resampling was done only on the training dataset after splitting it into train and validation sets. The test set was not resampled so that the testing could be done on original unaltered dataset in order to get true results.

### B. *Machine learning models*

Before defining a CNN model for classification, we first used conventional machine learning models for classification in order to assure that sufficient data is available for training a neural network. We trained SVM (Support Vector Machine),

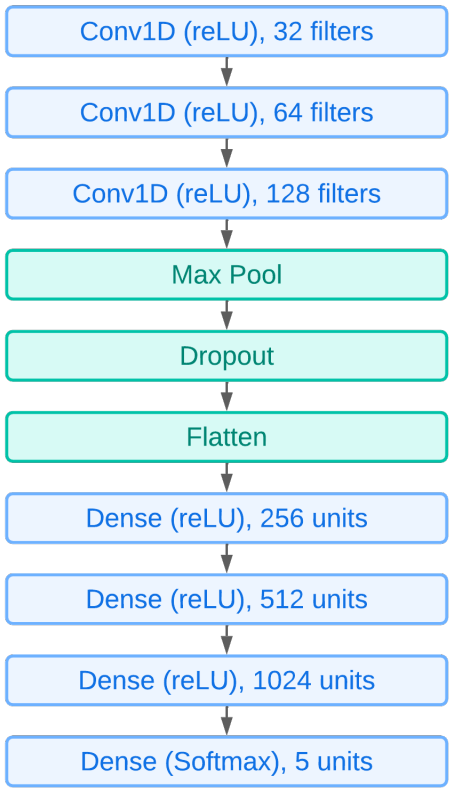



Fig. 1. Proposed CNN model architecture with 10 layers for classification of Myorcardial Infarction and Arrhythmia

ExtraTreesClassifier and KNN (K Nearest Neighbor) machine learning models for both MIT-BIH and PTB datasets using scikit-learn along with GridSearchCV. GridSearchCV is a utility to loop through predefined parameters in order to arrive at the best parameters for a model. The models were fitted using these best parameters and the results were compared with CNN model as described in section V.

### C. *Training CNN model for Arrhythmia*

After preprocessing the data and establishing that sufficient data is available, we trained a Convolutional Neural Network for detection of Arrhythmia. The proposed model not only detects but also classifies the ECG signal into distinct categories of Arrhythmia.

Our proposed architecture for the model of classification of ECG signals has been illustrated in Fig. 1. The model consists of 3 1-D convolution layers with 32, 64 and 128 filters and activation layer as 'ReLU' (Rectified Linear Unit). Multiple convolutional layers allow extraction of high level features but also increase the computational power, thus we use 3 convolutional layers. These three layers allow a hierarchical decomposition of the input signals and leads to better detection of details in the signal. This is followed by max pooling of kernel size 3. This max pooling layer selects the maximum elements from feature map and outputs the most eminent features of the previous layer. We then add

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 12,2022 at 06:44:54 UTC from IEEE Xplore. Restrictions apply.

dropout of 0.5 in order to reduce the chances of overfitting. The output of these is then flattened. Further, we add 3 dense layers with 256, 512 and 1024 neurons followed by an output layer of 5 nodes and SoftMax as the activation in order to predict the probabilities of output class.

TABLE I. COMPARING HEARTBEAT CLASSIFICATION RESULTS

| Work | Model | Accuracy (%) |
|---|---|---|
| **This Paper** | **Conv1D CNN** | **MI - 99.38<br>Arr – 98.57** |
| [4] Kachuee *et al.* | Deep Residual CNN | 95.9 |
| [5] Martis *et al.* | DWT + SVM | 93.8 |
| [6] Acharya *et al.* | Augmentation + CNN | 93.5 |
| [7] Li *et al.* | DWT + Random Forest | 94.6 |
| [8] Acharya *et al* | CNN | 93.5 |
| [9] Safdarian *et al.* | Naïve Bias Classification | 94.7 |
| [10] Kojuri *et al* | MLP ANN | 95.6 |
| [11] Sun *et al.* | Latent Topic MIL | 82.4 |
| [12] Liu *et al.* | PolyECG-S Features | 94.4 |
| [13] Sharma *et al* | SVM + (RBP/KNN) | 96 |

*MI – Myocardial Infarction*
*Arr – Arrhythmia*

TABLE II. CLASSIFICATION RESULTS OF DIFFERENT MODELS

| Model | Arrhythmia Accuracy (%) | MI Accuracy (%) |
|---|---|---|
| SVM | 94.72 | 94.50 |
| ExtraTreesClassifier | 97.49 | 96.56 |
| KNN | 97.40 | 93.57 |
| **Conv1D CNN** | **98.57** | **99.38** |

### *D. Training CNN model for Myocardial Infarction*

The proposed model as illustrated in Fig. 1 and described in IV-C was then trained for PTB dataset in order to predict Myocardial Infarction. The only difference in the model for MI and Arrhythmia is that the number of nodes for the very last SoftMax output layer is 2 in case of MI and 5 for Arrhythmia. The rest of the layers for the two models are indistinguishable.

## V. RESULTS & DISCUSSION

After training the models, we tested all the models on test dataset and discovered some interesting results. Table I mentions the average accuracies that were achieved by relevant related works and compares their final results with our proposed model. The performance of our proposed CNN model comes across better than the other works for both the diseases. Table II represents and compares the accuracy achieved by different proposed machine learning models we developed for Arrhythmia and Myocardial infarction classification.

### *A. Machine Learning*

All the machine learning models: SVM, KNN and ExtraTreesClassifier had an appreciable accuracy as mentioned in Table I. The best regularization parameters for SVM proved to be 10. KNN gave its best results when the nearest neighbours were 3. The best results for ExtraTreesClassifier were achieved when maximum depth was 20 and the number of trees were 30.

### *B. Conv1D CNN*

The CNN model performed very well for classification of signal for both the diseases. The model achieved an accuracy of 99.38% for Myocardial Infarction after running for 20 epochs and 98.57% for Arrhythmia after running for 18 epochs.

Training model history was saved to illustrate the accuracy and loss curves over the epochs for validation and training datasets. The training accuracy increased over the epochs and the loss for model decreased. Thus, the model has a good learning rate and high accuracy for both the diseases.

The accuracy and loss plot of training and validation data for myocardial infarction can be seen in Fig. 2 and Fig. 3, respectively. The accuracy and loss plot for training and validation data of Arrhythmia can be seen in Fig. 4 and Fig. 5, respectively. We used *scikitplot* for visualizing confusion matrices in order to summarise the performance of classification by comparing the target values with predicted values. Fig. 6 displays the confusion matrix for Myocardial Infarction. Fig. 7 displays the confusion matrix for Arrhythmia.

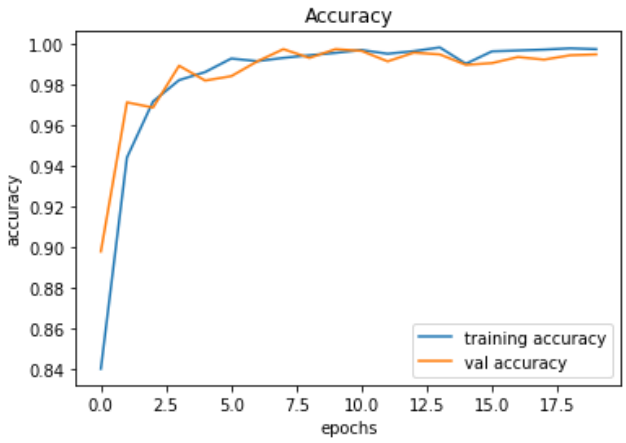



Fig. 2. Myocardial Infarction detection accuracy curve

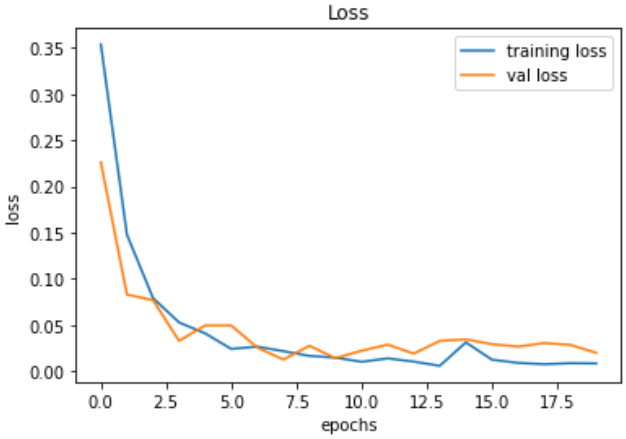



Fig. 3. Myocardial Infarction detection loss curve

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 12,2022 at 06:44:54 UTC from IEEE Xplore. Restrictions apply.

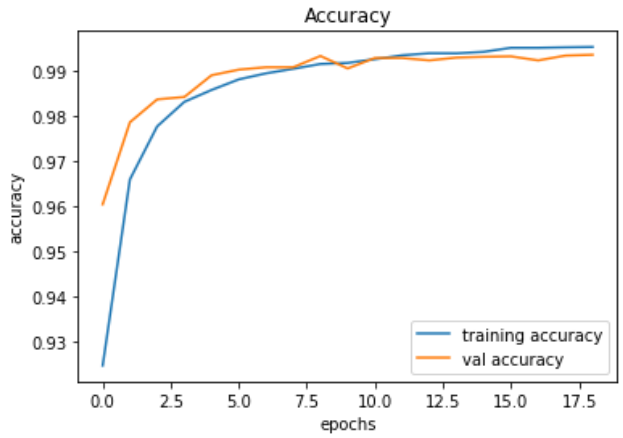



Fig. 4. Arrhythmia detection accuracy curve

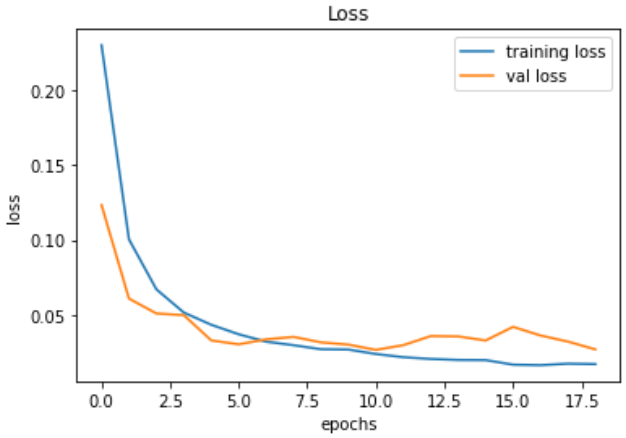



Fig. 5. Arrhythmia detection loss curve

## VI. Conclusion

Through this study we presented different machine learning and a CNN model for detection and classification of Arrhythmia and Myocardial Infarction from ECG signals. We trained SVM, KNN and ExtraTreesClassifier to assure that the data was sufficient to train a neural network. We then trained a convolutional neural network for Arrhythmia and Myocardial infarction classification and compared its result with conventional machine learning models. Our results indicate that the proposed CNN model has a good learning rate and a high accuracy of 98.57% and 99.38% in predicting Arrhythmia and Myocardial Infarction, respectively. The results also prove that it is indeed possible to push the boundaries of existing research by focusing on maximizing results instead of diversifying into new techniques that may or may not work. We hope that our study promotes deeper research in this field, with a more vertical approach and can help towards automation in medical systems.

## Advantages

We have managed to attain a high accuracy while maintaining relative simplicity in model design. This allows scope for future research – building upon the existing model architecture to achieve more optimal results. Moreover, resampling data in an intelligent manner has proved to result in noticeable improvements in model accuracy and precision making it simple but useful technique for classification of such diseases.

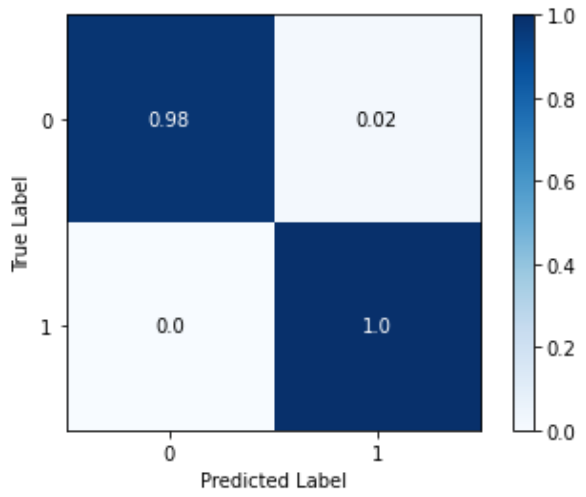



Fig. 6. Confusion matrix for predictions generated on test dataset of Myocardial Infarction

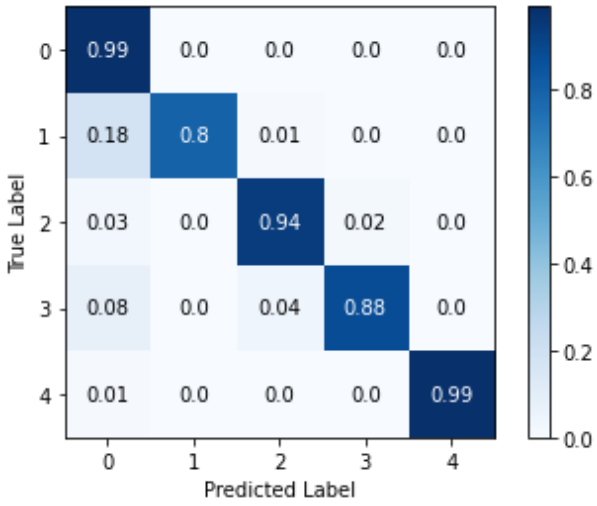



Fig. 7. Confusion matrix for predictions generated on test dataset of Arrhythmia

## Limitations

Even though there existed a total of 109446 samples in the MIT-BIH dataset, the number of training samples for *Supraventricular premature beat* were only 2223 and for the cases of *Fusion of Ventricular and Normal beat* were only 641. This was much lower in comparison to the number of samples for other categories. Even though we resampled the data to counter this and managed to achieve a high accuracy on a new unaltered test database, better results can be expected with more raw data.

## References

[1] P. Libby, P. Theroux, Pathophysiology of coronary artery disease, Circulation 111 (25) (2005) 3481–3488.

[2] H. Society, "Heart diseases and disorders," 2018. [Online]. Available: https://www.hrsonline.org/Patient-Resources/Heart-Diseases-Disorders

[3] V. Jahmunah, E.Y.K Ng, Tan Ru San, U. Rajendra Acharya, "Automated detection of coronary artery disease, myocardial infarction and congestive heart failure using GaborCNN model with ECG signals", Computers in Biology and Medicine, vol. 134, July 2021, 104457

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 12,2022 at 06:44:54 UTC from IEEE Xplore. Restrictions apply.

[4] Mohammad Kachuee, Shayan Fazeli, Majid Sarrafzadeh, "ECG Heartbeat Classification: A Deep Transferable Representation", 2018 IEEE International Conference on Healthcare Informatics (ICHI)

[5] R. J. Martis, U. R. Acharya, C. M. Lim, K. Mandana, A. K. Ray, and C. Chakraborty, "Application of higher order cumulant features for cardiac health diagnosis using ecg signals," International journal of neural systems, vol. 23, no. 04, p. 1350014, 2013.

[6] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych, and R. San Tan, "A deep convolutional neural network model to classify heartbeats," Computers in biology and medicine, vol. 89, pp. 389–396, 2017.

[7] T. Li and M. Zhou, "Ecg classification using wavelet packet entropy and random forests," Entropy, vol. 18, no. 8, p. 285, 2016.

[8] U. R. Acharya, H. Fujita, S. L. Oh, Y. Hagiwara, J. H. Tan, and M. Adam, "Application of deep convolutional neural network for automated detection of myocardial infarction using ecg signals," Information Sciences, vol. 415, pp. 190–198, 2017

[9] N. Safdarian, N. J. Dabanloo, and G. Attarodi, "A new pattern recognition method for detection and localization of myocardial infarction using t-wave integral and total integral as extracted features from one cycle of ecg signal," Journal of Biomedical Science and Engineering, vol. 7, no. 10, p. 818, 2014.

[10] J. Kojuri, R. Boostani, P. Dehghani, F. Nowroozipour, and N. Saki, "Prediction of acute myocardial infarction with artificial neural networks in patients with nondiagnostic electrogram"

[11] L. Sun, Y. Lu, K. Yang, and S. Li, "Ecg analysis using multiple instance learning for myocardial infarction detection," IEEE transactions on biomedical engineering, vol. 59, no. 12, pp. 3348–3356, 2012.

[12] B. Liu, J. Liu, G. Wang, K. Huang, F. Li, Y. Zheng, Y. Luo, and F. Zhou, "A novel electrocardiogram parameterization algorithm and its application in myocardial infarction detection.

[13] L. Sharma, R. Tripathy, and S. Dandapat, "Multiscale energy and eigenspace approach to detection and localization of myocardial infarction," IEEE Transactions on Biomedical Engineering, vol. 62, no. 7, pp. 1827–1837, 2015.

[14] Shayan Fazeli. ( 2018, June). ECG Heartbeat Categorization Dataset, Version 1. Retrieved October 2021 from https://www.kaggle.com/shayanfazeli/heartbeat/metadata.

[15] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.

[16] Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng in Med and Biol 20(3):45-50 (May-June 2001). (PMID: 11446209)

[17] Bousseljot R, Kreiseler D, Schnabel, A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische Technik, Band 40, Ergänzungsband 1 (1995) S 317

[18] Eduardo José da S. Luz, William Robson Schwartz, Guillermo Cámara-Chávez, David Menotti, ECG-based heartbeat classification for arrhythmia detection: A survey.

[19] Z. Masetic, A. Subasi, Congestive heart failure detection using random forest classifier, Comput. Methods Progr. Biomed. 130 (2016) 54–64. [9] J.A. Akoh, World J. Transplant. 1 (1) (2011) 4–12.

[20] V. Jahmunah, et al., "Computer-aided diagnosis of congestive heart failure using ECG signals – a review, Phys. Med. 62 (March) (2019) 95–104.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 12,2022 at 06:44:54 UTC from IEEE Xplore. Restrictions apply.

FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 481-486 | e-ISSN: 2347-470X

# Ensemble Deep Convolution Neural Network for Sars-Cov-V2 Detection

**Subrat Sarangi [1], Uddeshya Khanna[2] and Rohit Kumar[3]**

*[1,2,3]Department of Applied Mathematics, Delhi Technological University, Delhi, India, subratsarangi.dtu@gmail.com, uddeshya.khanna@gmail.com, rohitkumar@dtu.ac.in.*

***Correspondence:** Subrat Sarangi; Email: subratsarangi.dtu@gmail.com

**ABSTRACT-** The continuing Covid-19 pandemic, caused by the SARS-CoV2 virus, has attracted the eye of researchers and many studies have focussed on controlling it. Covid-19 has affected the daily life, employment, and health of human beings along with socio-economic disruption. Deep Learning (DL) has shown great potential in various medical applications in the past decade and continues to assist in effective medical image analysis. Therefore, it is effectively being utilized to explore its potential in controlling the pandemic. Chest X-Ray (CXR) images were used in studies pertaining to DL for medical image analysis. With the burgeoning of Covid-19 cases by day, it becomes imperative to effectively screen patients whose CXR images show a tendency of Covid-19 infection. Several innovative Convolutional Neural Network (CNN) models have been proposed so far for classifying medical CXR images. Moreover, some studies used a transfer learning (TL) approach on state-of-art CNN models for the classification task. In this paper, we do a comparative study of these CNN models and TL approaches on state-of-art CNN models and have proposed an ensemble Deep Convolution Neural Network model (DCNN).

**General Terms:** Neural Network, Deep Learning (DL), Covid-19, Chest X-Ray (CXR), Medical Image Analysis.

**Keywords:** Convolutional Neural Network (CNN), Deep Convolutional Neural Network (DCNN), DenseNet121, VGG19, Support Vector Machine (SVM).

ARTICLE INFORMATION

**Author(s)**: Subrat Sarangi, Uddeshya Khanna and Rohit Kumar.
**Received**: 30/04/2022; **Accepted**:20/07/2022; **Published**: 10/08/2022;
**e-ISSN**: 2347-470X;
**Paper Id**: IJEER100313;
**Citation:** 10.37391/IJEER.100313
**Webpage-link**:
https://ijeer.forexjournal.co.in/archive/volume-10/ijeer-100313.html

This article belongs to the Special Issue on **Recent Advancements in the Electrical & Electronics Engineering**

**Publisher's Note:** FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

## 1. INTRODUCTION

In March 2020, the World Health Organisation (WHO) declared COVID-19 as a pandemic and it is said to be one of the most detrimental pandemics in history. Initially, several names were used for the virus but WHO in January 2020 advocated the tentative name as a 2019-novel coronavirus (2019-nCov). Till 22 March 2020, the covid-19 epidemic had caused over 123 million infections and over 2.7 million deaths worldwide. The virus has varied symptoms ranging from some to critical illness and some of the common symptoms are cough, sore throat, loss of smell, difficulty in breathing, fever, dry cough, etc. The primary spread of this infection is through the air. A healthy person in close proximity to an infected person or in contact with a contaminated surface is highly probable to incur the infection. Therefore, it is always advised for an infected person to remain isolated during the infection period. Some infected patients don't develop any noticeable symptoms and don't get tested for the disease. These asymptomatic carriers tend to unknowingly spread the infection. Preventive measures can abate the risk of getting infected by Covid-19. It is advised to wear face masks and gloves in public places to avoid contact with contaminated surfaces and air particles.

Covid-19 patients suffer from lung infection and it can cause other lung issues such as ARDS (Acute Respiratory Distress Syndrome), pneumonia, and sepsis. It is essential for the world to fight covid-19 and screening potential covid-19 patients is one of the predominant tasks for taking a step against the covid-19 spread. Bio-Medical Image analysis is a popular research field and it has received a lot of attention in recent years. CXR images are widely available in various open databases for research and these databases have been used to classify CXR images to potential covid-19 cases using the DL technique. CXR images are preferred over CT images because CXR images are easily and cheaply available. DCNN technique has been widely used for classifying potential covid-19 positive CXR images by feature extraction and learning. The dataset is pre-processed and divided into training data, validation data, and testing data. Other hyperparameters such as optimizer, number of epochs, loss function, suitable batch size, are set for the model, and training data and validation data are used to learn the weights of the network. Furthermore, this trained model is used to classify the images in testing data.

## 2. SYSTEM MODEL

Rachna Jain et al. [1], in their paper, considered posteroanterior view scans of CXR for healthy and Covid-19 patients. The authors used three different DCNN models and compared these models using necessary evaluation metrics. The models considered were Xception Net (XN), Inception Net V3, ResNet.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 481-486 | e-ISSN: 2347-470X

It was concluded that the XN model gave the highest accuracy among the compared models. The dataset used was taken from the Kaggle repository and it constituted CXR scans of people affected with pneumonia, Covid-19, and healthy people. There were a total of 6432 CXR images, out of which 5467 images were considered for training and 965 images were considered for validation. The three models were trained using this dataset and used on the validation set for predictions. These models predicted whether the image in the validation set belonged to a normal, covid-19, or pneumonic person. Leaky RelU activation was used instead of ReLU to boost up the training system and avoid the issue of dead neurons. Categorical cross-entropy loss was used which optimizes the parameters used in the model. Moreover, Adam Optimizer (AO) was used with a learning rate of 0.001 for training. It was concluded that the XN model gave the highest testing data accuracy of 97% amongst the models the authors presented in their paper.

Mohd Zulfaezal Che Azemin et al. [2], in their study, used ResNet - 101, a CNN with 101 layers. The model's hyperparameters have been described briefly in *Table 1*. The dataset used for training and validation purposes was distilled from the CXR 14 dataset, which is a dataset of thoracic disorders amongst masses. The actual dataset had four labels namely: Fracture, Pneumothorax, Nodule - mass, Airspace Opacity. In this study, the authors made the intellectual decision of using only one label for training the model, airspace opacity which at the time was found to be linked to positive COVID - 19 cases. The authors based their study only on frontal CXR. The authors also introduced 5828 images labeled "no finding" to the dataset used in their study for the sake of simulating a population with a 2.57% spread rate. As the study was done at a time where there wasn't as much data available compared to a few months after the study was done, the study tends to have less desirable metrics compared to other models in our study. The performance metrics of the model are as follows: accuracy - 71.9 %, sensitivity - 77.3%, specificity - 71.8%.

Ioannis D. Apostolopoulos et al. [3], in their research, focused on evaluating the effectiveness of the models considered for the study. The models that were evaluated are complex and required large data sets for classification and feature extraction. However, the dataset used in the study is small so TL is used to train CNN models because of their capability to retain knowledge from one job and use it to perform another. The authors used two different CXR image datasets for their research. 224 covid-19 positive images, 700 confirmed common bacterial pneumonia images and 504 normal health patients' images were present in the first dataset. The second dataset was a slight modification of the first dataset. This dataset included 224 covid-19 positive cases images, 504 images of healthy patients, 400 bacterial, and 314 viral pneumonia patients. The study considered several state-of-art-CNN models and parameters tuned for TL purposes namely XN, Inception, MobileNet v2, Inception ResNet v2, and VGG19. These CNN models shared some common hyperparameters. All the convolutional network models used Rectified Linear Unit activation function (ReLU) and AO. The classification metrics observed in this study were confusion matrix, specificity and sensitivity. Furthermore, the authors calculated two different accuracies. The first accuracy, called the 3-class accuracy, was the total accuracy of the models in predicting the three classes. Nevertheless, 2-class accuracy was only related to Covid-19 patients. It was further observed that MobileNet v2 showed better results than VGG19 for specificity. Therefore, the authors concluded that MobileNet v2 was the best model for the specific classification task. The setup of the MobileNet v2 model was retained to perform the classification for the second dataset. Finally, the authors concluded that MobileNet v2 was the best model for the classification task amongst other models.

Arun Sharma et al. [4] used Artificial Intelligence-based techniques to classify CXR images for the screening of Covid-19 patients. A TL approach for the classification task to make the model efficient was implemented. The authors used CXR images for adults that were publicly available and performed 25 data augmentations on the dataset to increase its size for better analysis and 27 different datasets were created from it. Furthermore, the authors used all these datasets to train and validate 29 different models using the TL approach. One model from these 29 models was trained on the 1st dataset and 25 from the remaining models were trained on the 2nd to 26th datasets. The first convolutional layer (CL) used a filter size of 3 and 32 filters. The second CL used 64 filters, filter size 5, and the third CL used 128 filters and filter size 7. Finally, the last three models were trained with the last dataset with a different number of iterations and epochs than the previous 26 models. The three models with the highest accuracy were selected and presented as the best three models for the classification task.

Prabira Kumar Sethy et al. [5] used two different sets of datasets for their study. The first dataset contained positive COVID-19 X-ray images excluding MERS, SARS, and ARDS. However, the second dataset contained negative COVID-19 X-ray images. The authors considered these two datasets for DL feature extraction using several models separately. The models used in their study were DenseNet201, AlexNet, GoogleNet, VGG16, ResNet 101, ResNet18, ResNet 50, XN, InceptionV3, and InceptionResNetV2. After extracting the deep features from these models, the authors used the Support Vector Machine (SVM) to classify them. A wide range of classification metrics was used to test this study. These metrics on which the models were evaluated were Specificity, Accuracy, Matthews Correlation Coefficient (MCC), Sensitivity, False Positive Rate (FPR), Kappa and F1 score. It was finally concluded that the ResNet50 plus SVM model showed better results for classification than other considered models with an accuracy of 95.38 %.

Tulin Ozturk et al. [6] in their paper focused on building a DL model, DarkCovidNet, for the detection of Covid-19 cases using CXR images. The authors used their model for binary classification for images, the classes being Covid-19 and no findings. Moreover, they used their model for three-class classification, the classes being Covid-19, pneumonia and neither. The model uses an end-to-end architecture. The proposed model DarkCovidNet is inspired by the existing DarkNet model. DarkNet model is a classification model for real-time object detection and is a successful DL architecture.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 481-486 | e-ISSN: 2347-470X

The DarkCovidNet model has 17 CNN layers and each DarkNet layer has a CL, Batch Norm, and Leaky ReLU layers one after the other. The model used AO, cross-entropy loss function, the learning rate of 3e-3, and had a total of 1,164,434 parameters. This model was used for binary and multi-class classification using five-fold cross-validation. The authors split the dataset into 80% training and 20% validation data and trained the DarkCovidNet model for 100 epochs. Evaluation metrics used in this study were sensitivity, specificity, precision, F1-score, and accuracy. The model gave an average multi-class classification percentage accuracy of 87.02 % and the average binary classification percentage accuracy of 98.08%.

Shervin Minaee et al. [7] used a prominent concept of TL to fine-tune four popular DL models named ResNet18, ResNet50, DenseNet-121 and SqueezeNet, to predict Covid-19 positive cases from CXR images. These four models were evaluated and compared with each other to determine the model with the best performance. The authors also created a dataset COVID-Xray-5k from the Joseph Cohen dataset, COVID-Chestxray-dataset, and ChexPert dataset. This dataset was divided into 2,031 training samples and 3,040 testing samples and each image had a corresponding binary label that tells whether it is a Covid-19 positive case or not. The study used a TL approach because the labeled data for Covid-19 X-ray images was limited and it might not be possible to train a model from scratch. Each model went through 100 epochs for fine-tuning, learning rate of 0.0001, batch size of 20, used AO, and downsampling of images to 224*224. The authors found that the dataset was imbalanced which made sensitivity and specificity good evaluation metrics. Moreover, the authors also considered ROC and AUC for evaluation. Finally, it was concluded that for a sensitivity rate of 97.5%, these models achieved a high 90% average for the specificity metric. Furthermore, ResNet50 and SqueezeNet performed better than the other two models.

A Stacked Ensemble DL Model named CoVNet-19 was proposed by Priyansh Kedia et al. [8] The researchers collected 6,214 CXR scans from five different open-sourced datasets and divided the dataset into Training, Validation, and Test Set. Two pre-trained DCNN were used namely VGG-19 and DenseNet121 which minimized noise, bias, and variance problems. Therefore, both models were used for feature extraction from the X-Ray images. The hyperparameters used were 10 epochs, 0.001 learning rate, and AO. The trained model was a robust and highly efficient classification model. SVM classified the images and the model achieved 99.28% accuracy for the three-class classification and 99.71% for the binary classification. Moreover, the proposed model was evaluated using accuracy, F1 Score, precision, confusion matrix, recall, and MCC score.

Stefanos Karakanis et al. [9], in their paper, first implemented a Conditional Generative Adversarial Network (cGAN) to create artificial data, as at the time of their study very limited data was available. Employing cGAN to generate synthetic data, certainly has improved the overall performance metrics of their proposed models for COVID-19 classification. The discriminator for the implemented cGAN architecture constitutes a standard Convolution Neural Network. The authors used AO for Stochastic Gradient Descent. Followed by the generation of the synthetic CXR image data, the authors have presented two DL models for COVID-19 classification. For binary classification, Stefanos et al., have proposed a CNN, consisting of a uni-CL with filter size 32 and kernel 4 x 4. Further, they've used the ReLU Function and Max pooling layer for feature extraction with a dense layer sized 128. For the multiple classes architecture, they've used 5 CNN layers, each having a ReLU function. The first layer is similar to the binary classification model, but the second one consists of one stride for convolution with a filter size of 64 and 4 x 4 kernels. Following this was a max-pooling activation operation, and 3 more layers with 128, 128, and 256 filter sizes and equivalent kernel size. There are two more dense layers with 512 and 128 filter sizes respectively. The final layer uses the softmax function for multi-class output. The COVID-19 data has been extracted from a Github repository made available by researchers at Montreal University. Another dataset containing CXRs of normal and pneumonia cases available at Kaggle has also been used. The proposed model for binary classification on the modified datasets achieved a 100%, 98.3%, 98.7% sensitivity, specificity, and accuracy respectively. The multi-class model achieved 99.3%, 98.1%, 98.3% sensitivity, specificity and accuracy respectively. These although may seem very good performance metrics, a lot of further studies are required to check how good the proposed models actually are.
Asif Iqbal Khan et al. [10], have implemented a CNN model for the purpose of classifying COVID-19 infected patients from the rest, on CXR image data. The authors have proposed a model which can classify the CXR images into 3 classes of pneumonia, namely- Bacterial, Viral, COVID-19. The proposed model is called CoroNet and has been derived from the XN model architecture. The proposed model has 33,969,964 parameters of which 54,528 are non-trainable. Their study is based on 3 datasets, of which two were for training and the dataset used by Tulin Ozturk et al. in their paper, for testing purposes. Dataset 1 contained 290 COVID-19 infected CXR images and 2 contained 660 bacterial, 931 viral, and 1203 non-infected images, of which only 1300 were used for training purposes to keep the dataset balanced. The dataset used for testing purposes contains 157 COVID-19 images and around 1000 non-Covid-19 images. The proposed model was implemented on Keras and was trained on a dataset that was optimized using AO. CoroNet achieved an accuracy of 96.6%, precision of 93.17%, recall of 98.25%, and F1 score of 95.6% for the COVID-19 class. For further improved performance metrics the authors fine-tuned the model and reduced the 4-class architecture to 3 classes by clubbing the bacterial and viral pneumonia classes into one class which increased the overall accuracy of the model from 89.6% to 94.59%.

## 3. PROPOSED 1D-CNN MODELS

The end goal of our model is to differentiate between the CXR images of patients who have tested positive for Covid-19, a person who is physically fit and patients with Pneumonia. Our proposed model is based on DCNN framework for the accurate COVID-19 patients' detection with the use of trained DCNN on the collected datasets of CXRs. Our model is bi-stacked ensemble model.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 481-486 | e-ISSN: 2347-470X

In the primary level of our proposed model, we have collated two pre-trained models VGG19 and DenseNet-121. These models were chosen because of their high-performance metrics for the purpose of classification of the CXR images. The image noise was accounted for using the combinational characteristics from the two Deep Convolutional Neural Networks.

At the secondary level of our model, a SVM model was used for the classification purposes and trained on the features extracted from our primary model.

### 3.1 DenseNet121

This has 121 layers and advantages such as supporting reusing of features and solving the problem of vanishing gradient. This model was sourced and trained on the Imagenet dataset. As known, every layer of this model uses activation-maps of the preceding layer as the input. This considerably decreased the parameters which were trainable compared to a CNN model with equal layers.

### 3.2 VGG19

It consists of nineteen layers, with sixteen CLs and three dense layers. This model was sourced alongwith with the previously mentioned DenseNet 121 model. The input to our secondary layer of SVM model for the classification purposes has been obtained as a feature vector from the second last layer of this model which is a Dense Layer with thirty-two nodes.

### 3.3 Dataset Description

We have used multiple repositories which were available publicly. We then created one unified library of the datasets and made a random split between training, validation and testing set. The repositories that were used are mentioned in the *table*.

**Table 1. Datasets used in our proposed mode**

| Dataset | Reference | COVID-19 | Normal | Pneumonia |
|---|---|---|---|---|
| 1 | [11] | 230 | 1231 | 1400 |
| 2 | [12] | 15 | 965 | 1334 |
| 3 | [13] | 359 | 0 | 56 |
| 4 | [14] | 60 | 500 | 0 |
| 5 | [15] | 89 | 256 | 24 |
| Total | | 753 | 2952 | 2814 |

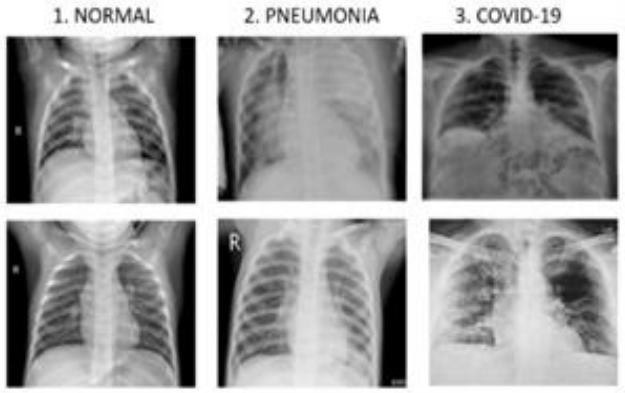



**Figure 1:** Random Images from our unified dataset

### 3.4 Model Hyperparameter

The reason behind using AO over others was the presence of finest properties from both the AdaGrad and RMSProp. It is significantly faster and highly reliable in reaching the Global minima with the pre-set hyperparameters. Binary Cross entropy and Categorical cross entropy were chosen as the loss function while training.

## 4. SIMULATION RESULT

After building our proposed model, we trained and tested it against the collected datasets. The results are further tabulated in detail in the subsequent sections.

### 4.1 Three Class Classification

**Table 2. Datasets used in our proposed mode**

| Training | P | R | F1 | Accuracy |
|---|---|---|---|---|
| COVID - 19 | 1.00 | 0.98 | 0.99 | 99.54% |
| Normal | 0.99 | 0.97 | 0.97 | |
| Pneumonia | 0.97 | 0.96 | 0.96 | |
| **Validation** | **P** | **R** | **F1** | **Accuracy** |
| COVID - 19 | 0.99 | 0.98 | 0.99 | 98.21% |
| Normal | 1.00 | 0.97 | 0.97 | |
| Pneumonia | 0.97 | 0.98 | 0.95 | |
| **Testing** | **P** | **R** | **F1** | **Accuracy** |
| COVID - 19 | 0.98 | 0.94 | 0.98 | 99.02% |
| Normal | 0.99 | 0.98 | 0.97 | |
| Pneumonia | 0.97 | 0.99 | 0.99 | |

### 4.2 Two Class Classification

**Table 3. Datasets used in our proposed mode**

| Training | P | R | F1 | Accuracy |
|---|---|---|---|---|
| COVID - 19 | 1.00 | 1.00 | 1.00 | 99.94% |
| Pneumonia | 0.99 | 0.99 | 1.00 | |
| **Validation** | **P** | **R** | **F1** | **Accuracy** |
| COVID - 19 | 1.00 | 0.99 | 0.96 | 99.55% |
| Pneumonia | 0.99 | 0.99 | 0.98 | |
| **Testing** | **P** | **R** | **F1** | **Accuracy** |
| COVID - 19 | 1.00 | 0.99 | 0.98 | 99.83% |
| Pneumonia | 0.97 | 1.00 | 1.00 | |

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 481-486 | e-ISSN: 2347-470X

## 5. MODEL COMPARISON & DISCUSSION

The aforementioned papers notably proposed either comparative studies of various CNN architectures or self-developed optimized models based on existing DL algorithms. The central methodology of almost all of these studies was to augment or collect the datasets fine-tune and optimize the models that included determining the number of layers in the selected neural network, minimum batch size, maximum epochs, and learning rate. The accuracy, specificity, precision, F1-score, and recall values were then compared with the other existing models. The models worked upon classifying the CXR images, present in the datasets, chiefly into 3 classes, Covid-19 Positive, Pneumonia and normal. The major drawback of most of the studies was the small size of training and testing sets, owing to the fact that most of these were conducted within the first 3 quarters of the year 2020. Small datasets raise concerns about overfitted and biased models. One of the solutions put forth in [4] was to generate 25 different types of augmentations using CLodDSA. Another concern with regard to the datasets is that they include CXR images from different age groups. Even though this makes classification of the images easy but the learning algorithm might become biased since the lung size of an adult age group is much larger than the paediatric age group. The models need to be trained accordingly to include this criterion. Almost all the models were trained on the datasets having limited diseases for classification. A better and more inclusive model would want to include other diseases and syndromes like SARS, ARDS, MERS, etc.

Models proposed [7], [3] and [4] use the TL CNN approach, a technique used to transfer the knowledge of a model usually trained on a larger dataset into a comparatively smaller dataset. This technique saves time and effort and is a great way to train, validate and test the CNN model and achieve high accuracy even with a limited number of images in the training set. The TL approach is then applied along with few of the best performing CNN models like Inception, VGG19, XN, MobileNet v2, and Inception ResNet v2. In [3], even though VGG19 has greater accuracy, MobileNet V2 outperforms it owing to greater specificity and hence is claimed as the best model with an accuracy of 97.40% (2-class) and 92.85% (3-class). MobileNet V2 has less number of False Negatives as compared to VGG19, which is crucial in this classification task since close to none suspected positive cases should be ideally classified incorrectly. Similarly, the model proposed in [10] predicts a low number of False Negatives, consequently increasing the recall value. In [5], the authors conclude that for models using the SVM algorithm for classification, ResNet50 gives the highest accuracy and specificity and the lowest FPR amongst other models considered by the author.

For the wide adaptability of the various models and to ensure quick processing and management of the CXR image dataset to aid the medical facilities, the models should be lightweight and free from sample bias. The Covid-Net model developed in [9] is a promising simple lightweight model for binary and multi-class classifications which uses CNN and ReLU function for feature extraction. The model has achieved an impressive accuracy of 98.7% in two class classification and 98.3% in three class classification. The model does not require pre-trained weights that explains the reduced model size, parameters, and complexity but on the other hand, would require training the model every time we have a new set of image dataset. Another shortcoming is that the dataset used to train the model chiefly consists of synthetically generated images with the aid of conditional cGANs. This can be resolved by training and testing the model on more real-world data to finetune it further. Our model uses Stacked Ensemble DL Model which makes use of 2 DCNNs and an SVC model which undoubtedly increases the overall accuracy of the model to 99.83% which is the highest of all the other papers reviewed but in turn, makes the whole model bigger and complicated that restricts its rapid adaptability. The model also limits the feature extraction to 32 for 1 DCNN and does not finetune every CL. A model should aim to work on real-time object and anomaly detection rather than being very deep. The model developed in [6] provides a novel addition in the form of the heatmaps that highlight the portion of concern in the CXR images. These heatmaps are for the reference of the radiologists to decipher the severity of the disease through manual analysis.

**Table 4. Comparison with other methods**

| Publication Time | Authors | Model | Accuracy |
|---|---|---|---|
| Aug 2020 | Mohd Zulfaezal Che Azemin et al. | ResNet 101 | 71.90% |
| Apr 2022 | Tulin Ozturk et al. | DarkCovidNet | 98.08% |
| Feb 2021 | Priyansh Kedia et al. | CovNet-19 | 99.71% |
| Dec 2020 | Stefanos Karakanis et al. | Covid-Net | 97.73% |
| May 2020 | Asif Iqbal Khan et al. | CoroNet | 96.60% |
| | Subrat Sarangi et al. | Proposed Model | 99.83% |

## 5. CONCLUSION & FUTURE SCOPE

The COVID-19 has posed a huge challenge in front of researchers, medical professionals, and scientists. The rapid and selfless response from some of the bright minds tried to share the burden of the already overwhelmed medical institutes. All the papers reviewed in our study provide models and combinations of algorithms to classify the CXR images the affected patients correctly. The problem of the limited size of the datasets available, overfitting, bias, inadequate parameters, and limited computational powers was skilfully handled by various models proposed.

The future scope of work includes training, validating, and testing our model on a large dataset made possible either through image augmentation or rapid testing for the virus. The model should also include various other pneumonia-based diseases to classify the images upon, in order to finetune the model. The model should be lightweight and shouldn't run on paid or bulky software so as to make sure that it can be used by a large number of medical agencies.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 481-486 | e-ISSN: 2347-470X

**Table 5. Comparison with best performing methods**

| Publication Time | Authors | Models Considered | Best Amongst Them | Accuracy of the Best Model |
|---|---|---|---|---|
| Oct 2020 | Rachna Jain et al. | 1. XN<br>2. Inception Net V3<br>3. RestNet | XN | 97.97% |
| Mar 2020 | Ioannis D. Apostolopoulos et al. | 1. Inception<br>2. VGG19<br>3. MobileNet V2<br>4. Inception ResNet V2<br>5. XN | MobileNet V2 | 97.40% |
| Jul 2020 | Arun Sharma et al. | 1. 24 Epoch-based model<br>2. 49 Epoch-based model<br>3. 101 Epoch-based model | 101 Epoch-based model using 60° rotated images | 66.67% |
| Mar 2020 | Prabira Kumar Sethy et al. | 1. VGG16<br>2. DenseNet 201<br>3. GoogleNet<br>4. ResNet 50<br>5. ResNet 18<br>6. Inception ResNet V2<br>7. XN<br>8. ResNet 101<br>9. AlexNet<br>10. Inception V3 | ResNet 50 | 95.38% |
| Apr 2020 | Shervin Minaee et al. | 1. ResNet 18<br>2. Squeezenet<br>3. ResNet 50<br>4. DenseNet 121 | SqueezeNet | 97.73% |

# REFERENCES

[1] R. Jain, M. Gupta, S. Taneja and D. Hemanth, "Deep learning based detection and analysis of COVID-19 on chest X-ray images", 2022.

[2] M. Che Azemin, R. Hassan, M. Mohd Tamrin and M. Md Ali, "COVID-19 Deep Learning Prediction Model Using Publicly Available Radiologist-Adjudicated Chest X-Ray Images as Training Data: Preliminary Findings", 2022.

[3] I. Apostolopoulos and T. Mpesiana, "Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks", 2022.

[4] A. Sharma, S. Rani and D. Gupta, "Artificial Intelligence-Based Classification of Chest X-Ray Images into COVID-19 and Other Infectious Diseases", 2022.

[5] Preprints.org, 2022. [Online]. Available: https://www.preprints.org/manuscript/202003.0300/v1/download. [Accessed: 30- Apr- 2022].

[6] T. Ozturk, M. Talo, E. Yildirim, U. Baloglu, O. Yildirim and U. Rajendra Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images", 2022.

[7] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani and G. Jamalipour Soufi, "Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning", 2022.

[8] P. Kedia, Anjum and R. Katarya, "CoVNet-19: A Deep Learning model for the detection and analysis of COVID-19 patients", 2022.

[9] S. Karakanis and G. Leontidis, "Lightweight deep learning models for detecting COVID-19 from chest X-ray images", 2022.

[10] A. Khan, J. Shah and M. Bhat, "CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images", 2022

[11] M. Chowdhury et al., "Can AI Help in Screening Viral and COVID-19 Pneumonia?" 2022.

[12] D. Kermany, K. Zhang and M. Goldbaum, "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification", Mendeley Data, 2022. [Online]. Available: https://data.mendeley.com/datasets/rscbjbr9sj/2. [Accessed: 30- Apr- 2022].

[13] J. Cohen, P. Morrison, L. Dao, K. Roth, T. Duong and M. Ghassemi, "COVID-19 Image Data Collection: Prospective Predictions Are the Future", arXiv.org, 2022. [Online]. Available: https://arxiv.org/abs/2006.11988. [Accessed: 30- Apr- 2022].

[14] "GitHub - agchung/Figure1-COVID-chestxray-dataset: Figure 1 COVID-19 Chest X-ray Dataset Initiative", GitHub, 2022. [Online]. Available: https://github.com/agchung/Figure1-COVID-chestxray-dataset. [Accessed: 30- Apr- 2022].

[15] "COVID-19 X rays", Kaggle.com, 2022. [Online]. Available: https://www.kaggle.com/andrewmvd/convid19-x-rays?select=X+rays [Accessed: 30- Apr- 2022].

© 2022 by Subrat Sarangi, Uddeshya Khanna and Rohit Kumar. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

*2022 2nd International Conference on Intelligent Technologies (CONIT)*
*Karnataka, India. June 24-26, 2022*

# Ensemble of Hybrid CNN Models for Brain Tumor Classification

Indu Singh
*Department of Computer Science & Engineering*
*Delhi Technological University*
Delhi, India
indusingh@dtu.ac.in

Rahul Kumar Yadav
*Computer Science & Engineering*
*Delhi Technological University*
Delhi, India
rahulkumaryadav25102000@gmail.com

Rahul Virahaniya
*Computer Science & Engineering*
*Delhi Technological University*
Delhi, India
rahulvirahaniya@gmail.com,

Noor Alam
*Computer Science & Engineering*
*Delhi Technological University*
Delhi, India
nooralam2151999@gmail.com

***Abstract*—A brain tumor develops due to the abnormal cell formation within the brain. The greatest challenge before initiating treatment is detecting and classifying tumors from brain MRI pictures since brain tumors have a wide range of appearances and there is a lot of overlap between tumor and normal tissues. In this paper, we propose a novel approach for the diagnosis of brain tumor using an ensemble of three hybrid CNN models. We used three different classifiers with CNN, Support Vector Machine, Random Forest and Multinomial Logistic Regression to classify brain tumors. In order to increase the accuracy, the hyper-parameters of the CNN model are tuned using KerasTuner. The proposed Ensemble Learning approach gives promising results in the classification of brain tumor when evaluated on Figshare brain tumor dataset and achieves an accuracy of 96.08% and thus is a trustworthy alternative to human experts for the brain tumor classification.**

***Keywords—Brain Tumor, Convolutional Neural Network, Random Forest, Multinomial Logistic Regression, Ensemble Learning***

## I. INTRODUCTION

Brain Tumor occurs due to the uncontrollable and irregular growth of the cells in the human brain. With time, these cells continue to multiply. These aberrant cells harm the healthy cells and continue to grow over time, thus disrupting typical human body functions. These cells divide uncontrollably, resulting in the formation of new daughter cells. The safest method to diagnose brain tumors is via Magnetic Resonance Imaging (MRI). MRI works by generating radio waves and high magnetic fields, which take a short image of the brain. As a result, correctly identifying the different types of brain tumors becomes the most crucial task. Hence, a cursory examination misclassification by human specialists can prove to be fatal, and the motivation to develop a model intensifies, which can act as a credible second opinion for human experts.

A meningioma tumor develops from the meninges, a layer of tissue that covers the brain and spinal cord. It is usually low-grade tumors, slow-growing, and unlikely to spread. Glioma tumors arise in the gluey supporting cells of the brain and spinal cord that help to keep the neurons in place and functioning well. Gliomas are malignant (cancerous), but some of them grow slowly. They are primary brain tumors, which means they develop in the brain's tissue. Gliomas usually spread outside the brain or spine, but they can be life-threatening because they are difficult to reach and cure with surgery and can grow in other parts of the brain. A pituitary tumor develops inside the pituitary gland, located behind the back of the nose near the brain over time, and can affect hormone levels in the body. They don't spread to other parts of the body and can cause an overproduction of hormones, causing other health problems.

Convolutional Neural Network is a feature extractor that is used for image classifications. It assigns weights and biases to the image features. CNN models can be trained to learn the characteristic features of different image classes through filters. Random Forest is a popular technique for both regression and classification. In classification tasks, it trains many decision trees and takes their majority vote for selecting a class. SVM can be used to solve regression and classification problems. SVM finds a hyperplane in an N-dimensional space that distinctly categorizes data points. Multinomial Logistic Regression is an extension of binary logistic regression, used for multiclass problems that have the possibility of multiple outcomes. It uses maximum likelihood estimation to determine the probability for categorical membership.

The proposed algorithm is a novel approach by ensembling three hybrid CNN models to accurately categorize brain tumors into glioma, meningioma, and pituitary tumor. Hyper-parameter tuning of the CNN model is done using KerasTuner to achieve better features and accuracy. The features extracted from the CNN model are provided to the classifiers, namely, Random Forest, SVM, and Multinomial Logistic Regression, for the classification. These three hybrid models are used together to generate a final prediction, which provides a better outcome than each individual model.

The rest of the paper is structured as follows. Section II highlights the literature review, while Section III provides a brief overview of the Brain Tumor Dataset's key features. The Section IV describes the proposed approach. The tests and results of the proposed algorithm are discussed in Section V. Finally, Section VI draws the conclusions and discusses future work.

2022 International Conference on Intelligent Technologies (CONIT) | 978-1-6654-8407-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CONIT55038.2022.9848012



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:57:26 UTC from IEEE Xplore. Restrictions apply.

## II. Related Work

Sanjeev Kumar et al. (2017) [1] introduced a technique for feature extraction using Discrete Wavelet Transform. To reduce the number of features, Genetic Algorithm was used and SVM was used for classification of brain tumor.

Mustafa R et al. (2018) [2] proposed a method for classifying brain tumor using MRI images, which uses statistical as well as neural network methods. A combination of 2D Gabor filter and 2D Discrete Wavelet Transform techniques were used for feature selection.

Amin Kabir et al. (2019) [3] proposed a CNN and Genetic Algorithm to classify brain tumors. GA gave the best model which is then bagged with an ensemble approach for decreasing the precision error variations. The CNN model reached 91.3% accuracy while the CNN-GA achieved 94.2% accuracy.

Nyoman Abiwinanda et al. (2019) [4] discovered an ideal CNN design for classification of brain tumor, which consisted of two convolution layers, activation layer (ReLU), and max pooling layer and finally followed by a layer of 64 neurons. The convolution layers had 32 filters of size 3x3 in this architecture.

S. Deepak et al. (2020) [5] classified medical pictures using a mix of SVM and CNN features. The completely automated approach was tested using MRI scans from the three kinds of brain tumors from the Figshare open dataset. A multiclass SVM was combined with CNN features for improved performance. A cross-validation process was carried out to test and evaluate the integrated system.

Angona Biswas et al. (2021) [6] proposed K-means clustering algorithm for preprocessing and ANN for the classification of brain tumors. 2D-DWT algorithm was used for feature extraction from the clustered images. In order to effectively reduce the number of features, the PCA algorithm was used.

## III. Figshare Brain Tumor Dataset

The Figshare Brain Tumor Dataset is a widely used dataset created by Jun Cheng of the School of Biomedical Engineering [7]. It contains 3064 T1-weighted contrast-enhanced brain tumor images and has images for three brain tumor categories. The categories include glioma tumor (1426 images), meningioma tumor (708 images) and pituitary tumor (930 images). It provides MAT files with each file containing a struct with the label indicating its category, id of the patient, data of the tumor image in 512*512 size, coordinates of the discrete points on brain tumor borders stored inside a vector, and mask images data having binary values where 1s indicate region of the brain tumor. In our research, we have used images and labels data from the dataset.

## IV. Proposed Approach

In this paper, we propose a novel approach for the classification of brain tumors into three classes: Glioma, Meningioma, and Pituitary Tumor. This approach accurately uses the ensemble of three hybrid CNN models to classify tumors into their respective categories. The hybrid CNN models used are CNN-RF, CNN-SVM, and CNN-Multinomial Logistic Regression.

The initial size of dataset images is 512*512 pixels which is then converted into 128*128 pixels for model training to reduce the computational cost of training. The ratio of training and validation used is 8:2, i.e., 2451 images for training while 613 images for validation.

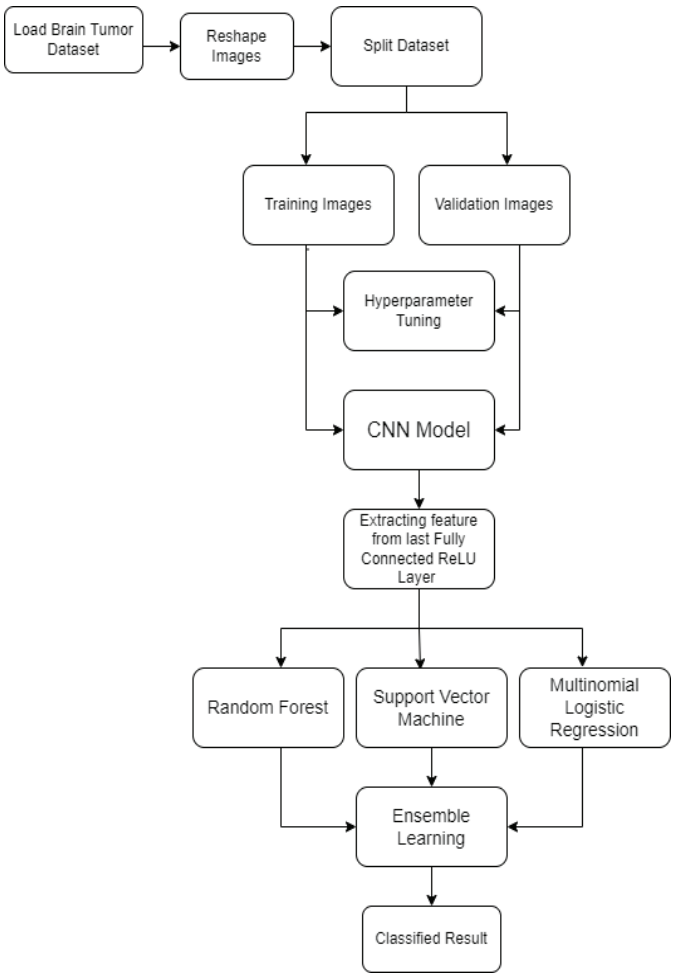



Fig. 1. Architecture of Proposed Approach

KerasTuner is used to obtain optimized hyper parameter values, which are then used to construct the CNN model. The CNN model consists of 5 convolutional layers. Convolutional layers take the input image and extracts features from it. Each of the convolutional layers has a ReLU activation function. Each convolutional layer is followed by a Max pooling layer to cut down on the number of samples obtained from the previous layer and reduce the feature map's size while still keeping important features. A dropout layer follows each Max pooling layer to avoid the problem of overfitting. The softmax layer at the end receives data from the last fully connected layer and uses the softmax function for performing classifications. After that, the fully connected layer having ReLU as the activation function is given as input to three different classifiers: Random Forest, SVM, and Multinomial Logistic Regression.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:57:26 UTC from IEEE Xplore. Restrictions apply.

At the end, the Ensemble learning technique is used to achieve higher accuracy of predictions than each of the individual hybrid CNN models. It combines these hybrid CNN models and results in an improvement in predictive performance. The hybrid CNN models and Ensemble learning technique are evaluated and compared by their performance metrics. The flowchart of the working procedure for our proposed approach is described in Fig. 1.

### A. CNN Model

#### 1) Hyper-Parameter Tuning

It is the process of selecting ideal hyperparameters for learning algorithms. Hyperparameters are model arguments, and their values are determined before starting the learning process. Forty percent of the dataset is provided to the tuner to obtain the value for the hyperparameters. KerasTuner is used for hyper-parameter tuning in the proposed approach.

#### 2) CNN Architecture

This section describes the CNN model's structures, layers, and parameters. The model has five 2D convolutional layers. A convolutional layer takes the input image and extracts features to produce a feature map. The number of channels for each convolutional layer and the filter size are obtained from the results of hyper parameter tuning. The channels for each layer are 314, 470, 32, 32 and 32 respectively, and each layer has a filter size of 5*5.

Each of the convolutional layers has a ReLU activation function. The ReLU activation function is computationally more efficient to compute than other functions like Sigmoid since ReLU only needs to pick max(0, x) instead of doing expensive exponential operations. A pooling layer of kernel size 5*5 is used after every convolutional layer to reduce the number of parameters to be considered by extracting representative features from inputs. This results in the reduction of computation and overfitting chances. The pooling layer downsamples the image in terms of width and height while channels remain the same. A dropout layer is used after each 2D convolutional and also after the fully connected layer having ReLU. The dropout layer is being used in order to prevent the overfitting problem. The dropout rate is kept at 0.2 as obtained from the result of hyper parameter tuning.

The final output of convolutional and pooling is flattened into a fully connected layer. These features are then combined, and the model is created. The outputs are classified into different classes using the softmax activation function. The outputs are normalized using the softmax function. The equation for the softmax function is given below:

$$s(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{1}$$

where n is the number of classifiers, $e^{x_i}$ is the input exponential function while $e^{x_j}$ is the exponential function for output.

### B. Classification

To improve the accuracy of the CNN model, various classifiers are used along with it and their performance metrics are evaluated.

#### 1) Support Vector Machine

This algorithm is a supervised learning method for performing analysis related to regressions and classifications. Here, every point is represented as some data item in n-dimensional space wherein the feature values give the values for that particular coordinate. The points are represented in space and mapped such that a hyperplane divides the examples of the different categories. The new values are then mapped into the same space, generating classification predictions depending on which side of the plane they lay on. Non-linear SVM is used in the proposed approach to classify brain tumors. The RBF kernel function is used to map the original data into a high-dimensional feature space and the equation for same is given below:

$$f(x_i, x_j) = \exp[\,\gamma||\,x_i - x_j||^2\,] \tag{2}$$

Intuitively, the hyperplane having the largest distance from the closest training data point of any class achieves decent separation. Since with higher margin, the generalization error of the classifier becomes smaller. The advantage of SVM over other algorithms is that it chooses the line that separates the data and is at the largest distance from the closest data points possible.

#### 2) Random Forest

Random forest is a meta estimator that prevents overfitting by averaging the results of several decision tree classifiers fitted to various dataset sub-samples. It is based on the ensemble learning idea. It collects predictions from each tree instead of depending on a single decision tree and predicts the final output on the basis of the vote majority.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{3}$$

The above formula is used to calculate Gini index and determines which class has probability of occurrence. C is representing the number of classes and $p_i$ is the relative frequency of observed class.

#### 3) Multinomial Logistic Regression

In logistic regression, logistic equations are used to estimate probabilities, subsequently used in statistical software to understand the relationship between dependent and independent variables. This form of analysis can assist in predicting the probability of an occurrence or a decision occurring. Multinomial logistic regression is its extension and is used in multi-class problems where the independent variables can have more than two classes. In case of multinomial logistic regression, the probability of x belonging to a class i is calculated as per the equation given below:



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:57:26 UTC from IEEE Xplore. Restrictions apply.

$$P\left(y^{(i)} = 1 | x, w\right) = \frac{\exp\left(w^{(i)^T} x\right)}{\sum_{j=1}^{m} \exp\left(w^{(j)^T} x\right)} \quad (4)$$

where $w^{(i)}$ is weight vector for class i and the $^T$ superscript represents vector transpose.

### C. Ensemble Learning

The ensemble learning technique uses multiple learning algorithms and obtains their combined performance. It is a powerful technique to improve the prediction results by combining various models into a single, reliable and better model. In our approach, Ensemble learning is used on three hybrid CNN models: CNN-RF, CNN-SVM, and CNN-Multinomial Logistic Regression. These models are layered together and generate a final prediction, giving a better outcome than the individual models.

**Algorithm 1** Algorithm for Proposed Approach

**Input: X_train:** Training images, **Y_train:** Training labels, **X_val:** Validation images, **Y_val:** Validation labels, **e:** Total number of epochs, **b:** Batch size
**Output:** Ensemble of hybrid CNN models

1. findHyperparameters():
2. for **layers** in range (2, 8):
3. model.add(conv2d filter in range (32, 512))
4. model.add(maxPooling in range (2, 5))
5. model.add(dropoutLayer in range (0.2, 0.5))
6. **randomSearch()** for best hyperparameters [maxTrials = 6, executionsPerTrail = 8]
7. Build **CNN** model with best hyperparameters
8. Compile and Train **CNN** model
9. for **i** ← 1 to **e** (for each epoch) do:
10. for **j** ← 1 to **b** (for each batch) do:
11. $\hat{\mathbf{y}}$ ← cnn(features)
12. **loss** ← crossEntropy(y, $\hat{\mathbf{y}}$)
13. optimization(**loss**)
14. accuracy()
15. **bestAccuracy** ← max(accuracy, bestAccuracy)
16. Evaluate **CNN** model
17. **rf** ← RandomForestClassifier()
18. Train hybrid **CNN-RF** model
19. **rf.fit**(X_train_out,**Y_train**)
20. **svm** ← SVM(C = 1.0) [C is regularization parameter]
21. Train hybrid **CNN-SVM** model
22. **svm.fit**(X_train_out,**Y_train**)
23. **lr** ← LogisticRegression(multi_class = 'multinomial')
24. Train hybrid **CNN-MLR** model
25. **lr.fit**(X_train_out,**Y_train**)
26. def **getStacking()**:
27. Base model is defined
28. **level0**.append(RandomForestClassifier())
29. **level0**.append(SVM())
30. **level0**.append(MultinomialLogisticRegression())
31. Meta learner model is defined
32. **level1** ← RandomForestClassifier()
33. Stacking ensemble is Defined
34. **model** ← **StackingClassifier**(estimators = **level0**, final_estimator = **level1**)
35. return model
36. **ensembleModel** ← **getStacking**()
37. **Ensemble model** is evaluated and performance metrics are calculated

Fig. 2. Algorithm for Proposed Approach

## V. Experiments And Results

To experiment with the proposed approach, the steps as shown in Fig. 2 were followed. The proposed approach was evaluated on the Figshare Brain tumor dataset. KerasTuner is used to find CNN model's optimal parameter values, such as the number of convolutional layers, dropout layers, learning rate, etc. The values of various hyper-parameter values received from the KerasTuner are shown in Table I. These values are then used to construct and train the CNN model. The number of epochs were kept 50 with batch size 32.

TABLE I. Hyperparameters Obtained From Kerastuner

| Hyperparameters | Value |
|---|---|
| Number of convolutional layers | 5 |
| Learning rate | 0.001 |
| Pool Size | 5 |
| Drop rate | 0.2 |
| Optimization Function | ADAM |

Fig. 3 depicts the training and validation accuracy and Fig. 4 depicts the training and validation loss for the CNN model. The model obtained validation accuracy of 94.77% and the validation loss was 0.22%.

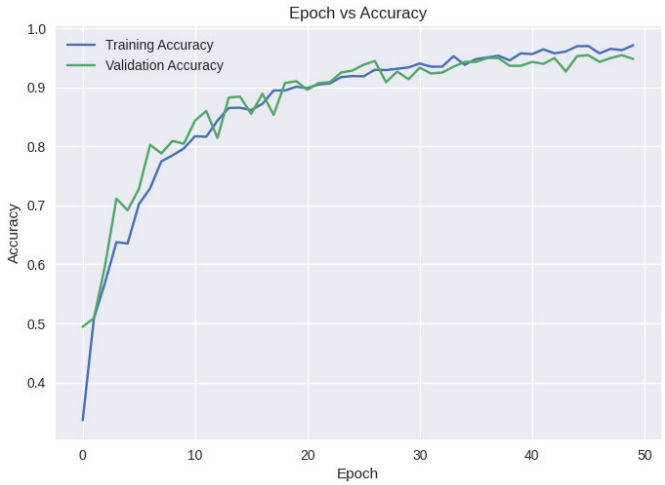



Fig. 3. Training and Validation Accuracy (CNN)



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:57:26 UTC from IEEE Xplore. Restrictions apply.

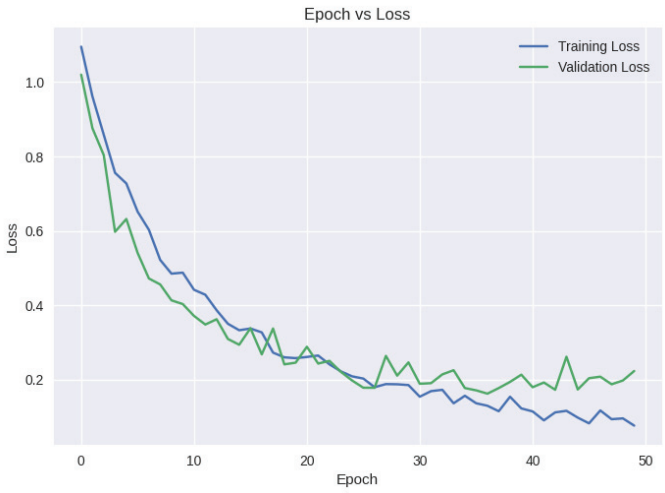



Fig. 4. Training and Validation Loss (CNN)

The loss values from Fig. 4 indicate how good the prediction of the model is on a single input. It can be observed that the errors are being efficiently decreased. The second last fully connected layer of the CNN model is provided to three different classification algorithms, namely Random Forest, SVM, and Multinomial Logistic Regression.

In the CNN-RF hybrid model, Random Forest Classifier was configured with parameters values 35 for maximum depth of the tree, 75 for the number of trees in forest and the criterion used was gini. In the CNN-SVM model, the regularization parameter C was set to 0.1, while the CNN with Multinomial Logistic Regression model had parameters C as 0.2, solver being newton-cg and maximum iterations for solver were set to 1000. After evaluating different classifiers in combination with CNN, the hybrid models CNN-RF, CNN-SVM, and CNN-Multinomial Logistic Regression achieved validation accuracies of 95.92%, 94.75%, and 95.75%, respectively.

In the Ensemble Learning technique, the Stacking Classifier was provided with an estimator list containing the three classifiers, the Random Forest classifier as the final estimator, and 5-fold cross validation was selected. The Ensemble learning technique achieved 96.08% accuracy. This technique achieved greater accuracy than the CNN-RF, CNN-SVM, and CNN with Multinomial Logistic Regression model, which demonstrates that the Ensemble Learning technique is a better, more accurate, and dependable method for classification of brain tumors.

The confusion matrix of CNN model and Ensemble learning in Fig. 5, and Fig. 6, respectively, show the predictions of these models for classifying the brain tumor. It provides a clear view of accurate predictions and deviations from actual predictions. The much higher values in the diagonal as compared to other indexes in the matrix shows the correctness of the model.

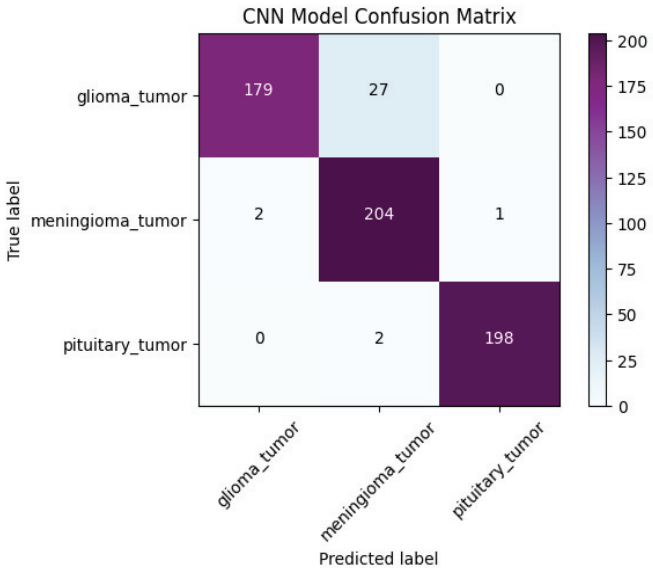



Fig. 5. Confusion Matrix (CNN model)

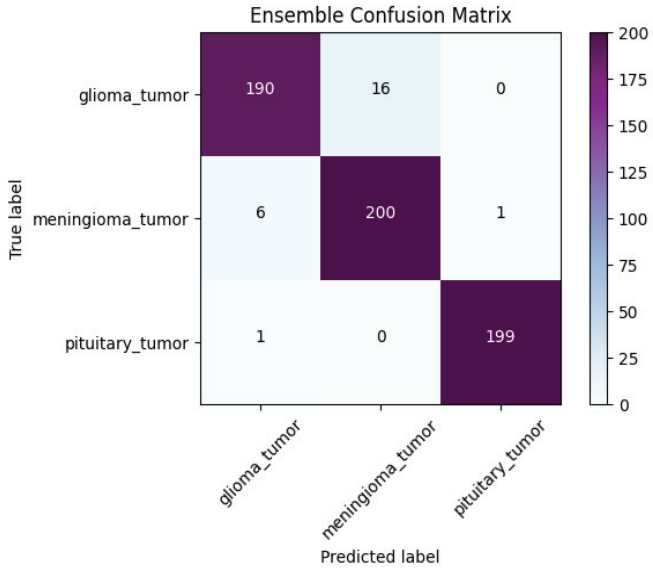



Fig. 6. Confusion Matrix (Ensemble Learning)

The performance metrics can be calculated using the true positives, true negatives, false positives and false negatives observed from the confusion matrix. In Table II, values of performance metrics of the Ensemble learning method indicate its better performance in predictions.

Precision, sensitivity, specificity and F1 score were evaluated with following formulas and recorded in Table II.

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Sensitivity\ (recall) = \frac{TP}{TP+FN} \quad (6)$$



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:57:26 UTC from IEEE Xplore. Restrictions apply.

$$Specificity = \frac{TN}{TN+FP} \quad (7)$$

$$F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (8)$$

where TP, TN, FP and FN represent the number of true positives, true negatives, false positives and false negatives respectively.

TABLE II. ENSEMBLE LEARNING PERFORMANCE

| Type | Precision | Sensitivity | Specificity | F1 Score |
|---|---|---|---|---|
| Glioma | 96% | 92% | 98% | 94% |
| Meningioma | 93% | 97% | 96% | 95% |
| Pituitary | 99% | 99% | 100% | 99% |

The results show the Ensemble learning method has the highest accuracy for classification. When compared with other authors work, the proposed approach achieves higher accuracy than previous works. The comparison between different authors, models, and their accuracies is carried out and the results were recorded in Table III.

TABLE III. COMPARISON WITH OTHER EXISTING METHODS

| Author Name | Approach | Accuracy |
|---|---|---|
| Pashaei et al. [8] | CNN + ELM | 93.68 % |
| Das et al. [9] | CNN | 94.39 % |
| Preethi et al. [10] | Wavelet Texture Features + Deep Neural Networks | 92 % |
| Ayadi et al. [11] | CNN | 94.74 % |
| Deepak et al. [5] | CNN + SVM | 95.82 % |
| Biswas et al. [6] | K-Means Clustering + ANNs | 95.40 % |
| **Proposed Approach** | **Ensemble of Hybrid CNN Models** | **96.08 %** |

The proposed Ensemble learning approach employing three hybrid CNN classifiers is clearly superior to other existing methods, as seen by the above comparison.

## VI. CONCLUSION AND FUTURE WORK

The proposed approach uses an Ensemble learning technique using three hybrid CNN models to achieve maximum accuracy in the classification of brain tumor images into their respective categories of glioma, meningioma or pituitary tumor. The model is trained and validated using the Figshare brain tumor dataset with 80:20 ratio. The images of the whole dataset are resized before being used for the training and validation purposes. The CNN model is formed using parameters obtained through hyperparameter tuning performed using KerasTuner. The CNN model reached an accuracy of 94.77%. Further, three classifiers, Random Forest, SVM, and Multinomial Logistic Regression, were used in combination with the CNN model to form three hybrid models. These hybrid models are layered together using the Ensemble learning technique to achieve better performance in brain tumor classification. Ensemble learning technique performed better than other models achieving an accuracy of 96.08%, while CNN-RF, CNN-SVM, CNN-Multinomial Logistic Regression models achieved 95.92%, 94.75%, and 95.75%, respectively. High accuracy of models is necessary for image classification, especially in medicine, as there can be a very minimal amount of samples present for different categories. The proposed approach was found to be a reliable algorithm for brain tumor classification when compared to previous research in the literature. This approach has significantly less need for preprocessing the input data, making its model robust and efficient.

For future work, we plan to work with 3D images and extend our approach by introducing metaheuristic algorithms to increase accuracy. Our future research will focus on a more precise formulation of unique hybrid swarm algorithms, which can further improve the system's performance.

## REFERENCES


[1] Kumar, S., Dabas, C., & Godara, S. (2017). Classification of brain MRI tumor images: a hybrid approach. Procedia computer science, 122, 510-517.

[2] Ismael, M. R., & Abdel-Qader, I. (2018, May). Brain tumor classification via statistical features and back-propagation neural network. In 2018 IEEE international conference on electro/information technology (EIT) (pp. 0252-0257). IEEE.

[3] Anaraki, A. K., Ayati, M., & Kazemi, F. (2019). Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms. biocybernetics and biomedical engineering, 39(1), 63-74.

[4] Abiwinanda, N., Hanif, M., Hesaputra, S. T., Handayani, A., Mengko,T.R.(2019).Brain tumor classification using convolutional neural networks. In World congress on medical physics and biomedical engineering 2018(pp. 183-189). Springer,Singapore.

[5] Deepak, S., & Ameer, P. M.(2021).Automated categorization of brain tumor from mri using cnn features and svm. Journal of Ambient Intelligence and Humanized Computing, 12(8), 8357-8369.

[6] Biswas, A., & Islam, M. S. (2021, January). Brain tumor types classification using k-means clustering and ANN approach. In *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)* (pp. 654-658). IEEE.

[7] Jun Cheng, 2017 Figshare dataset https://figshare.com/articles/brain_tumor_dataset/1512427

[8] Pashaei, A., Sajedi, H., & Jazayeri, N. (2018, October). Brain tumor classification via convolutional neural network and extreme learning machines. In 2018 8th International conference on computer and knowledge engineering (ICCKE) (pp. 314-319). IEEE.

[9] Das, S., Aranya, O. R. R., & Labiba, N. N. (2019, May). Brain tumor classification using convolutional neural network. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)* (pp. 1-5). IEEE.




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:57:26 UTC from IEEE Xplore. Restrictions apply.

[10] Preethi, S., & Aishwarya, P. (2019). Combining wavelet texture features and deep neural network for tumor detection and segmentation over MRI. Journal of Intelligent Systems, 28(4), 571-588.

[11] Ayadi, W., Elhamzi, W., & Atri, M. (2020, December). A new deep CNN for brain tumor classification. In 2020 20th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA) (pp. 266-270). IEEE.

[12] Afshar, Parnian, Konstantinos N. Plataniotis, and Arash Mohammadi. "Capsule networks for brain tumor classification based on MRI images and coarse tumor boundaries." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[13] Anaraki, A. K., Ayati, M., & Kazemi, F. (2019). Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms. biocybernetics and biomedical engineering, 39(1), 63-74.

[14] Hossain, T., Shishir, F. S., Ashraf, M., Al Nasim, M. A., & Shah, F. M. (2019, May). Brain tumor detection using convolutional neural networks. In 2019 1st international conference on advances in science, engineering and robotics technology (ICASERT)(pp. 1-6). IEEE.

[15] Telrandhe, S. R., Pimpalkar, A., & Kendhe, A. (2016, February). Detection of brain tumor from MRI images by using segmentation & SVM. In 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave)(pp. 1-6). IEEE.

[16] D. Zikic et al., "Segmentation of brain tumor tissues with convolutional neural networks," MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS), pp. 36–39, 2014.

[17] Lyksborg, M., Puonti, O., Agn, M., & Larsen, R. (2015, June). An ensemble of 2D convolutional neural networks for tumor segmentation. In the Scandinavian Conference on Image Analysis (pp. 201-211). Springer, Cham.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:57:26 UTC from IEEE Xplore. Restrictions apply.

Journal of Molecular Modeling (2022) 28:319
https://doi.org/10.1007/s00894-022-05310-9

ORIGINAL PAPER

# Epicatechin analogues may hinder human parainfluenza virus infection by inhibition of hemagglutinin neuraminidase protein and prevention of cellular entry

**Sidharth Bhasin**[1] · **Megh Nadar**[1] · **Yasha Hasija**[1]

Received: 17 January 2022 / Accepted: 30 August 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

## Abstract

Human parainfluenza viruses (HPIVs) are (−)ssRNA viruses belonging to *Paramyoviridaie* family. They are one of the leading causes of mortality in infants and young children and can cause ailments like croup, bronchitis, and pneumonia. Currently, no antiviral medications or vaccines are available to effectively treat parainfluenza. This necessitates the search for a novel and effective treatment. Computer-aided drug design (CADD) methodology can be utilized to discover target-based inhibitors with high accuracy in less time. A library of 45 phytocompounds with immunomodulatory properties was prepared. Thereafter, molecular docking studies were conducted to characterize the binding behavior of ligand in the binding pocket of HPIV3 HN protein. The physicochemical properties for screened compounds were computed, and the top hits from docking studies were further analyzed and validated using molecular dynamics simulation studies using the Desmond module of Schrodinger Suite 2021–1, followed by MM/GBSA analysis. The compounds CID:72276 (**1**) and CID:107905 (**2**) emerged as lead compounds of our in silico investigation. Further in vitro studies will be required to prove the efficacy of lead compounds as inhibitors and to determine the exact mechanism of their inhibition.

**Keywords** Drug discovery · HPIV · Molecular docking · Molecular dynamics · Parainfluenza · Phytocompounds

## Abbreviations

| | |
|---|---|
| ADME/T | Absorption, distribution, metabolism, excretion, and toxicity |
| GA | Genetic algorithm |
| HPIV | Human parainfluenza virus |
| HN | Hemagglutinin neuraminidase |
| MDSim | Molecular dynamics simulation |
| MM/GBSA | Molecular mechanics/generalized born model and solvent accessibility |
| RdRp | RNA-dependent RNA polymerase |
| rGyr | Radius of gyration |
| RMSD | Root mean square deviation |
| RMSF | Root mean square fluctuation |
| SASA | Solvent accessibility surface area |
| ssRNA | Single-stranded ribonucleic acid |
| VSGB | Variable dielectric generalized Born solvation method |

## Introduction

The most frequently occurring diseases among people of all age groups worldwide are acute respiratory illnesses. Children and older adults, especially those in developing countries, are the more vulnerable to such infections. HPIVs or human parainfluenza viruses represent the second most common cause after respiratory syncytial viruses, leading to acute respiratory tract infections in children [1]. There are four reported serotypes of HPIV: HPIV1 and HPIV3 belonging to *Respirovirus* and HPIV2 and HPIV4 of *Rubulavirus*. HPIV4 is less studied and is known to cause mild respiratory illnesses. HPIV1 and HPIV2 are associated with croup-related indications. HPIV3 is the most dangerous among the different serotypes of HPIVs and is responsible for indications ranging

✉ Yasha Hasija
yashahasija@dtu.ac.in

Sidharth Bhasin
sidharthbhasin@hotmail.com

[1] Department of Biotechnology, Delhi Technological University, Shahbad Daulatpur, Main Bawana Rd, New Delhi 110042, India

Published online: 16 September 2022

from bronchitis to pneumonia. Some atypical clinical indications include apnea, bradycardia, and respiratory distress syndrome [2]. HPIV3 infection has been reported to be one of the major causes of hospitalization, with 23,000 cases annually in the USA alone [3]. The financial burden of medical care expenses for HPIV-related bronchiolitis, croup, and pneumonia hospitalizations has been estimated to be approx. USD 43 M, USD 58 M, and USD 158 M, respectively [4]. Despite the clinical importance of HPIVs, there are currently no medications or vaccines available to cure or prevent the viral infection.

HPIVs are pleiomorphic (−)ssRNA viruses belonging to the *Paramyoviridaie* family. There are six main proteins involved in viral entry and replication. The matrix protein (M) is present on the interior side of the viral envelope, while the nucleocapsid protein (NP) covers the viral RNA. The fusion glycoprotein (F) and the hemagglutinin-neuraminidase (HN) glycoprotein present on the viral surface play a principal role in facilitating the cellular entry of virus [2, 5]. HN assists in the attachment of HPIV to sialic acid residues present on the host epithelial cells, while the F-protein causes the viral envelope to fuse with the host cell membrane to facilitate the transfer of viral (-)ssRNA. Once inside the cell, the NP cover is removed, and the uncovered RNA acts as the template for RNA-dependent RNA polymerase (P + L protein). The M protein helps in the assembly of the virus [6]. Additional proteins, unrelated to viral replication, are also reported for different HPIV serotypes: C proteins in HPIV1 and HPIV3; V protein in HPIV2; and D protein in HPIV3. These proteins, with the exception of D protein (unknown function), assist the viral progression by immunosuppression via a reduction in levels of type I IFN. After viral replication and packaging inside the cytoplasm, HN facilitates the cleavage of sialic acid residues to facilitate the exit of new HPIV molecules [2].

An effective strategy against parainfluenza can be targeting and inhibiting the cell surface protein HN to prevent the cellular entry of viruses. HN protein is a type II transmembrane protein that plays a pivotal role in viral entry and exit from host cells via receptor binding and cleavage. Furthermore, it is crucial to maintain the F protein's stability before receptor binding and activates it when attached to the host cell receptor [7, 8]. This study focuses on discovering the immunomodulatory phytocompounds as potential inhibitors of HPIV3-HN. This strategy will serve dual purposes of immunomodulation as well as viral inhibition. The chosen compounds are reported to have anti-cancer and anti-inflammatory effects. Furthermore, some compounds have also been reported to boost the immune system response against viral infections. Thus, this strategy can be more effective than administering a single antiviral therapy regimen. Furthermore, the natural compounds are highly effective with minimum side effects.

## Methods

### Protein preparation

The 3D structure of HPIV3 hemagglutinin-neuraminidase was downloaded from RCSB PDB with accession code 1V2I (Fig. 1) [9]. The macromolecule is a homo-2-mer consisting of 431 residues each. The unrelated molecules like water and ions were removed using DS Visualizer 2021. The structure was then saved in.pdb format for further studies. Due to the absence of a bound ligand in the original PDB structure, the binding site was ascertained by a combination of computational analysis and literature review. CASTp 3.0 was utilized to obtain all the binding pocket residues [10]. The key residues of the catalytic site were identified as R192, D216, E409, R424, R502, Y530, and E549 [11].

### Ligand library preparation

PubTator, a literature-based data mining tool from NCBI, was utilized to search for immunomodulatory phytocompounds [12]. A total of 45 compounds were obtained via manual curation and confirmation from retrieved literature. An overview of the chemical properties of these compounds is available in Table S1 (Supporting Information). The 3D structures of the listed phytochemicals were then retrieved in.sdf format from the PubChem databank [13]. Open Babel GUI v3.1.1 was then used to convert these structures into. pdb format for docking studies [14].

### Molecular docking studies

AutoDock 4.2.6 in combination with AutoDockTools GUI, was used to find the binding characteristics of screened

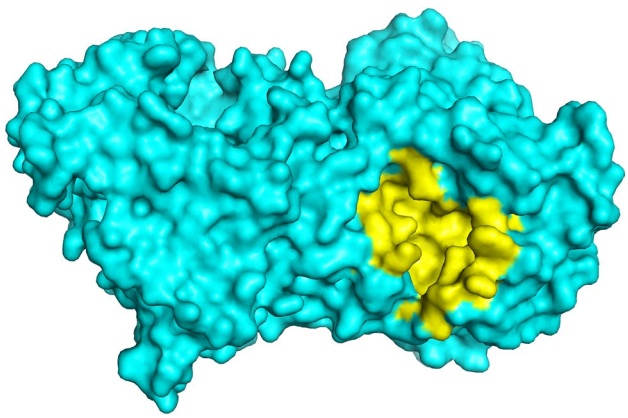


**Fig. 1** Protein surface of HPIV3-HN (PDB ID: 1V2I). The binding pocket is marked in yellow. Figure generated via the open-source version of PyMOL

ligands in the binding pocket of the protein [15]. The protocol was similar to the one recommended by Rizvi et al. (2013) [16]. The protein and ligand molecule were loaded in the environment and optimized by the addition of hydrogens and appropriate charges. A grid box of 50 × 68 × 52 was constructed around the active binding site, and AutoGrid was used for the creation of affinity maps. Lamarckian Genetic Algorithm (Lamarckian GA), a hybrid between GA and local search algorithm, was utilized for docking of ligands. The number of GA runs was set as 10 with population size as 150 and 27,000 generations; the rest of the parameters were set as software default. The docking studies yielded ten variable conformations of ligands in the binding pocket. The pose with highest negative binding energy per compound was selected for further analysis. The top hits were selected for interaction analysis using DS Visualizer 2020.

### Prediction of physicochemical parameters

The Canonical SMILES of the top compounds were downloaded from PubChem via PUG REST API and fed to the SwissADME web server available at http://swissadme.ch/ [17, 18]. SwissADME is a freely accessible web-based tool used to predict the physicochemical parameters of potential drug compounds. The parameters like molecular weight (MW), number of H-bond donors (H-Donor), number of H-bond acceptors (H-Acceptor), molar refractivity (MR), Moriguchi's calculated LogP values (MLogP), Wildman and Crippen's LogP values (WLogP), number of atoms (atoms), topological polar surface area (TPSA), and GI absorption were noted. The violations of both Lipinski's RO5 parameters viz. H-Donor ≤ 5, H-Acceptor ≤ 10, MW ≤ 500 Da, and MlogP ≤ 4.15, as well as Ghose's drug likeliness parameters, viz. MR ϵ [40, 130], atoms ϵ [20, 70], MW ϵ [160, 480] and WLOGP ϵ [0.4, 5.6], were determined [19, 20]. These rules predict a high probability of an oral drug being effective if the compound has fewer violations of the said parameters.

### MDSim studies

Molecular dynamics simulations (MDSim) were conducted on the top hits with the help of the Desmond module of Schrodinger Suite 2021–1 [21, 22]. The docked complex was loaded in the Maestro environment. The structure was pre-processed to detect any structural abnormalities, followed by assignment and optimization of hydrogen bonds at physiological pH 7.0, and default software settings were utilized. TIP3P water model and OPLS_2005 force field were used for building the system for MDSim. The buffer distance was set as 20 Å, while the cutoff distance for non-bonded interactions was configured as 4 Å. The charges were neutralized with the help of $Na^+$ and $Cl^-$ ions. The system was immersed in 0.15 M NaCl salt, and MDSim was run for 100 ns at 26.85 °C (= 300 K) and 1.01 bar. The parameters like PL-RMSD, P-RMSF, SASA, rGYR, and PL interactions were determined with the help of MD output. Afterwards, the script trj_essential_dynamics.py of Schrodinger Suite was used to run principal components analysis (PCA) on the MD trajectory to help in the examination of the collective motion of the backbone atoms of protein to compare the conformational variability of the HPIV3-HN protein when bound to different ligands.

### MM/GBSA analysis

The MM/GBSA analysis was performed for the last 10 ns of MD simulation using the Prime module of Schrodinger Suite 2021–1 [23–25]. This helps in the determination of free binding energies of the protein–ligand complexes [26]. The total binding energy represented by $\Delta G_{bind}$ is determined by the summation of the contribution of different energies like coulomb, covalent, electrostatic solvation, vdW, lipophilic, and H-bond interactions with corrections for self-contacts and π-π packing. The thermal_mmgbsa.py script of Schrodinger Suite 2021–1 was utilized to calculate individual MM/GBSA energies of each frame obtained from MD simulation starting from the 900th frame (representing 90 ns) with a step size of 100 ps, using the VSGB2.0 model and OPLS_AA forcefield [27].

## Results and discussion

### Molecular docking

The 45 immunomodulatory compounds of the phytocompound ligand library were docked in the binding site of HPIV3 HN (PDB ID: 1V2I) using AutoDock (Table 1). The compound BCX-2798, a reported inhibitor of HPIV3-HN obtained via in vitro and in vivo studies on HPIV3 infected mice, was taken as the reference compound [28–30]. The threshold function was the combination of binding energy (BE = − 6.18 kcal $mol^{-1}$) and inhibition constant (IC = 29.47 μM), corresponding to the values obtained for the reference compound. The compounds with more negative BE and lower IC than the reference have a higher efficacy probability than the inhibitor reported in the literature. A total of thirteen compounds satisfied the threshold function. The top five hits were selected for further analysis.

The least binding energy obtained for the reference BCX-2798 (CID:10088846) was − 6.18 kcal $mol^{-1}$. It forms a total of eight conventional H-bonds with five of the residues of the binding pocket, namely ARG192, ARG424, ARG502, TYR530, and GLU549 (Fig. 2a). It forms alkyl interactions with CYS214 and LEU558 and interacts with the rest of the pocket atoms via van der Waals forces. An unfavorable

**Table 1** Molecular docking of reference and experiment compounds

| Comp[a] | CID[b] | Name | BE[c] | LE[d] | IC[e] | IE[f] | VD[g] |
|---|---|---|---|---|---|---|---|
| - | 10088846 | BCX-2798 [h] | −6.18 | −0.26 | 29.47 | −10.06 | −7.74 |
| 1 | 72276 | (-)-epicatechin | −7.44 | −0.35 | 3.5 | −9.23 | −8.51 |
| 2 | 107905 | (-)-epicatechin gallate | −7.3 | −0.23 | 4.48 | −10.58 | −9.88 |
| 3 | 65064 | (-)-epigallocatechin gallate | −7.17 | −0.22 | 5.54 | −10.75 | −9.88 |
| 4 | 5,281605 | Baicalein | −7.16 | −0.36 | 5.68 | −8.35 | −8.12 |
| 5 | 44558901 | Salpichrolide C | −6.78 | −0.21 | 10.66 | −8.27 | −8.05 |
| 6 | 445154 | Resveratrol | −6.61 | −0.39 | 14.19 | −8.11 | −7.75 |
| 7 | 638024 | Piperine | −6.55 | −0.31 | 15.85 | −7.44 | −7.39 |
| 8 | 145858 | Flavylium | −6.52 | −0.41 | 16.56 | −6.82 | −6.69 |
| 9 | 12303662 | Phytosterol | −6.43 | −0.21 | 19.34 | −8.52 | −8.56 |
| 10 | 6264 | 1-fluoro-2,4-dinitrobenzene | −6.36 | −0.49 | 21.85 | −6.95 | −4.68 |
| 11 | 222284 | β-sitosterol | −6.35 | −0.21 | 22.18 | −8.44 | −8.41 |
| 12 | 9064 | Cianidanol | −6.35 | −0.3 | 21.99 | −8.14 | −7.52 |
| 13 | 5281437 | Costunolide | −6.33 | −0.37 | 22.89 | −6.33 | −6.03 |
| 14 | 73174 | Dehydrocostus lactone | −6.1 | −0.36 | 33.57 | −6.1 | −5.98 |
| 15 | 5280961 | Genistein | −6.1 | −0.31 | 33.8 | −7.29 | −7.02 |
| 16 | 370 | Gallic acid | −6.06 | −0.51 | 36.21 | −7.55 | −4.95 |
| 17 | 5280863 | Kaempferol | −6.03 | −0.29 | 37.74 | −7.53 | −7.12 |
| 18 | 122738 | Procyanidin B2 | −5.85 | −0.14 | 51.46 | −9.73 | −9.42 |
| 19 | 5281616 | Galangin | −5.82 | −0.29 | 54.59 | −7.01 | −6.57 |
| 20 | 5280343 | Quercetin | −5.81 | −0.26 | 55.5 | −7.6 | −7.19 |
| 21 | 5318998 | Licochalcone A | −5.78 | −0.23 | 58.16 | −8.16 | −7.99 |
| 22 | 5282164 | Gamma-oryzanol | −5.75 | −0.13 | 61.37 | −8.73 | −8.84 |
| 23 | 54242917 | 2-benzoyloxycinnamaldehyde | −5.63 | −0.3 | 74.28 | −7.12 | −7.02 |
| 24 | 5318169 | 2-hydroxycinnamaldehyde | −5.57 | −0.51 | 82.05 | −6.47 | −6 |
| 25 | 637858 | Piperlongumine | −5.52 | −0.24 | 90.18 | −7.01 | −6.85 |
| 26 | 5281787 | Caffeic acid phenethyl ester | −5.5 | −0.26 | 93.53 | −7.88 | −7.81 |
| 27 | 689043 | Caffeic acid | −5.41 | −0.42 | 108.22 | −6.9 | −5.47 |
| 28 | 5281426 | Umbelliferone | −5.35 | −0.45 | 119.76 | −5.65 | −5.3 |
| 29 | 10404245 | Macelignan | −5.25 | −0.22 | 141.25 | −7.34 | −7.27 |
| 30 | 382514 | Polyozellin | −5.22 | −0.16 | 150.08 | −7.6 | −7.38 |
| 31 | 84298 | Asperuloside | −5.17 | −0.18 | 162.23 | −8.15 | −8 |
| 32 | 5280460 | Scopoletin | −4.54 | −0.32 | 473.87 | −5.13 | −4.95 |
| 33 | 2758 | Eucalyptol | −5.02 | −0.46 | 208.69 | −5.02 | −4.84 |
| 34 | 174174 | Atropine | −4.73 | −0.23 | 343.3 | −6.52 | −7.21 |
| 35 | 644019 | Cannabidiol | −4.59 | −0.2 | 433.92 | −6.97 | −6.57 |
| 36 | 3000322 | Hyoscine | −4.58 | −0.21 | 435.71 | −6.37 | −6.32 |
| 37 | 22311 | Limonene | −4.58 | −0.46 | 440.23 | −4.88 | −4.89 |
| 38 | 637511 | Cinnamaldehyde | −4.44 | −0.44 | 560.22 | −5.03 | −4.9 |
| 39 | 445070 | Farnesol | −4.35 | −0.27 | 643.01 | −6.74 | −6.72 |
| 40 | 11223782 | Phyltetralin | −4.35 | −0.14 | 644.8 | −7.04 | −6.86 |
| 41 | 119 | γ-aminobutyric acid | −4.3 | −0.61 | 705.01 | −5.79 | −3.04 |
| 42 | 10621 | Hesperidin | −4.21 | −0.1 | 814.16 | −8.69 | −8.2 |
| 43 | 71312551 | Withanoside IV | −3.74 | −0.07 | 1.81 | −8.81 | −8.78 |
| 44 | 3014538 | 3-isopropenyl-6-oxoheptanal | −3.75 | −0.31 | 1.78 | −5.54 | −5.39 |
| 45 | 449303 | 4-phospho-D-erythronohydroxamic acid | −2.08 | −0.15 | 30.12 | −5.06 | −4.24 |

[a]Compound number (**1**, **2**, **3**…)

[b]PubChem compound ID

[c]Binding energy (kcal $mol^{-1}$)

[d]Ligand efficiency

[e]Inhibition constant (μM)

[f]Intermolecular energy (kcal $mol^{-1}$)

[g]Van der Waals H-bond desolvation (kcal $mol^{-1}$)

[h]BCX-2798 (CID:10088846) is the reference compound

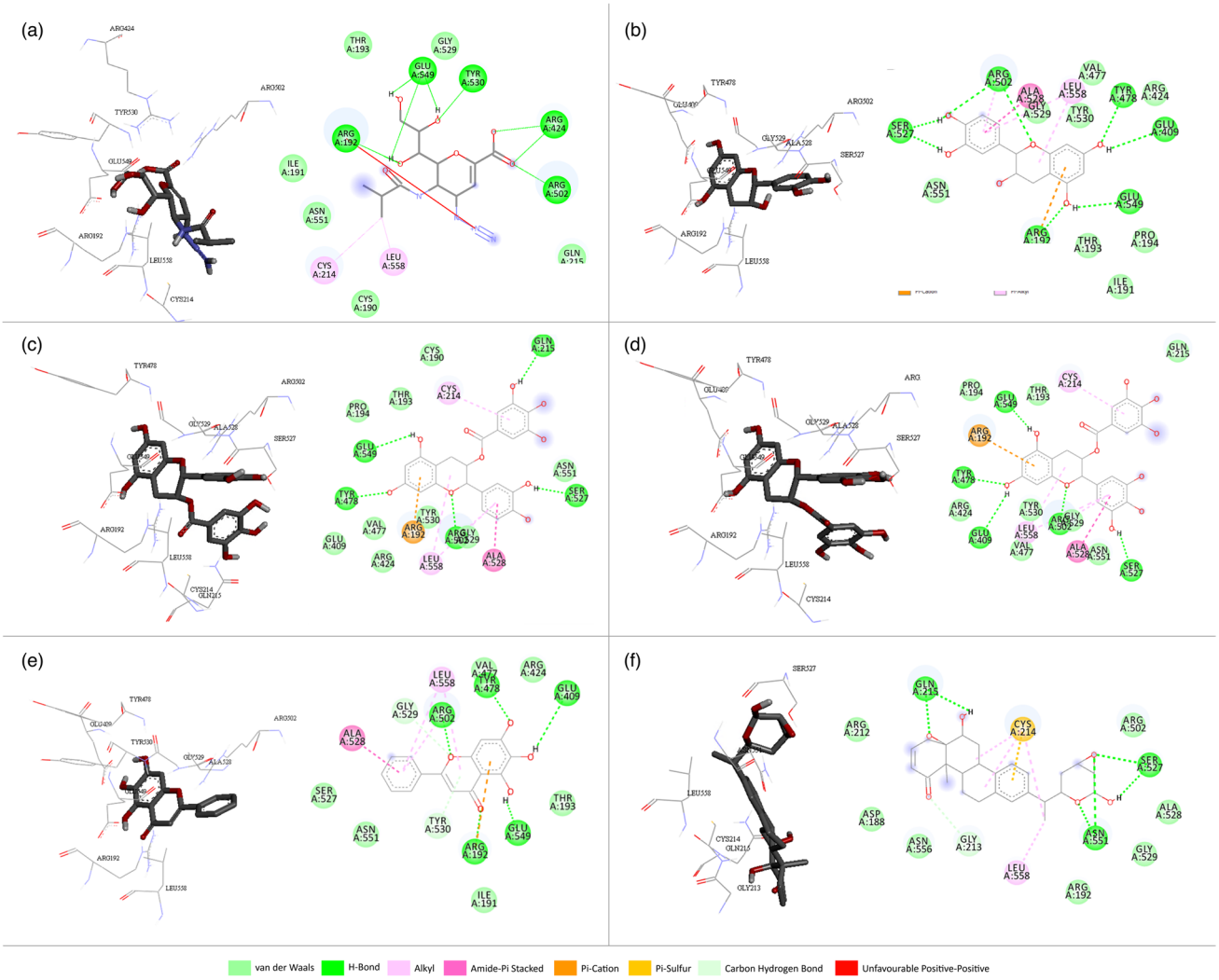



**Fig. 2** Interactions of different ligands with HPIV3 HN. (**a**) Reference compound BCX-2798 (CID:10088846). (**b**) Compound **1** (CID:72276). (**c**) Compound **2** (CID:107905). (**d**) Compound **3** (CID:65064). (**e**) Compound **4** (CID:5281605). (**f**) Compound **5** (CID:44558901). Left: best binding pose in 3D. Right: 2D interaction diagram

interaction is also observed in the form of a positive-positive clash between the ligand and ARG192.

Out of ten conformations of compound **1** (CID:72276; (-)-epicatechin), the least binding energy obtained was − 7.37 kcal $mol^{-1}$. The ligand formed eight conventional H-bonds with ARG192, GLU409, TYR478, ARG502, SER527, and GLU549 (Fig. 2b). It also formed an amide pi-stacked bond with ALA528 and a pi-alkyl bond with LEU558. Van der Waals interactions were observed with the remaining residues of the binding pocket. Compound **2** (CID:107905; (-)-epicatechin gallate) forms five conventional H-bonds with GLN215, TYR478, ARG502, SER527, and GLU549 (Fig. 2c). The ligand forms one pi-amide stacked bond with ALA528; one pi-cation interaction with ARG192; and two alkyl bonds with CYS214 and LEU558. Van der Waals interactions were observed with the remaining residues of the binding pocket. Compound **3** (CID:65064; (-)-epigallocatechin gallate) formed a total of five conventional H-bonds with GLU409, TYR478, ARG502, SER527, and GLU549 (Fig. 2d). The ligand interacts with ARG192 via amide-pi stacked interactions, CYS214 and LEU558 via alkyl interactions, and ALA528 via amide π-stacked interactions. The ligand interacts via Van der Waals interactions with the remaining residues of the binding pocket.

A sum total of seven types of interactions were observed between compound **4** (CID:5281605; baicalein) and protein (Fig. 2e). It formed conventional H-bonds with ARG502, TYR478, GLU409, GLU549, and ARG192. C-H bonds and π-donor hydrogen bonds were observed between the ligand and residues GLY529 and TYR530. The ligand also exhibited π-cation and amide-π stacked interactions with

ARG192, and ALA528, respectively. π-alkyl interactions were also seen between the ligand and residues ARG502 and LEU558. Compound **5** (CID:44558901; salpichrolide C) formed five conventional H-bonds with three residues of chain A of the protein: GLN215, SER527, and ASN551 (Fig. 2f). It also exhibited π-sulfur and alkyl interactions with CYS214 and LEU558, respectively. Van der Waals interactions were observed with the remaining residues of the binding pocket.

Compound **4** has the highest types of interactions, and the highest number of conventional H-bonds (= 8) were observed in compound **1**. The screened compounds have good contact with the binding pocket residues of HPIV3-HN protein as compared to the reference compound BCX-2798 (CID:10088846). The interactions of different compounds with various protein residues are summarized in Table S2 (Supporting Information).

## Prediction of physicochemical properties

The physicochemical parameters of top hits were predicted using the SwissADME server (Table 2). The parameters of two filters, viz. Lipinski RO5 and Ghose Rule, were considered. These rules warrant fewer anomalies of respective parameters for compounds to be potential drug molecules. The violations of Lipinski's parameters were in the range of 0 to 2 (the generally considered upper limit of violations is four), while no parameters of Ghose's rule were violated by any compound. These indicate that there is a high probability that the selected compounds have high drug-likeliness properties. The SwissADME server predicted the GI absorption for all compounds except CID:107905 (compound **2**) and CID:65064 (compound **3**) to be high, indicating that there may be some hindrance in the absorption of said compounds. This observation is also supported by the TPSA values of the compounds; generally, a value less than 140 $Å^2$ means higher chances of intestinal absorption [31, 32]. These compounds will have to be conjugated/encapsulated in a suitable drug delivery vehicle to increase the absorption and half-life of these compounds. A suitable linker and/or nanoparticle strategy can be devised for these compounds to function at their maximum efficiency, albeit their efficiency of action will have to be assessed. Overall, the identified leads follow both Lipinski's RO5 and Ghose rule, indicating that these compounds should exhibit excellent ADME properties and will make a suitable drug candidate for further testing.

## MDSim analysis

Molecular dynamics simulation (MDSim) studies were conducted on the top five hits from docking studies to determine the binding characteristics of ligand in the binding

**Table 2** Physicochemical properties of screened phytocompounds

| C. no. | CID | MW (g $mol^{-1}$) | H-Donor | H-Acceptor | MLogP | Lipinski[a] violation | MR ($m^3$ $mol^{-1}$) | WLogP | Atoms | Ghose[b] violations | GI absorption | TPSA ($Å^2$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 72276 | 290.27 | 5 | 6 | 0.24 | 0 | 74.33 | 1.22 | 33 | 0 | High | 110.38 |
| 2 | 107905 | 442.37 | 7 | 10 | 0.05 | 1 | 110.04 | 2.2 | 50 | 0 | Low | 177.14 |
| 3 | 65064 | 458.37 | 8 | 11 | −0.44 | 2 | 112.06 | 1.91 | 51 | 0 | Low | 197.37 |
| 4 | 5281605 | 270.24 | 3 | 5 | 0.52 | 0 | 73.99 | 2.58 | 36 | 0 | High | 90.9 |
| 5 | 44558901 | 440.53 | 3 | 6 | 1.86 | 0 | 117.65 | 2.34 | 38 | 0 | High | 99.52 |

[a]Lipinski parameters: H-Donor ≤ 5, H-Acceptor ≤ 10, MW ≤ 500 Da and MlogP ≤ 4.15

[b]Ghose parameters: MR ∈ [40, 130], atoms ∈ [20, 70], MW ∈ [160, 480], and WLOGP ∈ [0.4, 5.6]

pocket of the protein in the dynamic state. The potential and total energy parameters were calculated during the simulation time of 100 ns (Table 3). The reference compound exhibited the least potential energy and total energy (more negative) values of − 258,803.891 kcal $mol^{-1}$ and − 209,242.817 kcal $mol^{-1}$, respectively. Among the screened compounds, compound CID:65064 (**4**) had appreciable value w.r.t reference viz. − 258,452.161 kcal $mol^{-1}$ and − 208,884.881 kcal $mol^{-1}$, respectively. A lower value of energy parameters (more negative) generally indicates a higher binding affinity of the ligand with the binding site throughout the simulation time. Thus, the compounds have an appreciable binding affinity with the HN protein compared to the reference BCX-2798. Compound CID:44558901 (**5**) exhibited the least negative energy parameters: potential energy of − 258,254.222 kcal $mol^{-1}$ and total energy of − 208,699.129 kcal $mol^{-1}$.

## Structural deviation and compactness

The different parameters to determine the stability of protein structure upon binding of ligand and the compactness of protein–ligand binding are PL-RMSD, L-rGyr, and SASA. The raw data generated for these parameters is available in Dataset S1 (Supporting Information). RMSD or root mean square deviation plots are used to measure the average displacement of atoms from an initial reference point (Fig. 3). Protein RMSD (P-RMSD) gives information about the stability of a protein's structural conformation. Changes in the ballpark of 1–3 Å for P-RMSD variations are acceptable for Desmond MDSim. The P-RMSD plot for the 1V2I-CID10088846 (BCX-2798; reference compound) complex exhibited stability in the form of a plateau region during the simulation. The slight changes observed in the ligand RMSD having an increasing trend indicate slight alterations in the binding pose of the reference compound towards the most optimal pose. 1V2I-CID:5281605 (**4**) complex exhibited the best stability in the duration of 50–100 ns, followed by CID:65064 (**3**) and CID:72276 (**1**), depicted by the formation of a plateau region. The RMSD plot for the 1V2I-CID:65064 (**3**) complex continuously increased across the simulation time of 100 ns. The ligand and protein RMSD plot were in complete sync in complex with CID:107905 (**2**). The crests and troughs observed in RMSD plots are due to the internal vibrations of the complex. The P-RMSD plot for CID:44558901 (**5**) exhibited the least fluctuations towards the latter half of the simulation, but the drastic transition points indicated in the L-RMSD plot suggest low affinity of ligand in the binding pocket. All the other complexes exhibit stable RMSD values towards the end of the simulation, indicating the maintenance of stability and affinity of the ligand with the active site of HPIV3-HN protein throughout the simulation time.

The radius of gyration of ligands estimates the extendedness of the ligand. The plot should remain constant for the full simulation timeline (Fig. 4a). Additionally, a lower rGyr value indicates close packing of the PL complex. The reference compound BCX-2798 (CID:10088846) exhibited a stable rGyr value of 3.7 ± 0.04 Å. The lowest rGyr value was obtained for the compound CID:5,281,605 (**4**) (3.57 ± 0.03 Å), followed by CID:72276 (**1**) (3.75 ± 0.03 Å), CID:65064 (**3**) (4.23 ± 0.08 Å), and CID:107905 (**2**) (4.37 ± 0.06 Å). A few deviations in the form of troughs are seen in the case of CID:44558901 (**5**). These correspond to a similar change in the L-RMSD plot for the compound and may be due to changes in protein–ligand packing.

SASA or solvent accessible surface area measures the accessibility of the surface area to solvent and should remain stable (Fig. 4b). The most stable SASA values are seen for CID:72276 (**1**) (145.01 ± 20.30 $Å^2$), followed by CID:107905 (**2**) (180.50 ± 20.46 $Å^2$) and CID:65064 (**3**) (162.78 ± 20.79 $Å^2$). BCX-2798 (CID:10088846) (Reference) exhibited a SASA value of 150.19 ± 48.56 $Å^2$. The values rise for an initial period before stabilizing at a lower value towards the later part of the simulation, corresponding to a slight change in the ligand binding mode to gain the optimum configuration. The SASA values for CID:72276 (**1**) and CID:107905 (**2**) overlap towards the end of the simulation. The SASA values for CID:5281605 (**4**) exhibited dual regions of stability, the transition point to a higher SASA value corresponding to a similar transition in the L-RMSD. This may be due to slight alterations in contact between the ligand and protein, resulting in an increase in SASA values. The SASA values for CID:44558901 (**5**) depicted two major

**Table 3** Energy parameters for 1V2I-ligand complexes obtained by MDSim

| Compound | CID | Compound name | Potential energy (kcal $mol^{-1}$) | Total energy (kcal $mol^{-1}$) |
|---|---|---|---|---|
| Reference | 10088846 | BCX-2798 | − 258,803.891 ± 136.459 | − 209,242.817 ± 150.631 |
| 1 | 72276 | (-)-epicatechin | − 258,417.380 ± 144.911 | − 208,856.111 ± 157.102 |
| 2 | 107905 | (-)-epicatechin gallate | − 258,404.776 ± 145.745 | − 208,838.671 ± 157.742 |
| 3 | 65064 | (-)-epigallocatechin gallate | − 258,430.297 ± 148.251 | − 208,861.102 ± 160.306 |
| 4 | 5281605 | Baicalein | − 258,452.161 ± 141.930 | − 208,884.881 ± 155.156 |
| 5 | 44558901 | Salpichrolide C | − 258,254.222 ± 158.708 | − 208,699.129 ± 169.615 |

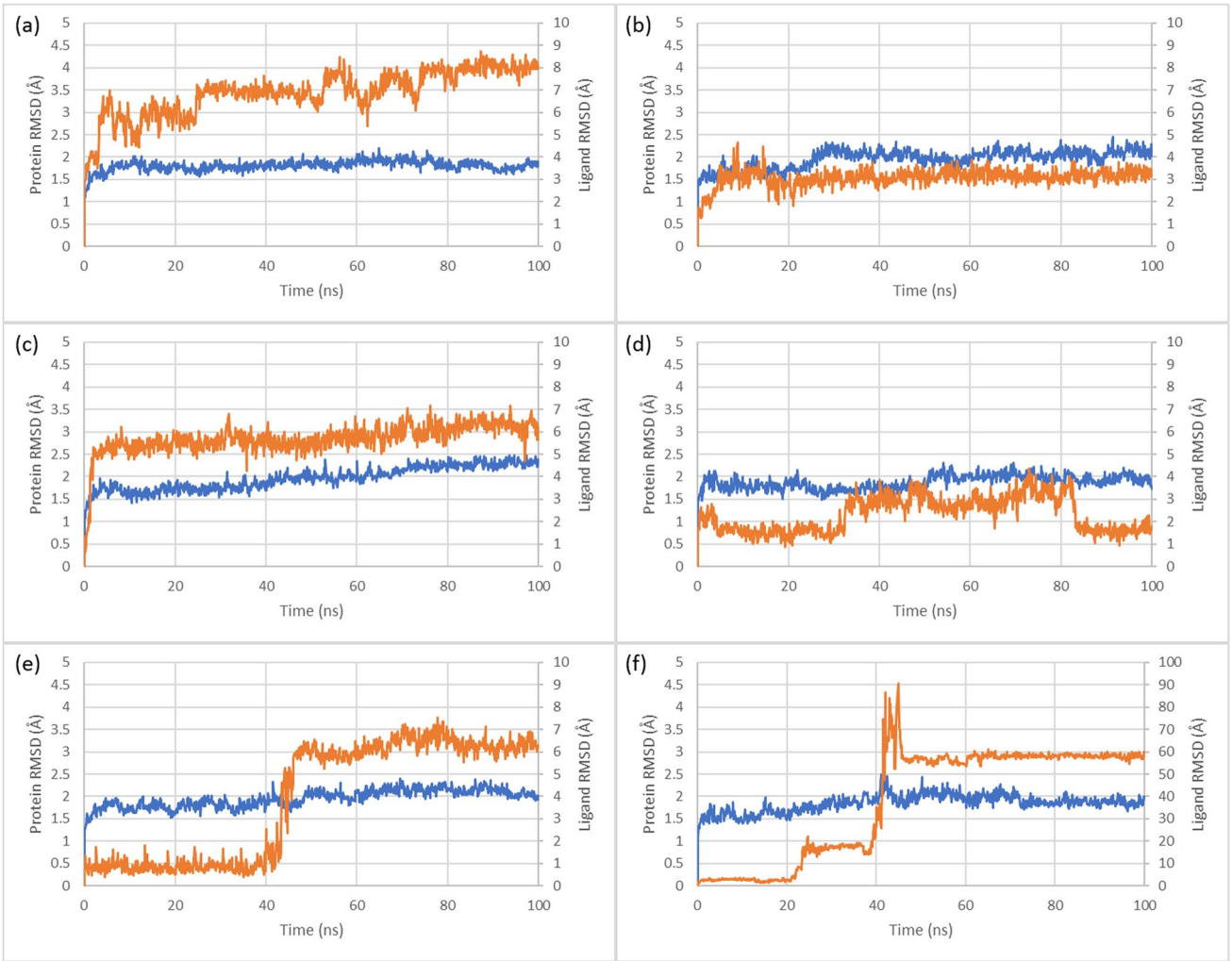



**Fig. 3** RMSD plots for protein–ligand complexes. (**a**) 1V2I-CID:10088846 (reference) complex. (**b**) 1V2I-CID:72276 complex. (**c**) IV2I-CID:107905 complex. (**d**) 1V2I-CID:65064 complex. (**e**) 1V2I-CID:5281605 complex. (**f**) 1V2I-CID:44558901 complex. P-RMSD is depicted by blue while L-RMSD is depicted in orange

fluctuations before stabilizing towards the middle and end of the stimulation. These fluctuations correspond to the major peaks and transition points observed in the L-RMSD plot for the compound, indicating an influx of solvent molecules into the binding pocket and interaction with inner residues at these instances. The SASA values for all the compounds equilibrized towards the end of the simulation, indicating the structural stability of complexes.

## Protein structural integrity

The number of residues forming the secondary structural elements of a protein is a measure of protein structural stability. The higher the SSE participation, the higher the protein–ligand complex's stability. The reference compound CID:10088846 (BCX-2798) witnessed the highest percentage of amino acids participating in secondary structure formation. Similar SSE participation was observed in CID:65064 (**3**) and CID:72276 (**1**), followed by CID:107905 (**2**) and CID:5281605 (**4**) (see Table 4). Furthermore, the percentage of SSE formed should remain stable and not exhibit high deviations throughout the simulation. The plots for SSE participation remained steady for all the compounds (Figure S1).

RMSF plots are a measure of the structural integrity of the residues (Figure S2). The residues participating in the formation of secondary structural elements (like β-strand and α-helix) exhibit low deviations due to the rigidity of these structures. The residues of the binding pocket exhibit lower fluctuations, indicating a smaller number of conformational changes in the binding pocket. Furthermore, the RMSF plots were running more or less parallel to the B-factor for all the compounds, indicating protein stability in the complexes. These observations for all the five compounds

**Fig. 4** Structure deviation and compactness plots. (**a**) rGyr plots for ligands. (**b**) SASA plots for 1V2I-ligand complexes

**Table 4** Percentage of residues participating in secondary structural element formation

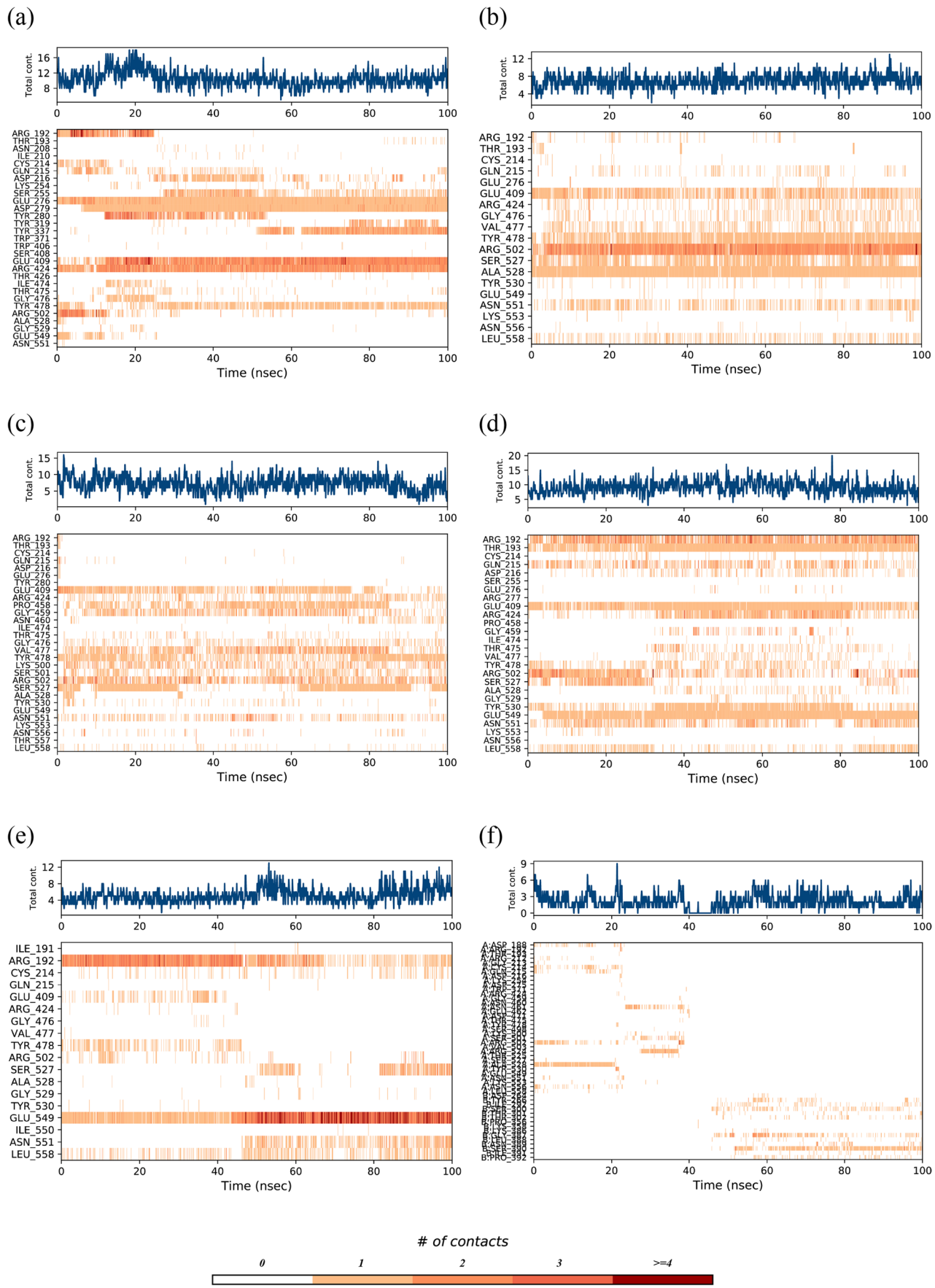
| Compound | CID | Compound name | % total SSE | % α-helix | % β-strand |
|---|---|---|---|---|---|
| (Reference) | 10088846 | BCX-2798 | 38.20 | 1.55 | 36.65 |
| 1 | 72276 | (-)-epicatechin | 38.08 | 1.52 | 36.56 |
| 2 | 107905 | (-)-epicatechin gallate | 37.55 | 1.30 | 36.25 |
| 3 | 65064 | (-)-epigallocatechin gallate | 38.18 | 2.43 | 35.75 |
| 4 | 5281605 | Baicalein | 37.25 | 1.36 | 35.89 |
| 5 | 44558901 | Salpichrolide C | 37.47 | 1.37 | 36.10 |

and the reference indicate good structural integrity of protein when bound to different ligands.

## MD protein–ligand interaction analysis

The main aim of MD simulations was to find out whether the interactions between the ligand and proteins remain throughout the simulation so as to confirm the presence of ligand in the binding site without any shred of doubt. For the reference compound BCX-2798 (CID:10088846), there is an incidence of weak contacts during the 100-ns simulation (Fig. 5). The reference makes strong contacts with residues GLU409 and ARG424 in the majority of the simulation time and with ARG192 in the beginning of the simulation. An analysis of the PL interaction timeline for the test compounds revealed CID:72276 (**1**), CID:107905 (**2**), and CID:5281605 (**4**) to have only one incidence devoid of strong contacts between the ligand and the protein. For CID:5281605 (**4**), strong interactions alternate between two residues: ARG192 and GLU549, with potent interactions observed towards the latter half of the simulation. CID:65064 (**3**) exhibits two incidences of no contacts but has the highest number of contact formations among the tested compounds. The number of contacts observed any time between 1V2I and the ligand averaged around 4–8 for CID:72,276 (**1**), CID:107,905 (**2**), and CID:5281605 (**4**), and 5–10 for CID:65064 (**3**). No appreciable PL contacts have been observed for CID:44558901 (**5**), indicating a low affinity of the compound with the active site of HPIV3-HN, and thus, the compound was ruled out for further analysis.

The PL histogram (Fig. 6) gives a better picture of the residues involved in interactions, revealing that for CID:107905 (**2**), none of the residues interacting with the ligand have an interaction fraction greater than one. CID:72276 (**1**),


(a)
Total cont.
ARG_192
THR_193
ASN_208
ILE_210
CYS_214
GLN_215
ASP_216
LYS_254
SER_255
GLU_276
ASP_279
TYR_280
TYR_319
TYR_337
TRP_371
TRP_406
SER_408
GLU_409
ARG_424
THR_426
ILE_474
THR_475
GLY_476
TYR_478
ARG_502
ALA_528
GLY_529
GLU_549
ASN_551
Time (nsec)
(b)
Total cont.
ARG_192
THR_193
CYS_214
GLN_215
GLU_276
GLU_409
ARG_424
GLY_476
VAL_477
TYR_478
ARG_502
SER_527
ALA_528
TYR_530
GLU_549
ASN_551
LYS_553
ASN_556
LEU_558
Time (nsec)
(c)
Total cont.
ARG_192
THR_193
CYS_214
GLN_215
ASP_216
GLU_276
TYR_280
GLU_409
ARG_424
PRO_458
GLY_459
ASN_460
ILE_474
THR_475
GLY_476
VAL_477
TYR_478
LYS_500
SER_501
ARG_502
SER_527
ALA_528
TYR_530
GLU_549
ASN_551
LYS_553
ASN_556
THR_557
LEU_558
Time (nsec)
(d)
Total cont.
ARG_192
THR_193
CYS_214
GLN_215
ASP_216
SER_255
GLU_276
ARG_277
GLU_409
ARG_424
PRO_458
GLY_459
ILE_474
THR_475
VAL_477
TYR_478
ARG_502
SER_527
ALA_528
GLY_529
TYR_530
GLU_549
ASN_551
LYS_553
ASN_556
LEU_558
Time (nsec)
(e)
Total cont.
ILE_191
ARG_192
CYS_214
GLN_215
GLU_409
ARG_424
GLY_476
VAL_477
TYR_478
ARG_502
SER_527
ALA_528
GLY_529
TYR_530
GLU_549
ILE_550
ASN_551
LEU_558
Time (nsec)
(f)
Total cont.
Time (nsec)
# of contacts
0
1
2
3
>=4

◂**Fig. 5** PL interaction timeline. (**a**) 1V2I-CID:10088846 (reference) complex. (**b**) 1V2I-CID:72276 complex. (**c**) IV2I-CID:107905 complex. (**d**) 1V2I-CID:65064 complex. (**e**) 1V2I-CID:5281605 complex. (**f**) 1V2I-CID:44558901 complex

CID:65064 (**3**), and CID:5281605 (**4**) remained in contact with the protein for the entire duration of 100 ns indicated by an interaction fraction greater than one. All four compounds exhibit some interactions with the key residues of the binding site throughout the simulation time.

The reference compound BCX2798 (CID:10088846) exhibits four significant H-bonds with residues GLU276, ASP279, GLU409, and ARG424 (interaction fraction > 0.75). An ionic bond was also observed between the reference and residues GLU409 and TYR337 for approximately 50% of the simulation time (interaction fraction = 0.5). The majority of the observed interactions with other residues are via the formation of water bridges between the reference and the protein, thus explaining the higher average number of contacts (8–12) observed as compared to the rest of the compounds.

An analysis of the 1V2I-CID:72276 complex revealed that the ligand forms H-bonds with TYR478, ARG502, SER527, ALA528, and ASN551. The interaction fraction for ARG502 was greater than 1.75, and interactions were observed employing water bridges and hydrophobic interactions apart from H-bonds. Upon analyzing the 1V2I-CID:5281605 complex, it was observed that the interaction fraction was more than 1.0 for residues ARG192 and GLU549, with significant interactions being H-bond and minor contributions observed in the form of water bridges. Transient contacts by means of water bridges and hydrophobic interactions are observed for the rest of the residues in the binding pocket.

An analysis of the 1V2I-CID:107905 complex revealed that this ligand forms the highest number of interactions with protein residues among the compounds under consideration. The ligand forms conventional H-bonds with protein residues, PRO458, GLY459, TYR478, ARG502, and SER527 for a significantly large amount of time than GLU409, ARG424, and ASN551. Interestingly, a transient ionic bond formation was observed with ARG424, ASN460, THR475, ARG502, and ASN551. The interactions among ligand and residues VAL477 and ARG502 were observed to be a combination of water bridges, H-bonding, and hydrophobic interactions.

In the case of the 1V2I-CID:65064 complex, it was seen that the ligand formed conventional H-bonds with THR193, GLU409, ARG502, and GLU549 for a significant amount of time, with THR193 and GLU549 for more than 85% of simulation time. The highest interaction fraction of 1.2 was observed for the ligand and residue ARG192, with a combination of interactions such as H-bond, water bridges, and hydrophobic interactions. Overall, the aforementioned observations confirm the presence of compounds **1**–**4** in the binding site of HPIV3-HN protein throughout the duration of 100 ns simulation by means of different interactions, viz. H-bonds, hydrophobic interactions, and water bridges (with transient ionic interactions observed in some compounds). Furthermore, these ligands interact with the key residues of the active site responsible for the entry of the virus into the cell, thus making these residues unavailable for the interactions required for viral entry.

## PCA analysis

The remaining four compounds were subjected to principal components analysis via essential dynamics approach to determine the conformational subspace occupied by the protein backbone when bound to different ligands. PCA helps in the assessment of the collective motion and conformational exploration of protein atoms and has been regarded as one of the essential tools to study the protein dynamics on the binding of small molecule inhibitors. The projections of protein C-α atoms along two pre-dominant eigenvectors were plotted against one another and with respect to the simulation time (Fig. 7). The protein exhibited similar, if not more, conformational restriction than the reference compound in all the cases except when bound to compound **3** (CID:65064), hinting towards less stability of this compound in the binding pocket with respect to the reference compound.

## MM/GBSA analysis

The MD simulation resulted in the generation of 1001 frames. The last 101 frames, corresponding to the last 10 ns of simulation, were processed and analyzed for the remaining three compounds. The most negative $\Delta G_{bind}$ value of -54.54 ± 4.19 kcal $mol^{-1}$ was observed for CID:107905 (**2**), while CID:5281605 (**4**) exhibited the least negative $\Delta G_{bind}$ of − 22.57 ± 1.86 kcal $mol^{-1}$. The reference, BCX-2798 (CID:10088846), exhibited a $\Delta G_{bind}$ value of − 27.97 ± 3.42 kcal $mol^{-1}$, in which the solvation component of $\Delta G_{bind}$ exhibited the most negative contribution while the Coulombic component exhibited the most positive contribution. Among all the test compounds, Coulomb interactions contribute the most towards the $\Delta G_{bind}$ value for Compound CID:72276 (**1**) vis-à-vis its total $\Delta G_{bind}$ value, followed by vdW (van der Waals) interactions. For compound CID:107905 (**2**), vdW interactions are more than double the contributions from Coulombic interactions. The total contributions of different energies in $\Delta G_{bind}$ value of different compounds are summarized in Table 5. Compounds **1** and **2** exhibited more negative binding energy than the reference BCX-2798 (CID:10088846). Compound **4** exhibited a considerably lower $\Delta G_{bind}$ value, in which vdW interactions

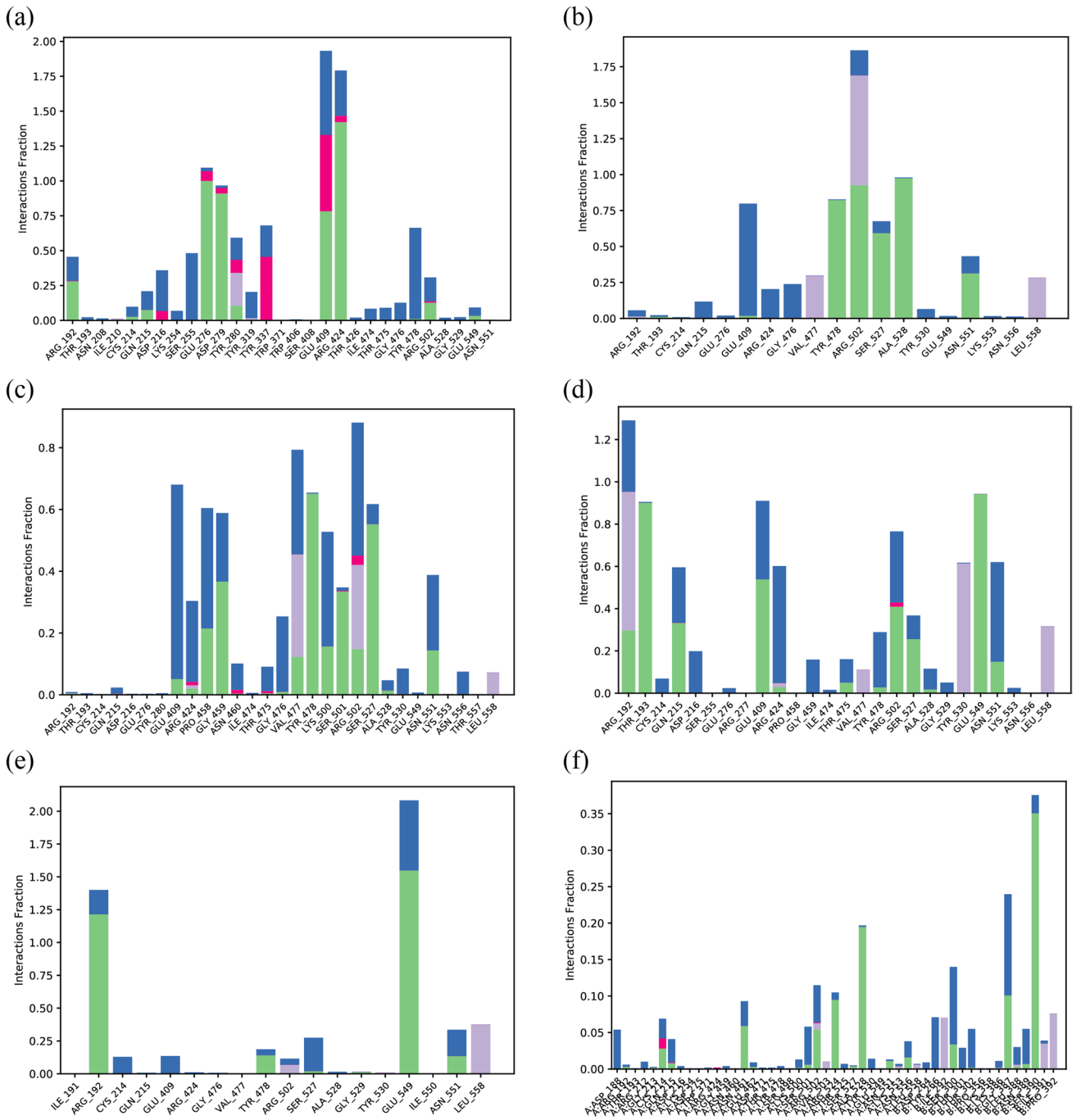



**Fig. 6** PL contact histogram. (**a**) 1V2I-CID:10088846 (reference) complex. (**b**) 1V2I-CID:72276 complex. (**c**) IV2I-CID:107905 complex. (**d**) 1V2I-CID:65064 complex. (**e**) 1V2I-CID:5281605 complex. (**f**) 1V2I-CID:44558901 complex (green depicts H-bond, purple depicts hydrophobic interactions, blue represents interactions via water bridges while pink demonstrates Ionic interactions)

contributed the most. Thus, the lead compounds emerging after MM/GBSA analysis were compounds CID:72276 (**1**), and CID:107905 (**2**).

All the lead compounds belong to the class of flavonoids and have been discovered in diverse plant species. Various studies have shown these compounds to possess antioxidant, antimicrobial, anti-inflammatory, antitumor, and cardioprotective activities [33, 34]. Flavonoids like orientin and vitexin have been proven to be inhibitory against HPIV3 [35]. Epicatechin has been shown to exhibit antimicrobial properties against *Moraxella catarrhalis*, *Streptococcus pneumoniae,* and *Haemophilus influenzae* [36]. Furthermore, these compounds have been proved as effective antiviral compounds in many in vivo and in vitro

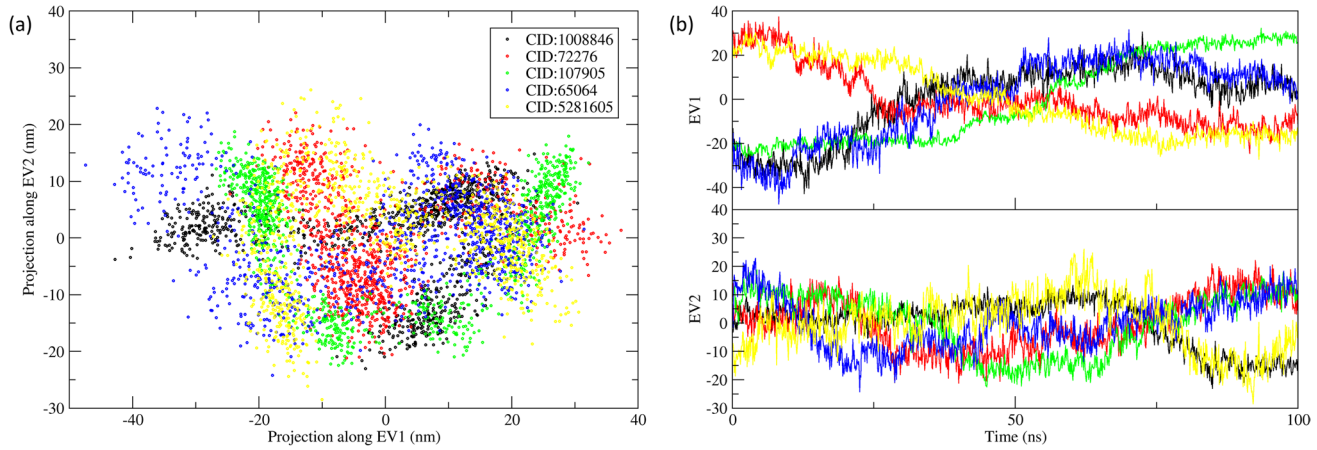



**Fig. 7** Principal component analysis. (**a**) 2D projection of HPIV3-HN protein complexed with CID:10088846 (black), CID:72276 (red), CID:107905 (green), CID:65064 (blue), and CID:5281605 (yellow) calculated after 100 ns of MD simulation. (**b**) Projections along the two modes plotted as a function of simulation time

**Table 5** Binding free energies ($\Delta G_{bind}$) (in kcal mol$^{-1}$) obtained via MM/GBSA analysis

| C. no. | CID | $\Delta G_{bind}$ | $\Delta G_{bind}$ (Coulomb) | $\Delta G_{bind}$ (H-bond) | $\Delta G_{bind}$ (lipophilicity) | $\Delta G_{bind}$ (vdW) | $\Delta G_{bind}$ (solvation) |
|---|---|---|---|---|---|---|---|
| (Reference) | 10088846 | −27.97 ± 3.42 | 94.99 ± 20.24 | 6.98 ± 1.83 | −3.60 ± 0.61 | −29.61 ± 2.70 | −90.54 ± 18.06 |
| **1** | 72276 | −43.13 ± 4.09 | −27.57 ± 4.71 | −2.56 ± 0.37 | −9.45 ± 0.62 | −24.46 ± 2.24 | 20.73 ± 2.31 |
| **2** | 107905 | −54.54 ± 4.19 | −21.23 ± 4.61 | −2.03 ± 0.5 | −15.18 ± 0.77 | −44.51 ± 2.11 | 28.13 ± 3.29 |
| **4** | 5281605 | −22.57 ± 1.86 | −14.40 ± 2.43 | −2.26 ± 0.31 | −3.83 ± 0.43 | −19.70 ± 1.57 | 17.89 ± 1.83 |

studies [37]. Catechins have been proven to exert inhibitory effects against the viruses HSV1, HPIV3, influenza A/B, and SV1 [38, 39]. Epicatechin and epicatechin gallate reduced the CPE exerted by HSV1 on Vero cells [40]. They have also been proven to be effective against HIV by preventing viral entry [39]. Even though the flavonoids have been proven to be inhibitory in numerous in silico and in vitro studies, a significant problem lies with the bioavailability of these compounds. Epicatechin has been shown to possess good ADME/T properties with an absorption rate of ~ 95% [41] and is thus a promising drug candidate. Structurally, all these compounds contain multiple hydroxyl groups, which play an essential role in H-bond formation with the inner key residues of the binding pocket. Similar docking orientation is observed in the case of both the compounds, with the flavin ring being placed inside the core cavity; with additional groups interacting with the neighboring pocket residues. The interaction is similar to various in vitro studies performed on neuraminidase inhibitors and is reported to prevent the proteolytic cleavage and preventing the sialidase activity of the influenza virus [42]. It was reported that the presence of C(4) = O and C(2) = C(3) bonds in these compounds is crucial in the neuraminidase-inhibiting activity of flavonoids [43]. Overall, the identified lead compounds may prove to be beneficial in inhibiting the progression of HPIV3 infection.

## Conclusions

HPIV3 hemagglutinin-neuraminidase macromolecule is a homo-2-mer consisting of 431 residues each. It is vital for viral entry into the cell and thus can be a target for suitable therapeutic intervention. In the current study, 45 compounds with immunomodulatory properties were selected from different plants. This library of immunomodulatory ligands was considered to impart the dual benefits of immunomodulation and inhibition of viral progression of HPIV3. This combinatorial therapy can be more effective than antiviral therapy alone. Furthermore, natural products in the form of plant extracts have been in use since ancient times for cutting various ailments and possess better ADME/T parameters with minimum known side effects. The compounds were docked in the binding site of HPIV3-HN protein using the Lamarckian Genetic Algorithm on AutoDock 4.2.6. The physicochemical properties were analyzed for the top hits of docking studies to determine their drug likeliness. The top five

hits were taken further for molecular dynamics simulation studies. No appreciable PL contacts were observed for CID:44558901 (**5**), and the compound was ruled out from further studies. CID:5281605 (**4**) exhibited protein–ligand contacts for the maximum duration. The protein exhibited the best SSE structure when bound to CID:65064 (**3**). The lowest SASA values were observed for CID:72276 (**1**). The lowest rGyr value was obtained for CID:5281605 (**4**). CID:107905 (**2**) demonstrated appreciable results in all the parameters of MDSim. Thus, the MD studies revealed CID:72276 (**1**), CID:107905 (**2**), CID:65064 (**3**), and CID:5281605 (**4**) to be potential compounds. PCA analysis predicted compound CID:65064 (**3**) to be less stable than the reference compound, and therefore, the compound was excluded from further analysis. The remaining three compounds were then taken forward for MM/GBSA analysis, which ruled out CID:5281605 (**4**) due to a significantly low $\Delta G_{\text{bind}}$ value compared to the reference compound BCX-2798 (a proven inhibitor for HPIV3-HN). The final lead compounds CID:72276 (**1**) ((-)-epicatechin) and CID:107905 (**2**) ((-)-epicatechin gallate) are computationally predicted to be effective against HPIV3 HN macromolecule. Further in vitro studies will be required to prove the efficacy of these compounds as inhibitors and to determine the exact mechanism of their inhibition.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s00894-022-05310-9.

**Acknowledgements** SB is thankful to Mr. Kishore Venkatesh and Ms. Shelvia Malik of Schrodinger Inc. for providing the temporary license of Schrodinger Suite 2021-1 for PCA and MM/GBSA Analysis. SB expresses his gratitude to Mr. Rajkumar Chakraborty, Ph.D. Scholar, Delhi Technological University, for his valuable input in the process of drafting the manuscript.

**Author contribution** Sidharth Bhasin: conceptualization, methodology, validation, writing — original draft preparation. Megh Nadar: data curation, investigation. Yasha Hasija: supervision, writing — review and editing.

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

**Software availability** All the tools and databases used in this study are open source and are freely accessible on the web unless otherwise specified. Discovery Studio Visualizer 2021 is proprietary software distributed for free by BIOVIA. Schrodinger Suite 2021–1 is proprietary fee-based software distributed by Schrodinger Inc. The Desmond module of Schrodinger Suite 2021–1 is distributed free of cost by D. E. Shaw Research under the Academic License.

## Declarations

**Conflict of interest** The authors declare no competing interests.

## References

1. Pawełczyk M, Kowalski ML (2017) The role of human parainfluenza virus infections in the immunopathology of the respiratory tract. Curr Allergy Asthma Rep 17(3):16. https://doi.org/10.1007/s11882-017-0685-2
2. Branche AR, Falsey AR (2016) Parainfluenza virus infection. Semin Respir Crit Care Med 37(4):538–554. https://doi.org/10.1055/s-0036-1584798
3. Schomacker H, Schaap-Nutt A, Collins PL, Schmidt AC (2012) Pathogenesis of acute respiratory illness caused by human parainfluenza viruses. Curr Opin Virol 2(3):294–299. https://doi.org/10.1016/j.coviro.2012.02.001
4. Abedi GR, Prill MM, Langley GE, Wikswo ME, Weinberg GA, Curns AT et al (2014) Estimates of parainfluenza virus-associated hospitalizations and cost among children aged less than 5 years in the United States, 1998–2010. J Pediatric Infectious Dis Soc 5(1):7–13. https://doi.org/10.1093/jpids/piu047
5. Moscona A (1997) Interaction of human parainfluenza virus type 3 with the host cell surface. Pediatr Infect Dis J 16(10):917–924. https://doi.org/10.1097/00006454-199710000-00003
6. Moscona A (2005) Entry of parainfluenza virus into cells as a target for interrupting childhood respiratory disease. J Clin Invest 115(7):1688–1698. https://doi.org/10.1172/JCI25669
7. Porotto M, Palmer SG, Palermo LM, Moscona A (2012) Mechanism of fusion triggering by human parainfluenza virus type III: communication between viral glycoproteins during entry. J Biol Chem 287(1):778–793. https://doi.org/10.1074/jbc.M111.298059
8. Marcink TC, Yariv E, Rybkina K, Mas V, Bovier FT, des Georges A et al (2020) Hijacking the fusion complex of human parainfluenza virus as an antiviral strategy. mBio. 11(1). https://doi.org/10.1128/mBio.03203-19.
9. Lawrence MC, Borg NA, Streltsov VA, Pilling PA, Epa VC, Varghese JN et al (2004) Structure of the haemagglutinin-neuraminidase from human parainfluenza virus type III. J Mol Biol 335(5):1343–1357. https://doi.org/10.1016/j.jmb.2003.11.032
10. Tian W, Chen C, Lei X, Zhao J, Liang J (2018) CASTp 3.0: computed atlas of surface topography of proteins. Nucleic Acids Research 46(W1):W363–W7. https://doi.org/10.1093/nar/gky473
11. Mizuta K, Tsukagoshi H, Ikeda T, Aoki Y, Abiko C, Itagaki T et al (2014) Molecular evolution of the haemagglutinin–neuraminidase gene in human parainfluenza virus type 3 isolates from children with acute respiratory illness in Yamagata prefecture. Japan 63(4):570–577. https://doi.org/10.1099/jmm.0.068189-0
12. Wei C-H, Kao H-Y, Lu Z (2013) PubTator: a web-based text mining tool for assisting Biocuration. Nucleic Acids Res 41(W1):W518–W522
13. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S et al (2020) PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 49(D1):D1388–D1395. https://doi.org/10.1093/nar/gkaa971
14. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. Journal of Cheminformatics 3(1):33. https://doi.org/10.1186/1758-2946-3-33
15. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS et al (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 30(16):2785–2791. https://doi.org/10.1002/jcc.21256
16. Rizvi SM, Shakil S, Haneef M (2013) A simple click by click protocol to perform docking: AutoDock 4.2 made easy for non-bioinformaticians. EXCLI journal 12:831–57
17. Kim S, Thiessen PA, Bolton EE, Bryant SH (2015) PUG-SOAP and PUG-REST: web services for programmatic access

to chemical information in PubChem. Nucleic Acids Res 43(W1):W605–W611. https://doi.org/10.1093/nar/gkv396
18. Daina A, Michielin O, Zoete V (2017) SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep 7(1):42717. https://doi.org/10.1038/srep42717
19. Ghose AK, Viswanadhan VN, Wendoloski JJ (1999) A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. J Combinatorial Chem 1(1):55–68. https://doi.org/10.1021/cc9800071
20. Lipinski CA (2004) Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol 1(4):337–341. https://doi.org/10.1016/j.ddtec.2004.11.007
21. Bowers KJ, Chow E, Xu H, Dror RO, Eastwood MP, Gregersen BA et al (2006) Scalable algorithms for molecular dynamics simulations on commodity clusters. Sc '06. 84–es. https://doi.org/10.1145/1188455.1188544.
22. Schrodinger Release 2021–1 (2021) Desmond molecular dynamics system: Schrödinger, LLC, New York, NY, 2021.
23. Jacobson MP, Friesner RA, Xiang Z, Honig B (2002) On the role of the crystal environment in determining protein side-chain conformations. J Mol Biol 320(3):597–608. https://doi.org/10.1016/S0022-2836(02)00470-9
24. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE et al (2004) A hierarchical approach to all-atom protein loop prediction. Proteins 55(2):351–367. https://doi.org/10.1002/prot.10613
25. Schrodinger Release 2021–1 (2021) Prime: Schrödinger, LLC, New York, NY, 2021.
26. Genheden S, Ryde U (2015) The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Discov 10(5):449–461. https://doi.org/10.1517/17460441.2015.1032936
27. Li J, Abel R, Zhu K, Cao Y, Zhao S, Friesner RA (2011) The VSGB 2.0 model: a next generation energy model for high resolution protein structure modeling. Proteins 79(10):2794–812. https://doi.org/10.1002/prot.23106
28. Alymova IV, Taylor G, Takimoto T, Lin TH, Chand P, Babu YS et al (2004) Efficacy of novel hemagglutinin-neuraminidase inhibitors BCX 2798 and BCX 2855 against human parainfluenza viruses in vitro and in vivo. Antimicrob Agents Chemother 48(5):1495–1502. https://doi.org/10.1128/aac.48.5.1495-1502.2004
29. Alymova IV, Watanabe M, Boyd KL, Chand P, Babu YS, Portner A (2009) Efficacy of the novel parainfluenza virus haemagglutinin-neuraminidase inhibitor BCX 2798 in mice - further evaluation. Antivir Ther 14(7):891–898. https://doi.org/10.3851/imp1420
30. Watanabe M, Mishin VP, Brown SA, Russell CJ, Boyd K, Babu YS et al (2009) Effect of hemagglutinin-neuraminidase inhibitors BCX 2798 and BCX 2855 on growth and pathogenicity of Sendai/human parainfluenza type 3 chimera virus in mice. Antimicrob Agents Chemother 53(9):3942–3951. https://doi.org/10.1128/aac.00220-09
31. Palm K, Stenberg P, Luthman K, Artursson P (1997) Polar molecular surface properties predict the intestinal absorption of drugs in humans. Pharm Res 14(5):568–571. https://doi.org/10.1023/a:1012188625088
32. Ertl P, Rohde B, Selzer P (2000) Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J Med Chem 43(20):3714–3717. https://doi.org/10.1021/jm000942e
33. Prakash M, Basavaraj BV, Chidambara Murthy KN (2019) Biological functions of epicatechin: plant cell to human cell health. J Funct Food 52:14–24. https://doi.org/10.1016/j.jff.2018.10.021
34. Nawrot-Hadzik I, Matkowski A, Kubasiewicz-Ross P, Hadzik J (2021) Proanthocyanidins and flavan-3-ols in the prevention and treatment of periodontitis-immunomodulatory effects, animal and clinical studies. Nutrients 13(1):239. https://doi.org/10.3390/nu13010239
35. Ahmad A, Kaleem M, Ahmed Z, Shafiq H (2015) Therapeutic potential of flavonoids and their mechanism of action against microbial and viral infections—a review. Food Res Int 77:221–235. https://doi.org/10.1016/j.foodres.2015.06.021
36. Pretorius JC, Magama S, Zietsman PC (2003) Purification and identification of antibacterial compounds from Euclea crispa subsp crispa (Ebenaceae) leaves. South African J Botany 69(4):579–86. https://doi.org/10.1016/S0254-6299(15)30298-2
37. Zakaryan H, Arabyan E, Oo A, Zandi K (2017) Flavonoids: promising natural compounds against viral infections. Adv Virol 162(9):2539–2551. https://doi.org/10.1007/s00705-017-3417-y
38. Aquino R, De Simone F, De Tommasi N, Pizza C (1995) Structure and biological activity of triterpenoids and aromatic compounds from medicinal plants. In: Atta ur R, editor. Studies in Natural Products Chemistry, Elsevier, p 113–52
39. Wang L, Song J, Liu A, Xiao B, Li S, Wen Z et al (2020) Research progress of the antiviral bioactivities of natural flavonoids. Nat Prod Bioprospecting 10(5):271–283. https://doi.org/10.1007/s13659-020-00257-x
40. Lyu SY, Rhim JY, Park WB (2005) Antiherpetic activities of flavonoids against herpes simplex virus type 1 (HSV-1) and type 2 (HSV-2) in vitro. Arch Pharmacal Res 28(11):1293–1301. https://doi.org/10.1007/BF02978215
41. Borges G, Ottaviani JI, van der Hooft JJJ, Schroeter H, Crozier A (2018) Absorption, metabolism, distribution and excretion of (−)-epicatechin: a review of recent findings. Mol Aspects Med 61:18–30. https://doi.org/10.1016/j.mam.2017.11.002
42. Müller P, Downard KM (2015) Catechin inhibition of influenza neuraminidase and its molecular basis with mass spectrometry. J Pharm Biomed Anal 111:222–230. https://doi.org/10.1016/j.jpba.2015.03.014
43. Liu A-L, Wang H-D, Lee SM, Wang Y-T, Du G-H (2008) Structure–activity relationship of flavonoids as influenza virus neuraminidase inhibitors and their in vitro antiviral activities. Bioorg Med Chem 16(15):7141–7147. https://doi.org/10.1016/j.bmc.2008.06.049

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

Springer Nature or its licensor 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.

AIP Conference Proceedings

# Estimation of air flow around the tall buildings of different shapes

Cite as: AIP Conference Proceedings **2520**, 030020 (2022); https://doi.org/10.1063/5.0102942
Published Online: 30 August 2022

**Rahul Kumar Meena, Ritu Raj and S. Anbukumar**

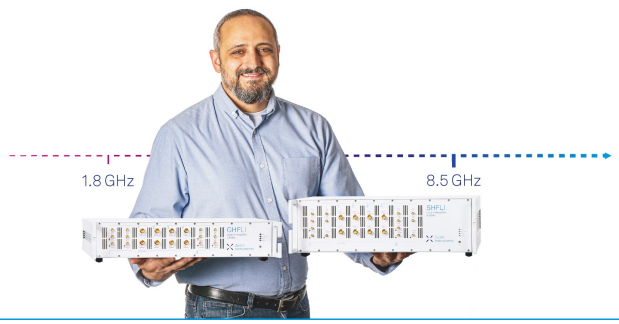
View Online Export Citation

**ARTICLES YOU MAY BE INTERESTED IN**

Real-time earthquake detection using Twitter tweets
AIP Conference Proceedings **2520**, 030014 (2022); https://doi.org/10.1063/5.0102903

Effect of turbulent inflow on the leading-edge protuberanced (LEP) wing: A time-series approach
AIP Conference Proceedings **2520**, 030001 (2022); https://doi.org/10.1063/5.0102943

Model of performance management analysis for an infrastructure project
AIP Conference Proceedings **2520**, 030018 (2022); https://doi.org/10.1063/5.0103271




Trailblazers. New
Meet the Lock-in Amplifiers that measure microwaves.
Zurich Instruments
Find out more



© 2022 Author(s).

# Estimation of Air Flow Around the Tall Buildings of Different Shapes

Rahul Kumar Meena [a], Ritu Raj [b] and S Anbukumar [c]

*Delhi Technological University, Delhi, India, 110042*

[a] *rahulkumar_phd2k18@dtu.ac.in*
[b] *rituraj@dtu.ac.in*
[c] *sanbukumar@dce.ac.in*

*Corresponding author:* [a] *rahulkumar_phd2k18@dtu.ac.in*

**Abstract.** Tall building is the growing pattern of cities, these buildings are highly vulnerable to wind loads. Wind loads effects on the high-rise building can be found using wind tunnel testing and CFD tools. Wind tunnel facilities are limited all around the world, so CFD tools can be incorporated. In this study ANSYS CFX is used to investigate the effects of wind on different shape of high-rise buildings. Results of numerical simulation are compared with international standards of various countries. Stream line pattern and pressure contours on different faces are discussed in this paper. It is found that building shape and size are having the direct effect on the pressure distribution on various surfaces of building model. Wind ward face pressure will always be positive in nature.

## 1. INTRODUCTION

High-rise residential buildings have become more widespread in many cities as a result of increasing population and rising urbanization, displacing broad regions of vernacular dwellings. These constructions are vulnerable to time- varying loads induced by winds, earthquakes, and other natural calamities, in addition to gravity loads. Tall buildings have a high sensitivity to wind excitation.[1]–[4]. In general, a structure cannot be considered a regular configuration when planning for wind loading. To achieve the structural pressure coefficients and force coefficients for other structures subjected to wind-induced loads, designers consult related wind load standards[5]–[9]. On the other hand, these standards offer details for plain cross-sectional configurations with a limited number of wind incidence angles. These codes do not provide information on wind loadings for structures with different configurations. As a result, wind tunnel research on models of such forms is popular. Raj and Ahuja [10] conducted an experimental study on cross-shape tall buildings with various cross-sectional shapes, observed that base shear, base moments, and twisting moments are influenced by wind incidence angle and affected by cross-sectional shape. Vafaeihosseini et al.[11]used computational fluid dynamics to find wind effects on high-rise buildings gives that the most efficient shape for tall buildings is rectangular plan shape. Hajra and Dalui [12] performed the mathematical-based research of interference effect on octagonal plan configuration high structure using CFX (ANSYS), for $0^0$ wind incidence angle using *k-ε,* SST, and *k*-ω model, these three models analysis shows nearly identical results. Meena et al.[13] discussed the effects of wind on the bracing system used in various types of multi-storeyed steel buildings. Pal et al.[14] carried out a comparative study on square and fish plan shaped tall building of wind-induced mutual interference effect, maximum efficiency in terms of wind-induced mutual interference pressure, and base shear ix exhibited by square plan shaped model at full blockage condition. Verma et

*Proceedings of the International Conference on Research Advances in Engineering and Technology - ITechCET 2021*
AIP Conf. Proc. 2520, 030020-1–030020-9; https://doi.org/10.1063/5.0102942
Published by AIP Publishing. 978-0-7354-4384-6/$30.00

al.[15] analyzed the effects of wind load on tall building octagonal shape, using CFD simulation for $0^0$,$15^0$ and $30^0$ wind incidence angle, found that CFD can be used to predict wind-related phenomena on building and other types of structure. Nagar et al.[16] carried out an experimental study of wind- induced pressure on tall buildings of different shapes from $0^0$ to $90^0$ wind incidence angle at an interval of $30^0$, found that wind pressure decreases up to 600 wind incidence angle, "H" plan shaped building is subjected to a higher pressure than the square Model.

Wind tunnel testing requires a costly setup and sophisticated instruments to measure a range of filed variables wind velocity, pressure loads, turbulent intensity etc. It is a main limitation that such measurements are a few selected points within the test section, severally restring the unsteady complex phenomena such as vortex shredding, turbulence wake.

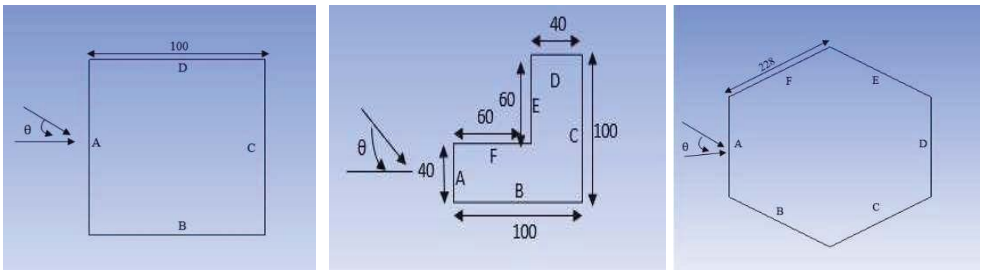
CFD work by dividing a space into a grid contains a large number of cells. The grid of the cells is surrounded by boundaries that simulates the surfaces and opening, that enclose the space, the pressure of the boundaries, the air movements of at opening within the cell is than set to a starting condition, which is closed to the real environments. These conditions are determined using boundary model that predicts boundary conditions, the more accurate the boundary model, and the closer to starting condition is to the final position predicted to the model, the faster the model will run and more accurate the output is likely to be. CFD results are highly dependent on the knowledge of the person setting up the model and interpreting the results, as the input information is wrong, the output information will be as well, CFD has no substitute for common sense.

The wind forces acting on a building and the motion that arises are directly influenced by the shape of the building. By carefully designing the construction components and geometry of tall buildings, wind excitation is reduced and costs are reduced. Passive aerodynamics adjustments in the form of building shape are one of the most potent and adaptable construction approaches for considerably minimizing the impacts of wind forces by changing/altering the flow pattern around the buildings.

## 2. MATHEMATICAL MODELLING BY ANSYS CFX

**Model**

To study the effects of wind, different shape of tall building model are considered in this study, in Fig. 1 (a),(b), and (c). with a geometrical scale of 1:200 is considered.




(a) (b) (c)

**FIGURE 1.** Models with Dimension

**Boundary Conditions**

The domain boundary must be specified to describe a problem with a unique solution. Boundary circumstances that aren't clearly established are having a big impact on your solution. Identifying the solver, the inlet, outlet wall, and other boundary conditions The domain, boundary conditions are employed to carry out the numerical simulation, according to Zidan et al.[17]].

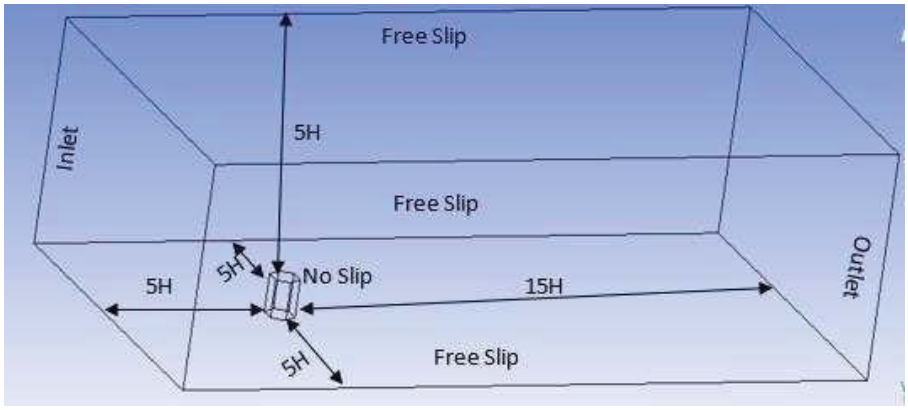



**FIGURE 2.** Domain

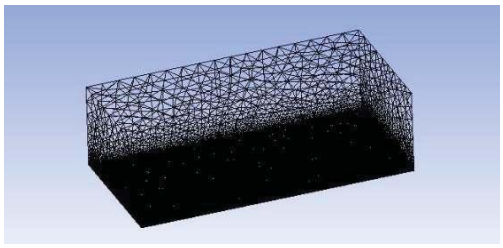


Fig. 3(a) Domain meshing

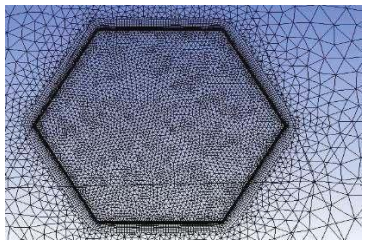


Fig. 3(b) Inflation

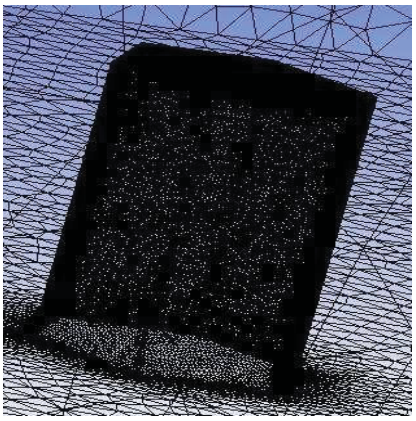


Fig. 3(c) Building meshing

Fig. 3(d) Edge meshing

**FIGURE. 3** Types of meshing

The purpose of meshing is to improve the accuracy of the solution generated during simulation. This can be done manually or automatically with ANSYS CFX. Meshing for distinct portions can be applied in the manual technique, and meshing size is chosen based on the situation. To reduce the anomalous flow, inflation is applied to all models in the CFD simulation. Domains with tetrahedron meshing, structure, and ground meshing are more delicate in size, as seen in Fig.3. It improves the precision of the solution. Figure 3(d) shows edge meshing, and Figure 3(b) shows inflation, both of which are employed to reduce the abnormal flow.

### VALIDATION

Before starting the numerical analysis of building the validity of the ANSYS CFX package is checked. For this reason, a square plan shaped building of dimension 100 mm × 100 mm.

Table 1 Comparison of face average values of coefficients of pressure

| Wind loading code | Face- A | Face-B | Face-C | Face-D |
|---|---|---|---|---|
| By ANSYS CFX | 0.87 | -0.54 | -0.78 | -0.78 |
| ASCE 7-10 | 0.8 | -0.5 | -0.7 | -0.7 |
| AS/NZS-1170.2(2002) | 0.8 | -0.5 | -0.65 | -0.65 |
| IS: 875 (part3)(2015) | 0.8 | -0.25 | -0.8 | -0.8 |

## 3 RESULTS AND DISCUSSION

### 3.1 Velocity Profile and Turbulent Profile

Wind blow due to the pressure difference over a different part of the earth and Coriolis force. Due to the ground roughness atmospheric boundary layer is formed near the ground. And the wind becomes highly turbulent due to ground obstacles. The degree of ground roughness and drag due to local projections that resist wind flow determines the vertical profile of wind speed. Gradient height is the height at which the drag effects disappear, and gradient velocity is the corresponding velocity. The atmospheric boundary layer is the height at which topography influences wind speed**.** As per Power Law, the wind speed profile, as shown in Fig. 4, within the atmospheric boundary layer is given by equation (1)

$$\frac{U}{U_H} = \left(\frac{Z}{Z_H}\right)^{\alpha} \tag{1}$$

Where U is the horizontal wind speed at an elevation Z; $U_H$ is the speed at the reference elevation $Z_H$, which was 10 m/s, α is the parameter that varies with ground roughness which is 0.147 for terrain category 2, and $Z_H$ is 1.0 m.

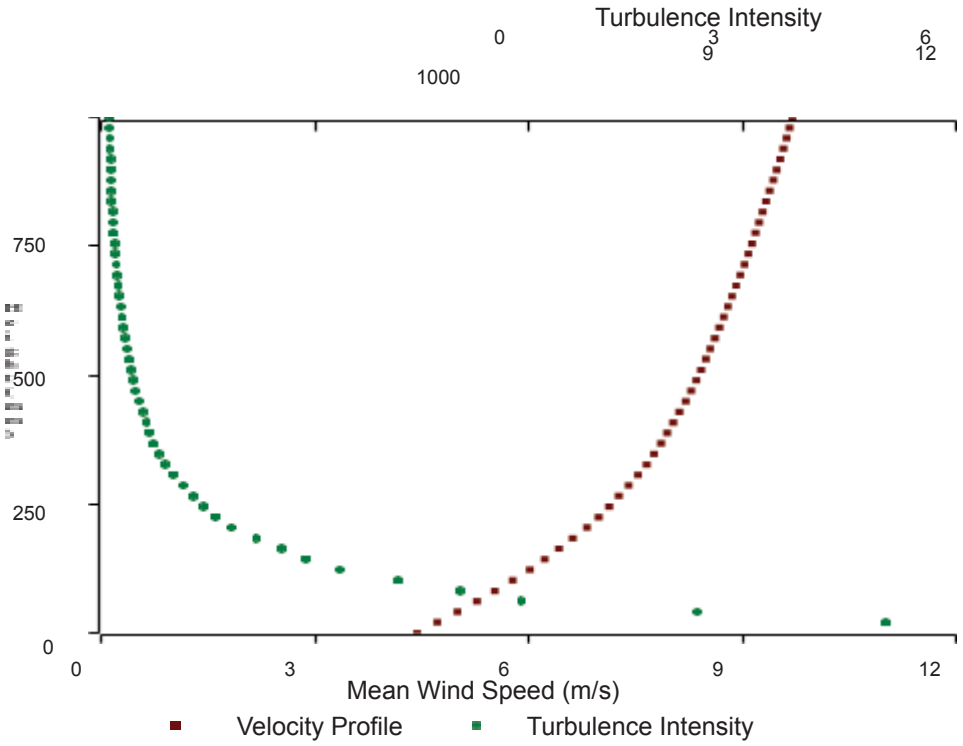



**FIGURE 4.** Variation of Velocity and Turbulence Intensity with height

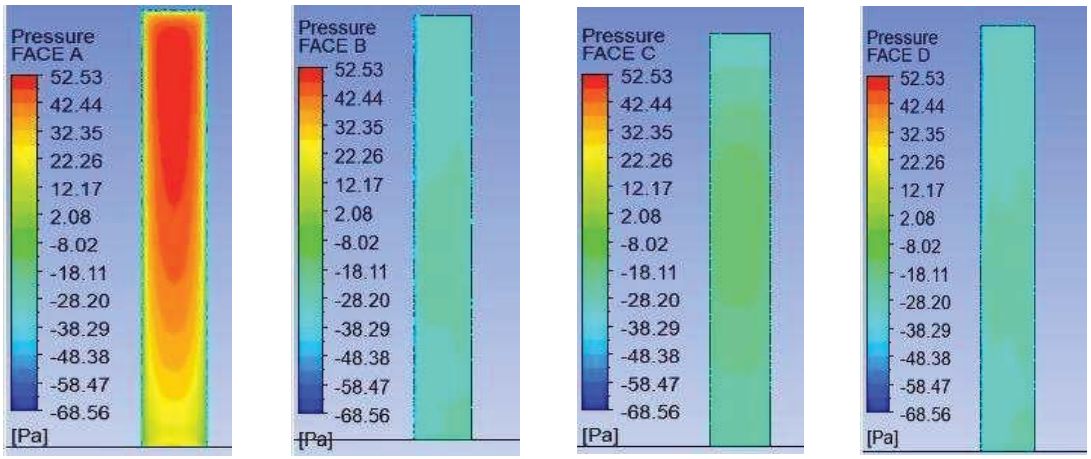
Fig.4 shows that wind speed near the surface is very low, it's because of the friction between the surface and moving air. The wind profile obtained in this study is similar to the wind profile provided in ASCE 7 and previous research done in this area.

## Pressure Contour

The pressure on the windward face is positive for the models, as shown in Fig.5,6 and Fig. 7, and for the windward face and leeward face, it is negative. The graphical representation of pressure on every face of both model's A,B and C are shown in Fig. 5,6 and Fig. 7, respectively.

## Velocity Streamlines

A streamline is an imaginary line in a fluid whose tangent at any point represents the direction of the velocity of a fluid particle at that location. Because various flow patterns are observed for all three models, and it is due to the variation in the plan shape of the building, the streamline for $0^0$ wind incidence angle for model A, model B, and model C in the plan and elevation is shown in Fig. 8. In the wake zone, vortices are formed. On the leeward side of the building models, it will cause negative pressure. On the downstream side of building models for model B and model C, a vast recirculation zone of air can be seen.




Face A Face B Face C Face D

**FIGURE 5. Pressure Contour for Model A at $0^0$ wind Incidence Angle**

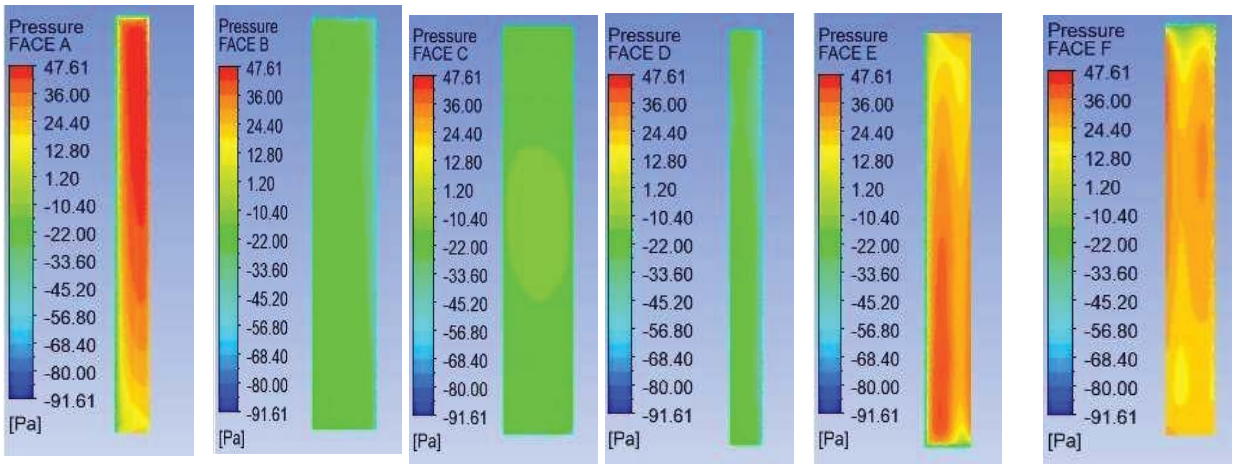



**FIGURE 6 Pressure Contour for Model B at $0^0$ wind Incidence Angle**

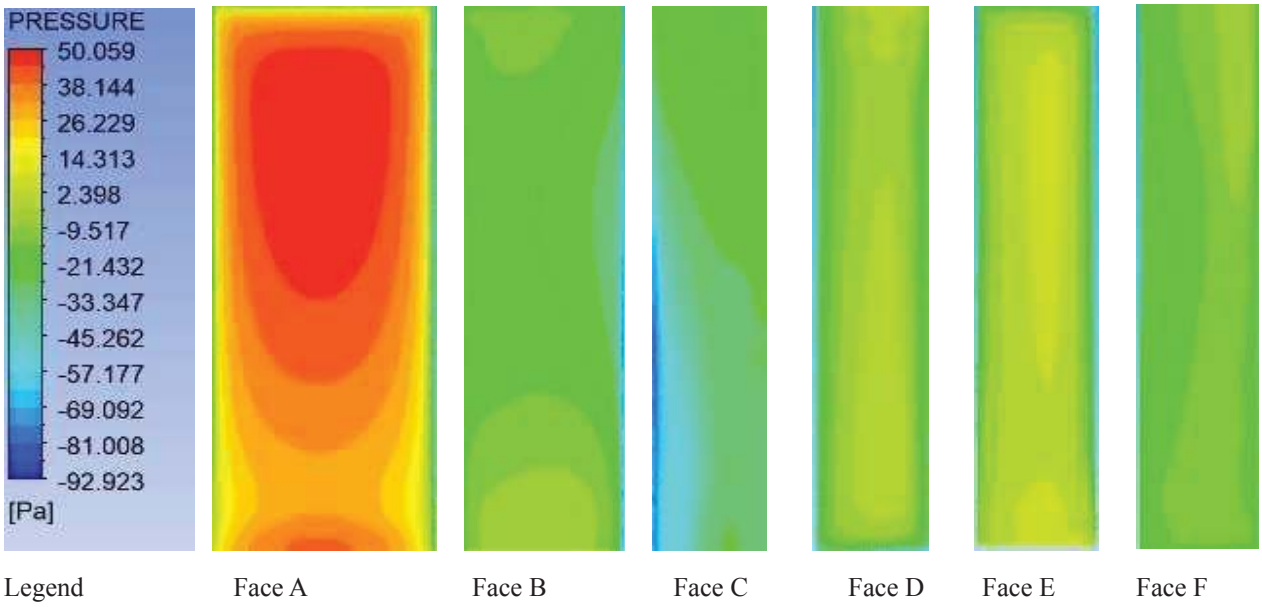



**FIGURE 7. Pressure Contour for Model C at $0^0$ wind Incidence Angle**

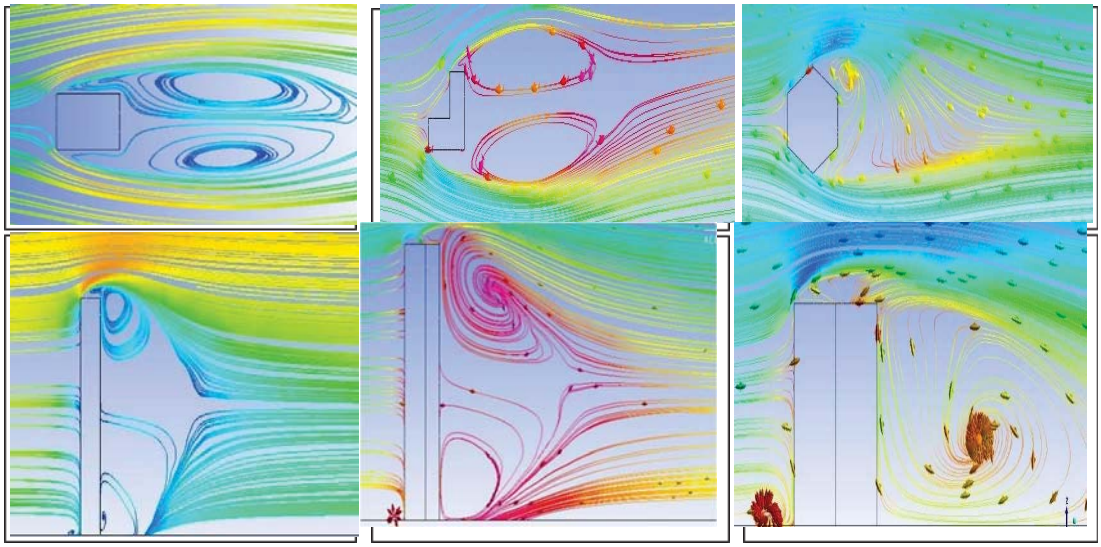


**FIGURE 8.** Velocity StreamLines in Plan and Elevation

### Pressure Coefficient

The mean pressure coefficient '$C_{p\ mean}$' is calculated from the equation (2)

$$C_{p\ mean} = \frac{p - p_o}{\frac{1}{2}\rho U^2} - \text{H} \qquad (2)$$

where $p$ is the pressure derived from the needed point, $p_o$ is the reference height static pressure, $\rho$ is the air density (1.225 kg/m$^3$), and $U_H^2$ is mean wind velocity at the structure reference height.

The mean Cp for square model (building model-a), L shape (building model-b) and hexagon shape building (building model-c) are presented in fig. 7, 8 and fig.9 and compared with various international standards.

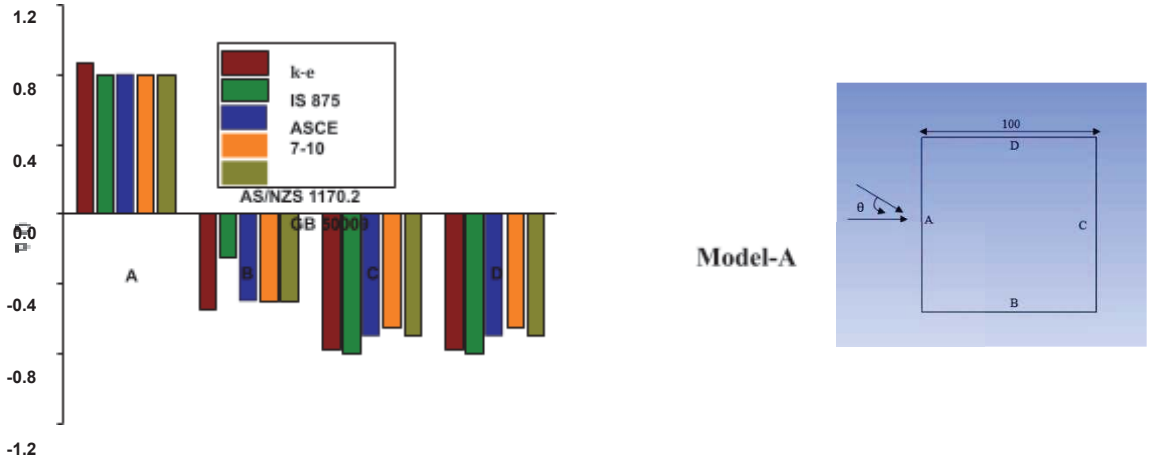



**FIGURE 9.** Mean Cp for Square Building model- A

Mean Cp calculated for square shape building model are compared with different international standards and found in the range of various codes, for different faces. L shape building model-b, has three face in positive pressure distribution in fig. 10 and for hexagon shape building model-c, in fig. 11 depict that only face a is having positive pressure distribution. Fig. 9 also shows mean Cp for square building model-a, only wind ward face has positive pressure distribution, rest all faces are under the effect of negative pressure distribution that is due suction on lee ward side and side faces of building model.

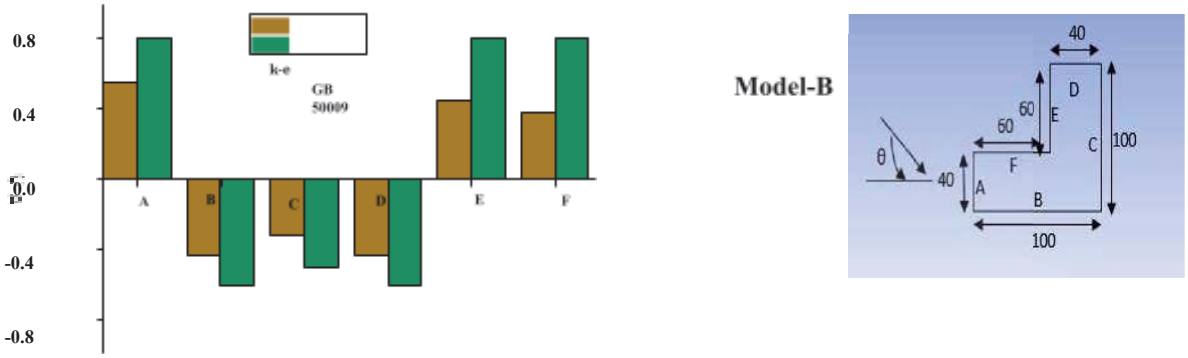



**FIGURE 10.** Mean Cp for L-shape Building model- B

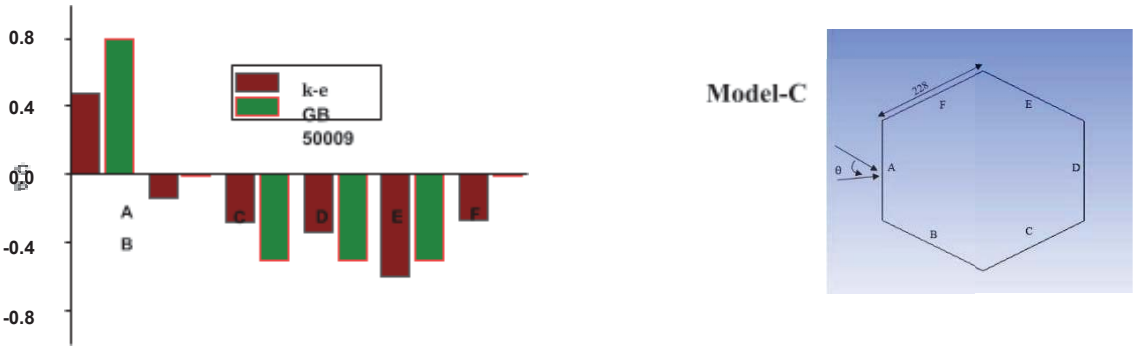



**FIGURE 11.** Mean Cp for Hexagon Building model- C

## CONCLUSION

At $0^0$ wind incidence angles, the pressure contour, mean pressure coefficients, and velocity streamlines for model-a,b and model c designed and studied in this work. This study is simulated using the k-ε model. The following are the research's major findings:

- Face A is always showing positive pressure for all the models, whereas the other face has negative pressure
- The pressure distribution of all faces of the building is the aim to conduct this study, and the pressure contour of all faces is presented in this paper.
- Changing in the plan shape of building considerably affects the suction on leeward and side faces of the building models.
- The maximum pressure coefficient for model B(L-shape building) is 0.8 on windward face A and a minimum of 0.78 on leeward face D
- The velocity streamlines are shown using the figure in the plan and elevation.
- The work's precision depends upon meshing the geometry model and defining the flow physics as in boundary layer wind tunnel.

## ACKNOWLEDGEMENT

The authors are thankful to **Delhi Technological University** for providing the research facilities and funding to support this study. authors would like to express their appreciation to Manvendra Verma, Indrajeet Singh, Deepak Singh, Dinesh Reddy, Abhishek Prakash Paswan, Harshit Kumar Jayant for their precious help with the numerical simulation and to DTU civil engineering faculty, who encouraged and supported the authors during the entire study.

## REFERENCES


[1] B. S. Taranath, *Wind and Earthquake Resistant Buildings*. 2004.

[2] E. Simiu and D. Yeo, *Wind effects on structures*, Fourth Edi. John Wiley & Sons Ltd, 2019.

[3] M. H. Günel and H. E. Ilgin, *Tall buildings: Structural systems and aerodynamic form*. 2014.

[4] G. Solari, *Wind actions and effects on structures*. 2019.

[5] IS: 875 (2015), *Indian Standard design loads (other than earthquake) for buildings and structures-code of practice,part 3(wind loads)*. 2015.

[6] ASCE: 7-10(2013), *Minimum Design Loads for Buildings and Other Structures. Structural Engineering Institute of the American Society of Civil Engineering, Reston*. 2013.

[7] GB 50009-2001, *NATIONAL STANDARD OF THE PEOPLE'S REPUBLIC OF CHINA*. 2002.

[8] ETHIOPIAN STANDARD, *ES ISO 4354 (2012) (English): Wind actions on structures*, vol. 2012. 2012.

[9] AS/NZS:1170.2(2011), *Structural Design Actions - Part 2: Wind actions. Standards Australia/Standards New Zealand, Sydney*. 2011.

[10] R. Raj and A. K. Ahuja, "Wind Loads on Cross Shape Tall Buildings," *Journal of Academia and Industrial Research (JAIR)*, vol. 2, no. 2, pp. 111–113, 2013.

[11] C. C. Colloquium, "Computational Fluid Dynamics Approach for Wind Analysis of Highrise Buildings," no. January, 2013.

[12] S. Hajra and S. K. Dalui, "Numerical investigation of interference effect on octagonal plan shaped tall buildings," *Jordan Journal of Civil Engineering*, vol. 10, no. 4, pp. 462–479, 2016.

[13] R. K. MEENA, G. P. Awadhiya, A. P. Paswan, and H. K. Jayant, "Effects of Bracing System on Multistoryed Steel Building," 2018, doi: 10.1088/1757-899X/1128/1/012017.

[14] S. Pal, R. Raj, and S. Anbukumar, "Comparative study of wind induced mutual interference effects on square and fish-plan shape tall buildings," *Sādhanā*, vol. 0123456789, 2021, doi: 10.1007/s12046-021-01592-6.

[15] D. S. K. Verma, A. . Roy, S. Lather, and M. Sood, "CFD Simulation for Wind Load on Octagonal Tall Buildings," *International Journal of Engineering Trends and Technology*, vol. 24, no. 4, pp. 211–216, 2015, doi: 10.14445/22315381/ijett-v24p239.

[16] S. K. Nagar, R. Raj, and N. Dev, "Experimental study of wind - induced pressures on tall buildings of different shapes," *Wind and Structures, An International Journal*, vol. 5, pp. 441–453, 2020.

[17] Y. F. Abu-Zidan, "Verification and Validation Framework for Computational Fluid Dynamics Simulation of Wind Loads on Tall Buildings," no. March, 2019.

# Facial Micro-Expression Recognition (FMER) using Model Compression

Nikhil Singh
*Research Scholar: Department of Electronics & Communication Engineering*
*Delhi Technological University*
Delhi, India
nikhil_2k19phdec04@dtu.ac.in

Rajiv Kapoor
*Professor: Department of Electronics & Communication Engineering*
*Delhi Technological University*
Delhi, India
rajivkapoor@dce.ac.in

2022 International Conference on Electrical, Computer and Energy Technologies (ICECET) | 978-1-6654-7087-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECET55527.2022.9872920

*Abstract*— **Emotion recognition that utilizes facial expressions is critical for AI systems like social robots to interact effectively with humans. In the actual world, however, it is considerably more difficult to distinguish face micro-expressions (FMEs) than it is to recognize facial general-expressions with complex emotions. It's a form of facial expression that lasts only a few seconds, has a modest magnitude, and usually local movement. However, because of these properties of MEs (micro-expressions), obtaining ME data is challenging, which is a constraint when using deep learning algorithms to recognize FMEs. The facial action coding system (FACS) can likewise encode these FMEs which represents that there is some relation between Action Unit (AU) and FMEs. To demonstrate this, a knowledge transfer scheme for ME recognition that compress and transfers information from an AU is given in this manuscript. Experiments are carried out on four publicly available ME database. Our model outperforms state-of-the-art systems, according on the findings of the experiments.**

***Keywords- Deep Neural Network, FME-Recognition, Knowledge Transfer, Model Compression.***

## I. INTRODUCTION

Emotion detection technology based on facial expression, action, and voice has recently been investigated as a means of assisting intelligent systems. Recognizing emotions associated with macro-expressions is not difficult in general [1]. People rarely reflect their feelings on their faces in practical terms. Therefore, the most active research area is on FME emotion recognition. [2], [3], [4], [5], [6]. In terms of face expression, even the same emotions might have vastly different quantitative values. For ultimate emotion recognition, we must manage even such a FME. Regrettably, there haven't been many previous investigations on FMER. Furthermore, datasets for FMER research are typically derived from broadcast content or films in which actors or actresses purposefully construct their facial expressions [7].

978-1-6654-7087-2/22/$31.00 ©2022 IEEE

The concept of micro-expression was initially proposed in this project in 1966. Micro-expression was first described in 1969. In a dialogue between a psychotherapist and a depressed client, they discovered that the client who grins frequently has several very painful emotions. MEs is a term used to describe quick, involuntary, and impulsive facial gestures. It frequently arises when people are experiencing strong emotions. ME is a sort of expression that lasts for a short period of time. MEs normally last 1/25 to 1/3 second. For macro-expression recognition, crucial temporal selection is beneficial [8]. As a result, we used three different temporal selection strategies to recognize MEs. The FACS, is a widely wielded programming system in the discipline of macro-affect recognition, can also encode micro-expression. It has a series of pre-programmed source code sheets, all of which represents some certain facial movement and is referred to as an AU (action unit).

The following is a summary of previous FMER research: An FME recognition technique was proposed in [9] using a temporal interpolation model and a random forest. The author presented a hierarchical spatio-temporal descriptor for controlling feature weight by looking for subtle facial muscle action [10]. Many FMER research used LBP-TOP (Local Binary Patterns on three orthogonal planes) and similar models [11]. For micro-expression analysis, suggested the Local binary pattern with six intersection points (LBP-SIP). The key contribution of this study is that it diminishes the aspects of features and enhances feature extraction efficacy. Using main directional mean optical flow, [12] devised an FMER system (MDMO). They used sparse coding to extract the atomic feature representing the region of interest from the optical flow data and then categorized the result using SVM. A ME grand challenge (MEGC) has been run for a few years to stimulate the competitive development of FME recognition algorithms [13].

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2022 at 05:03:06 UTC from IEEE Xplore. Restrictions apply.

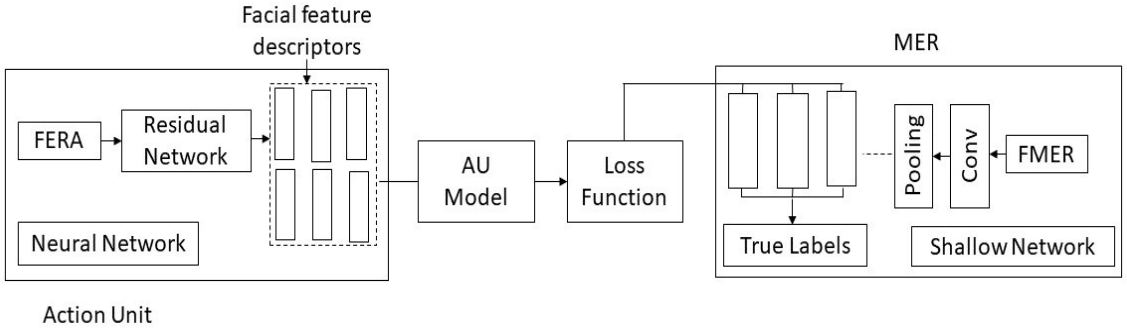



Fig.1. Block Diagram of Proposed Algorithm

Deep learning schemes are very common in the domain of affective computing. Deep learning also necessitates abundance of data to train with. When data for training purpose in not sufficient, in such circumstances we always pre-train the system on a related database before fine-tuning it on the driven database. For FMEs the size issue is always there.

To overcome the aforementioned problems, we offer a unique model compression scheme that compress and transfers multi-model information from AU for MEs detection having ground truth of critical temporal sequences. An objective function is used to construct a tutor-tutee correlative structure. Finally, knowledge relating to multi-level AU is collected for ME recognition based on critical temporal sequences.

The rest of this paper is laid out as follows. Section 2 delves into the specifics of the suggested methodology. Section 3 describes the experiments in full. Section 4 concludes with our conclusions and plans for future work.

## II. METHODOLOGY

### A. *Pre-Processing*

Temporal phase selection, face identification and synchronization, and image size normalization are all part of the data pre-processing procedure. ME representation is, in general, a phased and physiologically limited structure. It features a general temporal phase structure that includes terms like offset, peak, onset and neutral, among others

To draw out and recognize the face, we use the Viola Jones algorithm. Through affine transformation, all sequences are set to this basis face. The frames have been reduced in size to 160 x 160 pixels.

### B. *Transfer Learning Framework*

We take into account the properties of FMEs as well as the known relationship between AU and MEs, and apply face AU domain knowledge to the FMEs domain. In this manuscript a model compression based FMER system is proposed. It consists of a shallow NN and a pre-trained instructor DNN. The goal of developing this context was to apply the deep learning scheme to a tiny ME database and use AU information for MER.

1. *Pre-trained DNN*

The residual network is divided into three categories. Each group contains four blocks, the structure of which is depicted in Fig.2. The following formula can be used to express a residual block with identity modeling:

$$p_{l+1} = p_l + R(p_l, t_l) \quad (1)$$

where, $p_l$ is the input & $p_{l+1}$ is the output, residual is given by R and $t_l$. Our model's loss is determined as follows:

$$F_{inst} = \beta_1 F_1 + \beta_2 F_2 \quad (2)$$

Where, $F_{inst}$ is the neural network leading loss, $F_1$ & $F_2$ is the detection and face identification loss of AU. To balance the sub-task $\beta_1$ & $\beta_2$ is introduced.

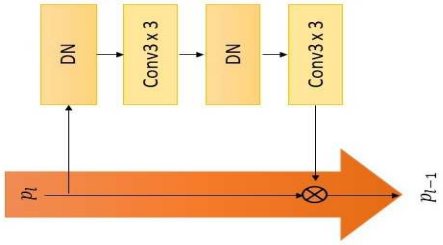



Fig.2. Residual block architecture

$F_{inst}$ is the result of the last two fully connected layers' output. The shallow neural network is then guided by these features [14].

2. *Shallow Neural Network*

The training of a deep network with many parameters necessitates a lot of computing [15], [16]. However, the micro-expression data set is insufficient to facilitate deep network training. As a result, we're attempting to reduce network size while preserving performance. We look at the way of model compression. $F_{tut}$ is made up of the output features from the last two completely linked layers [17]. In

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2022 at 05:03:06 UTC from IEEE Xplore. Restrictions apply.

this paper, we define MEs as a sort of expression that uses characters similar to those used in ordinary expressions. As a result, for a given input image, $F_{sh}$ should be similar to $F_{inst}$, which is the kernel of given network. Hence loss is given by:

$$F_{tut} = \eta_1 F_1 + \eta_2 F_2 \quad (3)$$

Where $F_{sh}$ is the shallow network leading loss.

## III. Performance Evaluation

### A. Training Model Procedure

Dataset FERA 2017 was used to train the DNN database after data pre-processing, and the best model was kept. The $F_{inst}$ was then extracted as prior data to help the shallow network in reviewing MER knowledge on the focused MEs database using the pre-trained model. The output from the shallow network's last two completely connected layers was then retrieved and spliced. The final findings were then obtained using SVM. At a given time, one participant is chosen as testing data, while the others are used as training data. Finally, the classification performance is computed by averaging the identification outcomes for every subject. This is also in line with the comparison method's technique, which we'll discuss later.

### B. Qualitative Comparison

In terms of accuracy, mean absolute error (MAE) and F-Measure qualitative comparison is done.

#### 1. Accuracy

The accuracy of a model is determined by how well it recognizes correlations and patterns between variables in a dataset using the input, or training data.

$$Acc = \frac{\sum_{i=1}^{z} TP_i}{\sum_{i=1}^{z} (TP_i + FP_i)} \quad (4)$$

where, P is precision, R is recall, z is classes of MEs

#### 2. MAE

A measure of uncertainties between paired observations describing the same phenomenon is called mean absolute error (MAE). It is given as:

$$MAE = \frac{\sum_{i=1}^{z} |FP_i - TP_i|}{z} \quad (5)$$

#### 3. F-Measure

The F-score, often known as the F-measure, is a measurement of a test's accuracy. It is calculated by dividing the number of true positive results by the entire number of positive results, including those that were mistakenly recognized, and recall by the number of true positive findings divided by the total number of samples that should have been detected as positive. It is written as follows:

$$F - Score = \frac{2PR}{P+R} \quad (6)$$

TABLE I: State of art comparison with CASME Dataset

| Reference | CASME Dataset | | |
|---|---|---|---|
| | *Accuracy* | *MAE* | *F-Score* |
| [22] | 65.45 | 7.72 | 0.58 |
| [23] | 57.89 | 8.36 | 0.61 |
| [24] | 72.61 | 7.58 | 0.67 |
| Proposed Scheme | **74.22** | **7.22** | **0.69** |

TABLE II: State of art comparison with SAVEE Dataset

| Reference | SAVEE Dataset | | |
|---|---|---|---|
| | *Accuracy* | *MAE* | *F-Score* |
| [22] | 72.44 | 7.77 | 0.69 |
| [23] | 74.53 | 8.19 | 0.58 |
| [24] | 76.06 | 7.55 | 0.71 |
| Proposed Scheme | **77.12** | **7.47** | **0.85** |

TABLE III: State of art comparison with SMIC Dataset

| Reference | SMIC Dataset | | |
|---|---|---|---|
| | *Accuracy* | *MAE* | *F-Score* |
| [22] | 72.06 | 7.55 | 0.63 |
| [23] | 74.72 | 7.52 | 0.77 |
| [24] | 71.47 | 8.19 | 0.71 |
| Proposed Scheme | **75.05** | **7.32** | **0..80** |

TABLE IV: State of art comparison with SAMM Dataset

| Reference | SAMM Dataset | | |
|---|---|---|---|
| | *Accuracy* | *MAE* | *F-Score* |
| [22] | 73.65 | 8.03 | 0.79 |
| [23] | 70.88 | 7.69 | 0.73 |
| [24] | 86.74 | 7.47 | 0.83 |
| Proposed Scheme | **90.22** | **6.81** | **0.86** |

## IV. Conclusion

This paper presents a model compression scheme based facial micro-expression recognition scheme that compress and transfer information from AU to MER. The results of the experiments reveal that our model outperforms the most recent FMER approaches. The study's findings suggest that temporal

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2022 at 05:03:06 UTC from IEEE Xplore. Restrictions apply.

selection is linked to MER. In the future, an end-to-end learning framework that considers temporal selection and FMER at the same time will be investigated.

## REFERENCES


[1] D. Y. Choi and B. C. Song, "Facial Micro-Expression Recognition Using Two-Dimensional Landmark Feature Maps," in IEEE Access, vol. 8, pp. 121549-121563, 2020, doi: 10.1109/ACCESS.2020.3006958.

[2] A. C. Le Ngo, Y.-H. Oh, R. C.-W. Phan and J. See, "Eulerian emotion magnification for subtle expression recognition", *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 1243-1247, Mar. 2016.

[3] Y. Wang, J. See, Y.-H. Oh, R. C.-W. Phan, Y. Rahulamathavan, H.-C. Ling, et al., "Effective recognition of facial micro-expressions with video motion magnification", *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21665-21690, Oct. 2017.

[4] R. Zhao, Q. Gan, S. Wang and Q. Ji, "Facial expression intensity estimation using ordinal information", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3466-3474, Jun. 2016.

[5] K. Zhang, Y. Huang, Y. Du and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks", *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193-4203, Sep. 2017.

[6] S. Xie and H. Hu, "Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks", *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 211-220, Jan. 2019.

[7] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The extended Cohn-Kanade dataset (CK): A complete dataset for action unit and emotion-specified expression", *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, pp. 94-101, Jun. 2010.

[8] Sun, Bo, et al. "Affect recognition from facial movements and body gestures by hierarchical deep spatio-temporal features and fusion strategy," Neural Networks, vol.105, pp.36-51, 2018.

[9] T. Pfister, X. Li, G. Zhao and M. Pietikainen, "Recognising spontaneous facial micro-expressions", *Proc. Int. Conf. Comput. Vis.*, pp. 1449-1456, Nov. 2011.

[10] Y. Zong, X. Huang, W. Zheng, Z. Cui and G. Zhao, "Learning from hierarchical spatiotemporal descriptors for micro-expression recognition", *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3160-3172, Nov. 2018.

[11] Huang, Xiaohua, et al. "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," Neurocomputing, vol. 175, pp. 564-578, 2016.

[12] Y.-J. Liu, B.-J. Li and Y.-K. Lai, "Sparse MDMO: Learning a discriminative feature for spontaneous micro-expression recognition", *IEEE Trans. Affect. Comput.*, Jul. 2018.

[13] J. See, M. H. Yap, J. Li, X. Hong and S.-J. Wang, "MEGC 2019—The second facial micro-expressions grand challenge", *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, pp. 1-5, May 2019.

[14] Li, Chengcheng, Zi Wang, and Hairong Qi. "An Efficient Pipeline for Pruning Convolutional Neural Networks." In 2020 *19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 907-912. IEEE, 2020.

[15] Wang, Zi. "Zero-Shot Knowledge Distillation from a Decision-Based Black-Box Model." arXiv preprint arXiv:2106.03310 (2021).

[16] He, Yang, Guoliang Kang, Xuanyi Dong, Yanwei Fu, and Yi Yang. "Soft filter pruning for accelerating deep convolutional neural networks." arXiv preprint arXiv:1808.06866 (2018).

[17] Wang, Zi. "Data-Free Knowledge Distillation with Soft Targeted Transfer Set Synthesis." arXiv preprint arXiv:2104.04868 (2021).


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2022 at 05:03:06 UTC from IEEE Xplore. Restrictions apply.

Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Fake News Detection in Different Applications: State-of-the-art and Future Directions

**Aman Jolly** ( aman.jolly1994@gmail.com )
Delhi Technological University
**Shailender Kumar**
Delhi Technological University

**Research Article**

**Keywords:** Deep learning model, Fake news detection, Reinforcement learning, Ensemble learning, Semi-supervised learning, Supervised learning, technique

**Posted Date:** September 7th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-2024922/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

# Abstract

Social media platforms such as Telegram, Twitter, WhatsApp, and Facebook have grown indispensable for information exchange and delivery in recent years. Humans have a tendency to trust any piece of content that circulates on social media without questioning its authenticity or origin. Fascinated people sometimes spread fake news on social media because of its convenience for information sharing. Additionally, it is used for personal advantages, such as amending government policies and defaming eminent personalities. Hence, to avoid the fatal outcome of spreading fake news, different methods are developed for its identification that achieves high accuracy. This paper reviews conventional artificial news detection methods to solve the mentioned issue. The Systematic Literature Review, often known as SLR, is a procedure that assesses review papers connected to detecting fake news. A technique known as unsupervised learning was used for this article review, along with ensemble learning, semi-supervised learning, reinforcement learning, and supervised learning. This methodology's effectiveness is evaluated regarding the precision of the measurements taken. In conclusion, the deep learning model generates practical outputs and outperforms classical machine learning models.


# 1. Introduction

Researchers have recently studied fake news detection (FND) and its consequences for people. [1]. To classify false news as any content designed to convince people to rely on erroneous facts, we must first define what constitutes misleading news. Fake news is detrimental to society and individuals in general. [2]. For starters, fraudulent messages might disrupt or undermine the message biology system's validity balance. Individuals are forced to tolerate erroneous or biased beliefs as a result of fake news. Fake news or proselytizers are frequently used to spread political messages or influence people. Fake news has a long-term impact on how people interact with and respond to genuine information. Building a system that can detect false news on social media quickly decreases the impact of fake news. ([3] [4]). Identifying bogus news at various social levels is difficult [5]. The categorization of selection targets seen in this manner, initiating or modifying specific messages or information within an individual organization, includes a well-informed source for understanding the actual expectation or significance of the information communicated and determining whether it is genuine or fake. Approve it to proceed with the decision of whether it is genuine or counterfeit. Artificial Intelligence (AI) has undergone significant changes in recent years [6] in data advances and designs, such as natural language processing, robotic surgery, virtual personal assistants, developing and using efficient transportation systems, and possibly the most significant impact on our lives [7]. Nonetheless, new technologies and features can be used to erase bogus news for the greater good by developing individualized data and making it more enticing in the phony crusades. [8]. As a result of the demand for dependable and trustworthy data, individuals generally do not bow to the ambitions of those interested in controlling reality. A few academics have argued that libertarian legislators use false news to undermine power. The current study, for example, suggests that false news damages previously trustworthy data sources. by mainstream scholars or in specialized sectors such as reporting. Fake news can be used to spread disinformation or propaganda

[9]. However, it usually impacts the general public's feelings, normal reactions, investigation, and data from some sources, raising annoyance and shock and, without a doubt, causing paranoia. Opinions and harsh one-sided content harm government-managed pensions. Then, to combat fake news, researchers employ techniques such as text characterization, network analysis, polls, and evaluations [10].

### 1.1. The study's motivation:

Customers are irritated by allegations of wholesale fraud, lack of validity and classification, bogus profiles, and so on. Thus, there is a need to govern the misleading material that penetrates the online world. False information spread by online entertainment damages people's reputations and affects society's faith in the news ecosystem, jeopardizing cultural stability. It is difficult to distinguish false news since the terminology used in reports is remarkably comparable to those used in actual news. [11] As a result of faulty message identification,

### 1.2. Challenges:

The most difficult component of distinguishing bogus news from true news is determining whether a piece of data is founded on facts. Reality is a fundamental concept that incorporates what happened previously and eventually to someone or something. Recognizing the importance of data necessitates, on all counts, a desire for efficiency, assuming that computers are given entire control over what data is conveyed to whom, when, and via which channel. Notably, many virtual entertainment pieces follow the basic premise of portraying something. As a result, establishing an editorial standards center is critical. [12].

## 2. Fake News Detection (Fnd)

Fake news is information that is designed to mislead readers purposefully. Meaning and authenticity are the two most important aspects of this notion. Fake news has two characteristics: it contains provable inaccuracies from the interview, and it is provided with the goal of fooling the audience. Misleading material disseminated through Internet-based entertainment can have serious consequences, such as undermining public trust in the news ecosystem, causing harm to a person or organization, or instilling fear in the general population, all of which can have an impact on the overall stability of society [13].

**Table 1:** Fake news categories

| S. No | Categories | Description |
|---|---|---|
| 1 | Fabricated material | Novel content that is designed to be offensive and deceptive |
| 2 | Imposter content | If the actual source can be impersonated |
| 3 | Misleading content | Utilizing data to assist clarify an issue |
| 4 | Satire | It contains the capability to deceive and has no malicious purpose |
| 5 | Manipulated content | Original images or information which can be tampered with to mislead the public |
| 6 | False content | The authentic material can be connected by erroneous metadata |
| 7 | False connection | The data cannot be supported by captions, images, and headlines |

Information can be represented as an array of tuples containing the headers and text of a specified number of messages with $A = \{(a^H, a^T)\}$. The proof-of-concept challenge of differentiating false news establishes if a piece of data is fake or not. Techniques used to curate data separate real news from fake news. On the other hand, news media may use subtle strategies, for example, to force the customer to accept something that production realities do not. By all accounts, it is also conceivable to move material from legitimate sources, although the evidence is lacking. In addition, misleading features of fake news employ altered content. Fake news uses fake stories, images, and context to mislead readers. [14].

### 2.1. Various News Data Types

Different types of information, including in the following section, are used to build intelligent narratives. People usually get their news in one of four ways, all of which are consistent with:

- **Text:** Language is employed to assess the meaning of a series of texts, focusing on the cooperative message. Because language is more than words and expressions, grammar, tone, and etiquette are all used to their fullest extent in the speaking exam.
- **Multimedia:** With the use of media, it is possible to combine music, images, videos, and designs into a single project. Because of how it is presented visually, the audience is immediately drawn to it.

### (c) Hyperlinks/Embedded content:

Clients can interface to various origins using hypertext links, and the reason for a news story can be used to gain users' Confidence. Writers use virtual entertainment credits to synthesize a preview of significant web-based entertainment posts such as YouTube videos, Facebook, tweets, and sound cloud cuts [15].

**(d) Audio:** Sound is one of the 221 elements of the Interactive Media Classification, an independent mechanism for news origins. This mediator considers various media, such as radio management, digital

recordings, and broadcasting companies, which are used to convey messages to a larger audience.

### 2.2. Categories of fake news detection

Three types differentiate fake news.:

- **Fabrication:** Fabricated news typically involves deliberately omitting information from a different source. The source is rationally aware of the story's error. The creation of fabricated reports depends on inaccurate content [16].
- **Tactic:** This advertisement uses sophisticated misdirection tactics to deceive the general public. Various outlets are spreading fake news. Few people can imagine believing the story to be true. Such information can be tracked to different places like false information about Donald Trump, through political competition on various Internet-based entertainment platforms, including Facebook, Twitter, and online magazines, so bound by the extensive population to believe it;
- **Parody:** Fake news is introduced as entertaining by a source. Because of the irony, it makes people unfamiliar with the beginning of the subject. Some individuals may mistakenly believe it to be valid [17].

The approaches for identifying false news are reviewed in this article using the SLR method. The following is a detailed description of the SLR.

## 3. Systematic Literature Review Method

Various false news expectations have recently been proposed in light of their preconditions. Information classifications are used to evaluate artificial news predictions. The results are separated by application and identified using graphical and factual depictions. This false news identification is powered by a particular written audit method, such as an exam with a survey and a quantitative and subjective research plan. This Work is also organized under the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [18] statement. PRISMA is used to conduct a comprehensive literature review on fake news anticipation. Figure 1 depicts the PRISMA flow chart.

The technique of identifying fake news is fundamental in industry and society. A formal letter survey is a basis for examining new forms and models, recognizing and closing in on future change identification goals. A provision for banning or adding collected papers should be introduced. General considerations and constraint conditions and their information are presented in Table 2.

**Table 2:** Inclusion and Exclusion condition

| S. No | Factors | Condition | Condition explanation |
|---|---|---|---|
| 1 | Inclusion | Closely Related (CR) | The research is entirely related to fake news prediction |
| | | Partially Related (PR) | To descript the challenges, trends, and issues related to fake news, that paper intends to describe the detection |
| | | | Research related to fake news detection without mentioned applications |
| | | | Fake news detection-based review and survey paper |
| 2 | Exclusion | Loosely Related | Fake news prediction only presented in keywords or references |
| | | | Fake news prediction is only presented in cited expression |
| | | | Fake news prediction is only presented in the future requirement, future perspective, and future research direction |
| | | | Fake news prediction is only presented as an example manner |
| | | Non-Related (NR) | The definition of fake news prediction is not coordinated to different applications |
| | | | A paper is not an academic paper. For example, conference reviews, forewords, content, and editorial materials |
| | | Without full text (WF) | A work of writing that, when read in its entirety, cannot be judged fairly. |
| | | | The only parts of a piece of writing that are available in English are the keywords, the abstract, and the title. |
| | | Search Engine Reason (SER) | A paper that cannot be evaluated based on its full text. |
| | | | The only parts of a piece of writing that are available in English are the keywords, the abstract, and the title. |

In formal written auditing, three major auditing standards are seen as introduced,

### 3.1. Data collection with evidence:

To gather the information that should include summary results, the key summary is the prediction of fake news, and the data set will include individual support text clarifications of the study [19]. Figure 3 depicts the logical learning of documents.

### 3.2. Objective review method:

At least two examiners should review the gathered papers to determine whether they should be excluded or included. Two reviewers should verify the summarized information to determine the suitability of the piece. On two occasions, two inspectors could not agree, and a third inspector made the final decision. Because of the censorship system's assessment and avoidance, fake news expectations are divorced from their future carriers and suggestions [20]. The PRISMA flowchart includes a fake news prediction that was used to analyze phony news forecasts. The survey cycle is completed by considering four cycles, which are as follows:

- Identification
- Screening
- Eligibility
- Included

### 3.3. Identification:

This collection includes study and exam papers gathered from an open-source framework. To manage a comprehensive set of exam papers, catchphrases such as "Fake News Prediction Using AI," "Fake News Prediction Using Deep Learning," "Fake News Prediction," "Fake News Prediction," and "Applications of Fake News Prediction" was employed. For example, in six studies, for example, the efficient pursuit relates to "IEEE Investigation," "Science Direct," "Scopus," "Meeting," "Studio," and "Web of Science." Furthermore, these materials were gathered between the years 2003 and 2022. The collected papers are in various settings. They should, for example, be written in English and circulated through diaries, books, or websites. Personal, topical, or observational language should be used to introduce search subjects. With these circumstances and rumors in mind, 250 papers have been gathered.

### 3.4. Screening:

When gathering documents with inferred notes and journals, the articles are screened using an appropriate technique based on keywords. Figure 2 explains the collection sheet in detail. The documents gathered are being utilized to analyze the design of fake news systems. In light of their approval, the screening approach removed unsuitable articles from the collected records. In addition, 75 documents were removed from the system.

### 3.5. Eligibility:

If unwanted documents are removed from the gathered documents, the remaining copies will be sent after the eligibility status has been verified. Articles are picked for audit by checking this eligibility level. Table 2 [21] clearly states the exclusion and consideration criteria for determining eligibility.

### 3.6. Included:

In the light of Table 2, documents are selected for approval after checking the eligibility status. Quantitative and subjective effects associated with fake news recognition are examined in this section. The whole cycle is completed; 175 papers are selected to explore the results from the position of the fake news anticipation method.

# 4. Fake News Detection Techniques Related To Review Work

This section classifies fake news scientifically. We use deep learning and AI to group fake news. The fake message detection procedures are introduced in Figure 4. The identification process of fake messages is described in Figure 5.

## 4.1. Supervised Learning

In supervised learning, iterative and characteristic-based problems are frequently used. The model is constructed from a collection of product-specific data, with each piece of data programmed to yield an acceptable result. When trained on a preset dataset, the model performs best. It is possible to address regression-based challenges by producing predictions of stable or actual properties using aspects of easily available information. Surprisingly, the categorization totally isolates each item of data from the perspective of its class labels [22].

### Support Vector Machine

The majority voting method was suggested by Dharmaraj R. Patil et al. [23] as a way to identify fake news. They examined many text-based traits of both genuine and fake messages in that paper. Additionally, a false news dataset that was openly accessible was used. It had 20,800 false messages, 20,800 real messages, and 10,413 false messages that were labelled with coordinates 0 and 1. The proposed method was tested using the AI classifiers Naive Bayes, SGD, SVM, AdaBoost, Extra Trees, Random Forest, XGBoost, Logistic Regression, and Decision Tree. A few SVM-based fake news detection systems are examined in Table 3.

**Table 3:** SVM-based fake news detection techniques

| S. No | Author | Dataset | Accuracy | Methodology |
|---|---|---|---|---|
| 1 | Yazdi *et al.* [24] | LIAR, BS detector, Buzzfeed news | 94.19% | K means clustering with SVM |
| 2 | Presidio *et al.* [25] | Self-aggregated dataset | 82% | SVM with cleaning, stop word removal, and tokenization |
| 3 | Ahmed *et al.,* [26] | LIAR dataset | 87% | n-gram, TF-IDF vectorizer, stop word removal, stemming with SVM classifier |
| 4 | Suchitra desolate *et al.,* [27] | Buzzfeed | 87.154% | SVM with credibility analysis |
| 5 | Ahmed *et al.,* [28] | Dataset collected from LIAR | 83% | SVM with TF-IDF |
| 6 | Agarwal *et al.,* [29] | LIAR | 61% | SVM with a systematic approach |

Machine learning computations and evidence for distinguishing bogus news have been introduced by Sajjad Ahmed et al. [30]. Three classifiers were used: Nave Bayes, Passive Aggressive, and Support Vector Machine. At the location of bogus news, the primary categorization is not totally accurate. By combining AI and text-based modification, we can construct classifiers that can recognize fake news and arrange news material. Text characterization removes individual text fragments and groups them together. Sanaa elyassami et al. [31] presented a research machine learning model that determines if messages are fraudulent or real. Five classifiers were built using Naive Bayes techniques, Logistic Regression, Gradient Boosting, Support Vector Machine, and Random Forest. The models were built using integrated open-source datasets gathered from various online sources and covering various geographies. To provide fake newsgroup models and extract meaningful information from news items, text lemmatization, vectorization, and tokenization were used. The impact of the democratic system on the display of group learning models was studied. The five classifiers' exhibition accuracy was assessed using F1-score, precision, and recall. Asha J et al. [32] reported n-gram analysis and machine learning algorithms to detect bogus news. The evaluation and interaction of Term Frequency-Inverted Document Frequency (TF-IDF) and Term Frequency (TF) with six machine learning computations were carried out. These calculations made use of Logistic Regression (LR), Decision Trees (DT), Stochastic Gradient Descent (SGD), K-Nearest Neighbor (KNN), and Support Vector Machines (SVM).

### *Convolutional Neural Network*

Mohammad HadiGoldani et al. [33] used Convolutional Neural Networks (CNN) to present several randomness and establishment models for false news recognition. We distinguish between fixed and non-fixed word embeddings, which allow us to assimilate continuously prepared and renewed words during the manufacturing phase. Our proposed approaches are evaluated using two recent large datasets

from the field: ISOT and LIAR. The best designs speed up processing by 7.9% on the ISOT and 2.1% on the LIAR dataset test set compared to traditional methods.

Chenggang Song et al. [34] suggested a crossmodal attention residual and multichannel convolutional neural network-based multimodal fake message detection system (CARMEN). A Crossmodal Attention Residual Network (CARN) can accurately remove vital data linked to a distinct source modality while following one type of data from the accurate modality. A multichannel convolutional neural network (MCN) can pull out literary element representations from individual and combined text-based data, overcoming the effect of the crossmodal ensemble's fusion data.

Hager Saleh et al. [35] present an improved convolutional neural network model for detecting bogus news (OPCNN-FAKE). This section compares the OPCNN-FAKE evaluation to that of the Long Short-Term Memory (LSM), the Recurrent Neural Network (RNN), and the (LSTM). The four databases containing bogus news were taken into account while using the six classic machine learning algorithms: Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), K Nearest Neighbour (KNN), logistic regression (LR), and Decision Tree (DT). The parameters of DL and ML were chosen using hyperope optimization, and grid search optimization approaches, respectively. TF-IDF and N-gram were also used to extract essential properties from common datasets for traditional machine learning. At the same time, glove word embedding is used to define features in the feature matrix structure for DL models.

### *Decision Tree*

Noman Islam et al. [36] proposed an intelligent arrangement by recognizing the authenticity of information using traditional language manipulation processes. This research, in particular, provides a novel approach with three stages: status discovery, creator validity confirmation, and an AI-based sequence to validate the news's integrity. Some AI processes, such as SVM, logistic regression, random forest, and decision trees, were used in the pipeline's final phase.

Shikun Lyu et al. [37] used AI computations to investigate the exhibition, including Fake News Tracker, doc2vec, SVM, and decision trees. The findings of this research show that SVM and decision trees effectively distinguish bogus news with 95% accuracy. Generally, the decision tree technique outperforms SVM. Topics for further research are preferred. Figure 6 depicts an analysis of this guided learning. In terms of accuracy, the CNN model works well from the point of view of selection.

### *Logistic regression*

N. Leela Siva Rama Krishna et al. [38] used a support vector machine technique to examine the text property and incorporated a logistic regression. For accurate false news detection, contrast text-based property correctness and SVM calculation with logistic regression (LR). Methodology and resources: The test for fake news identification in the proposed study was performed using AI calculations such as LR and SVM calculations, with G powers of 80% and 0.05, respectively (N=311). O. V. Chandra Shekhar Rao et al. [39] differentiate bogus news, offer a model of events reversed, and a logistic regression classifier.

This classifier performed admirably. According to previous research, when classifying organizations. This method used TF-IDF highlight to increase the false message model's accuracy. Regarding the article's news content, this program tries to identify fake news using NLP and machine learning. Python Flask was used to transform the model into a web interface after constructing an appropriate machine learning model to distinguish between real and fake communications. Table 4 compares supervised learning analysis to traditional false news detection strategies.

**Table 4:** Few papers related to supervised learning

| S. No | Author | Dataset | Accuracy | Methodology |
|---|---|---|---|---|
| 1 | Julio *et al.,* [40] | Buzzfeed News articles | Naive Bayes, SVM, and KNN | 89% |
| 2 | Abdullah *et al.,* [41] | 2010 earthquake dataset, which contains 20360 data | RNN, LSTM, Logistic Regression, Naive Bayes and SVM | 89.34% |
| 3 | Supanya *et al.,* [42] | Twitter API data which contains 948373 messages | Naive Bayes, neural network, SVM | 99.90% |
| 4 | Jain *et al.,* [43] | RSS feeds and different news websites | NLP language, SVM, and Naive Bayes | 93.50% |
| 5 | Agarwal *et al.,* [44] | Kaggle.com databases of news articles | LSTM, CNN, Naive Bayes | 94% |
| 6 | Ahmed *et al.,* [45] | ISOT dataset, news articles from reuters.com and kaggle.com | Linear SVM and unigram features | 92% |

### *Naive Bayes*

Shafayat Bin Shabbir Mugdha et al. [46] To detect bogus news in Bengali, the Gaussian Naive Bayesian Classifier was introduced. The proposed method employs a data set generated in an unusual fashion via causal links, as well as a Gaussian Naive Bayes algorithm. In the computation, both text characteristics based on TF-IDF and an extra tree classifier are used. They have also conducted substantial studies on various AI computations and characteristics. Parita Jain et al. [47] proposed a new method for determining if the collected messages are genuine or not. This was accomplished through the use of a technique known as Natural Language Processing (NLP). Various systems, such as text clustering, characterization, and so on, were also applied, and the outcomes were thoroughly investigated. Information was gathered from many sources, and several algorithms, such as SVM, Naive Bayes, and LSTM, were utilized to determine whether or not the message was correct. To detect bogus news in Bengali, the Gaussian Naive Bayesian Classifier was introduced. The proposed method employs a data set generated in an unusual fashion via causal links, as well as a Gaussian Naive Bayes algorithm. In the

computation, both text characteristics based on TF-IDF and an extra tree classifier are used. They have also conducted substantial studies on various AI computations and characteristics. Parita Jain et al. [47] proposed a new method for determining if the collected messages are genuine or not. This was accomplished through the use of a technique known as Natural Language Processing (NLP). Various systems, such as text clustering, characterization, and so on, were also applied, and the outcomes were thoroughly investigated. Information was gathered from many sources, and several algorithms, such as SVM, Naive Bayes, and LSTM, were utilized to determine whether or not the message was correct.

### 4.2. Semi-supervised Learning

One significant disadvantage of supervised learning is that it requires labelled data, which can be difficult and time-consuming to obtain. The main disadvantage of unsupervised learning is that it can only be utilized for a limited range of applications, such as customer segmentation, personality analysis, recommender systems, and so on. To address these issues, a type of artificial intelligence called semi-supervised learning employs a considerable amount of unlabeled data and a limited amount of labelled data. The following are the goals of semi-supervised learning: The first strategy, known as "inductive semi-supervised learning," aims to predict the next batch of experimental data. The second method is transudative semi-supervised learning, which attempts to identify the identity of anonymous data. Semi-dynamic learning models employ skepticism when dealing with unlabeled data. To help identify fake news, Rohit Kumar Kaliyar et al. [48] have implemented restricted rooms within a specific organization as well as the content of news stories. When combined with classic AI models, standard factorization techniques are insufficient and ineffectual for detecting fake news. It is necessary to plan a robust deep learning model utilizing the tensor factorization method. Using a collaborative network tensor factorization algorithm, we concatenate message content with a tensor to generate a passive representation of message content and friendship structure.

### 4.3. Unsupervised Learning

The essential distinction the use of an unlabeled training data set distinguishes supervised learning from unsupervised learning (with unsupervised learning, classes are not assigned values). There is no information available about the crucial system. NN, DBN, and other sophisticated unsupervised learning algorithms are included. It is feasible to uncover important and valuable clusters within unlabeled input data using these learning techniques, which can then be utilized to classify systems. This section will discuss some of the most innovative and recent unsupervised learning methods for classifying fake news. Despoina Mouratidis et al. [49] presented a novel paired text input method. This innovative deep neural network learning architecture enables variable input fusion across network layers. On Twitter, several input methods, such as word embeddings and linguistic and network accounting aspects, detect bogus posts automatically. A full test setup also runs classification tests on news items and topics generated inductively from tweets.

Capsule neural networks have been shown to be successful at spotting fake news by Mohammad Hadi Goldani et al. [50]. In this instance, various embedding models should be used for messages with varying

durations. Fixed word embeddings are utilized for news pieces that are lengthier or more in-depth. Selectable word embeddings, on the other hand, enable progressive training and the incorporation of changes during the training process.

Shiwen Ni et al. [51] identify bogus news and defend its proliferation on social media. A novel neural network-based method known as Multi-View Attention Networks (MVAN) has been presented. By adding the textual semantic emphasis and propagation structural focus from the MVAN model, our model may derive clues and information from both the content of the initial tweet and how it spreads. Furthermore, the model's two attentional processes may detect suspicious people in a campaign context as well as crucial cue phrases in false news texts.

Vidit Jain *et al.* [52] have presented a straightforward network design combining context embedding, word embedding, and focus techniques with available relevant metadata that was used on several real-world datasets to demonstrate the effectiveness and efficiency of these models. On the LIAR dataset, the model achieved an accuracy of 46.36 %, outperforming the state-of-the-art by 1.49 %. The unsupervised learning-based techniques are given in table 5. The different method of supervised training is analyzed and illustrated in figure 7 Additionaly, the conventional techniques of DAGA-NN, Deep contextualized learning, hybrid deep learning, simple network architecture, MVAN, and capsule neural network is achieved the 78%, 72%, 88%, 81%, 72%, and 68% respectively. In figure 7, the Temporal propagation technique is achieved efficient outcomes of 90% accuracy.

**Table 5:** Few papers related to unsupervised learning

| S. No | Author | Dataset | Accuracy | Methodology |
|---|---|---|---|---|
| **1** | Fernandez *et al.,* [53] | LIAR | 48.5% | stacked CNN with ANN |
| **2** | Wang *et al.,* [54] | Weibo and Twitter | 82.7% | Event Adversarial Neural Network (EANN) |
| **3** | Zhang *et al.,* [55] | Twitter account details | 68.12 | Deep diffusive network |
| **4** | Singhania *et al.,* [56] | Self-aggregated dataset | 96.24% | 3HAN |
| **5** | Ruchansky *et al.,* [57] | Weibo and twitter | 95.3% | RNN |
| **6** | Chopra *et al.,* [58] | FNC-dataset | 85.07 | NN-LSTM |

Jamal Abdul Nasir *et al.* [59] have introduced a hybrid deep learning model that combines recurrent and convolutional neural networks to classify fake news. The model was successfully evaluated on two fake news datasets (ISO and FA-KES), and the detection results were noticeably superior to those of existing

non-hybrid methods. It would be encouraging to see results from more tests of the proposed model's generalizability across different datasets.

Mohammadreza Samadi *et al.* [60] presented the input messages embedded using three classifiers previously trained using various models. To benefit from the deep contextual representation and deep neural classification that these models offer, combine single-layer perceptron (SLP), multi-layer perceptron (MLP), and convolutional neural networks (CNN) after the embedding layer. Pre-trained models like Funnel Transformer, BERT, RoBERTa, and GPT2.

Chenggang Song *et al.* [61] have presented a new paradigm for detecting fake news based on temporal propagation that integrates structure, content semantics, and data. In particular, the model can simulate real-world news's temporal growth patterns into a continuous-time graph. A dynamic diffusion network environment.

Hua Yuan *et al.* [62] have presented to solve the problem, the "Domain-adversarial and graph-attention neural network" (DAGA-NN) model was used. Its key benefit is the ability to train a model with few (or no) samples in a text environment with numerous instances or domains using only partial domain sample data. Fake message detection tasks are caused by cross-domain identification or message content evolution.

### 4.4. Reinforcement Learning

RL seeks to maximize reward versus supervised learning and unsupervised learning. Algorithms make better choices to connect to the best available route or path. These choices are based on the behaviors required to engage with the environment. Simply put, reinforcement learning uses feedback from the system at each step to get better results. The performance of several RL algorithms, such as Markov Decision Process (MDP) Feedback, can improve LSTM, Q-Learning, and RNN. Reinforcement learning helps classify fake news. This section presents the latest reinforcement learning approaches for fake news classification. The reinforcement learning model is analyzed and illustrated in figure 8. RNN and deep structured semantic technique, BDANN, FNDNet, and deep learning model achieved 68%, 58%,55%, and 68% accuracy. Based on the analysis, Bi-directional LSTM-RNN attained 72% accuracy.

### RNN- LSTM

Pritika Bahad *et al.* [63] have presented a Recurrent Neural Network with Bi-directional LSTMs for Predicting Fake News.

Two datasets of unstructured news articles were used to evaluate the model. Bidirectional LSTM outperforms CNN, simple RNN, and unidirectional LSTM in spotting fake news.

Shrutika S. Jadhav *et al.* [64] They have used a deep structured semantic model and enhanced recurrent neural networks to create a system capable of recognizing and classifying content associated with fake news. The proposed method identifies critical characteristics of fake news without requiring prior

knowledge and achieves a 99 percent accuracy. The proposed method for analyzing the performance of a system is based on three things: sensitivity, accuracy, and specificity.

Tong Zhang *et al.* [65] have presented BERT-Based Domain Adaptive Neural Network Multi-Model Fake Message Detection as a Complete Model (BDANN). Multi-model feature extraction, a domain classifier, and a fake message detector are the three primary modules that make up BDANN. The pre-trained BERT model is used to extract text features by multi-model feature extraction, and the pre-trained VGG-19 model is used to extract image features. The features synthesized from the received components are sent to the detector to detect fake messages. The primary responsibility of a domain classifier is to convert multiple model features of different instances into a common feature space.

Rohit Kumar aliyar *et al.* [66] found misleading information using a deep convolutional neural network (FNDNet). This model (FNDNet) was developed to automatically learn discriminative properties for classifying fake news through multiple hidden layers built on a deep neural network. Instead of being based on manually generated features, this model uses a neural network to construct its model. In this scenario, a deep convolutional neural network (CNN) was utilized to extract various features at each layer.

Yin-Fu Huang *et al.* [67] presented a system for identifying fake news based on deep learning. To begin with, several different training models are utilized to preprocess and analyze the messages. The next step is constructing a proposed ensemble learning model for identifying fake news by merging four distinct models. These models are the LIWC CNN, the depth LSTM, the N-gram CNN, and the embedding LSTM. The Self-Adaptive Harmony Search (SAHS) method was used to find the best weights for the ensemble learning model. This helped make it even easier to spot fake news.

### 4.5. Ensemble Learning

Human culture has long practiced functioning as a democratic society and making decisions based on the opinions of many people or experts. Over the last few decades, academic interest in machine learning and artificial intelligence in multi-classifier systems, also known as ensemble learning techniques, has grown. The emphasis on ensemble systems is more than adequate because they may have such enormous beneficial effects. It should also be emphasized that ensemble systems have numerous practical applications in a variety of sectors. Several studies, both theoretical and experimental, have proven that ensemble approaches produce more accurate findings than single model approaches.

Vijay Srinivas Tida et al. [68] effectively provided an integrative approach for identifying bogus news samples by merging publically available datasets. This study made use of three publicly available datasets, including ISOT and others from the Kaggle website. The model was created using Google's Bidirectional Encoding Representation of Transformers (BERT) basic uncase model with pre-trained weights that were not changed during training or in preprocessing processes, such as deleting words with lengths shorter than three. To create the final model, the pre-trained Google BERT base uncased model for each dataset is tweaked, and the model with the greatest performance across all three datasets is chosen. Both message content and fake news detection have been presented by Rohit Kumar Kaliyar et

al. [69]. A social network considers the echo chambers of individuals who have similar opinions on social media. By combining information from the news, users, and communities, a tensor expressing social context was created. The association between social media user profiles and news stories is known as social context. Following the tensor and message content combination, coupled matrix-tensor factorization was used to accurately reflect both the message content and the social environment. To valid ate the suggested methodology, BuzzFeed data was used as a test dataset. The factors revealed during decomposition have been utilized to classify messages as features. The Deepfake deep neural network model and the XGBoost ensemble machine learning classifier were utilized in the investigation. Work on classification.

Figure 9 depicts and analyses the ensemble procedures. The BERT model, CNN, LSTM, and CNN achieved 75%, 81%, and 91% accuracy, respectively, in the analysis. According to the findings, the CNN model has a high level of accuracy.

Muhammad Umer et al. [70] demonstrated the use of PCA and chi-square two-dimension reduction methods in a CNN-LSTM hybrid neural network. This Work lowered the dimensions of the feature vectors before passing them to the classifier. The FNC website provided a dataset divided into four categories: agree, disagree, discuss, and unrelated. Sachin Kumar and colleagues [71] created a deep learning algorithm to detect bogus news.

To generate correct and false news databases, this study collected 1356 news occurrences from Twitter and PolitiFact. In this study, CNNs, LSTMs, ensemble approaches, and attention mechanisms are used. Our investigation shows that the attention-grabbing technique using CNN + bilateral LSTM ensemble network has an 88.78% success rate.

# 5. Performance Evaluation

Using a dataset of authentic and fraudulent articles, evaluate the most cutting-edge algorithms for detecting false news. It contains 7796 recorded extracts of diverse situations such as press releases, campaign speeches, and radio or television shows. Each report has an authenticity label, a title, and a context annotation. Eighty percent of the dataset is utilized for training the model, while the remaining twenty percent is used for testing. The dataset is divided into two parts: training and testing. To validate false news identification, three databases are considered: LIAR [72], fake news datasets [73], and corpus-based datasets [74]. Figure 11 depicts some examples of word-based databases.

According to the findings, the deep learning model produced efficient results in terms of accuracy. The main methodologies are used to study three different types of databases. Figure 12 depicts the exploration and illustration of liar databases for various machine learning applications. According to the findings, CNN has been used to detect fake news. Similarly, the k-NN employs a 10% percentage ratio, whereas the LSTM employs a 10% proportion ratio: Bi-LSTM is used 10% of the time; Naive Bayes is used 11% of the time; Adaboost is used 10% of the time; Decision tree is used 9% of the time; Logistic regression is used 10% of the time; SVM is used 10% of the time, and GAN is used 10% of the time. Many

articles use naive Bayes to tackle fake news detection through the liar database. Figure 13 depicts an analysis of corpus databases for various machine learning applications. According to the findings, CNN used an 11% proportion ratio to detect bogus news. Similarly, k-NN is used 9% of the time; LSTM is used 11% of the time; Bi-LSTM is used 12% of the time; Naive Bayes is used 11% of the time; Adaboost is achieved 9% of the time; Decision tree is achieved 8% of the time; Logistic regression is performed 9% of the time; SVM is used 9% of the time; GAN is used 11% of the time. Many articles employ the Bi-LSTM to tackle fake news identification using a corpus database. Figure 14 depicts an analysis of fake news databases for various machine learning techniques. According to the findings, CNN used an 11% proportion ratio to detect bogus news. Similarly, k-NN is used 9% of the time; LSTM is used 11% of the time; Bi-LSTM is used 12% of the time; Naive Bayes is used 11% of the time; Adaboost is used 10% of the time; A decision tree is used 9% of the time; logistic regression is used 9% of the time; SVM is used 9% of the time, and GAN is used 11% of the time. Many papers use CNN, GAN, Naive Bayes, and Bi-LSTM to detect fake news in a corpus database.

Either the users themselves or the structure of the network should form the basis of an intervention strategy for fake news. From the perspective of the consumer, the efficacy of interventions to combat fake news is contingent on the unique roles people play in the propagation of false information [87] .

i. Powerful spreaders of fake news on a social media platform, if they are stopped, will result in a more effective intervention rather than focusing on individuals who have little or no social impact.
ii. Fact-checker (one who intervenes to stop the spread of false information on social media) by including links in their posts or comments.
iii. Adversarial user (one who is against the given opinion) shall be identified and be penalized if found guilty of spreading fake news; ( In contrast to the manual approach of user-based fake news intervention, network-based fake news intervention is automatic and blocks the propagation paths of fake news in an effort to stop it from spreading. In other words, network-based fake news intervention is more effective than user-based fake news intervention. It accomplishes this by analyzing the network structure of the distribution of bogus news and making predictions regarding how it will spread in the future.

**Table 6:** Study of the State-of-the-Art in Empirical social media Fake News Detection

| References | Input Data | Methodology's | Results |
|---|---|---|---|
| Castillo et al. [76] | An event hosted on Twitter | Using a comprehensive feature set, a decision tree-based model was developed to predict the reliability of news sources. | A level of accuracy of 86% is reached by the supervised classifier. |
| Rubin et al. [77] | RST analyzed news transcripts | The SVM classifier is used with RST and VSM to classify news as true or fraudulent. | A level of accuracy of 86% is reached by the supervised classifier. |
| Ferreira et al. [78] | Emergent dataset rumors/news articles | Multiclass logistic regression model for stance categorization | The achieved accuracy of 73% was 26% higher than that of EOP. |
| Janze et al. [79] | BuzzFeed's dataset of verified Facebook posts | Developed a paradigm for exploratory research based on hybrid methodologies of ELM and UGC, making use of the visual, behavioral, cognitive, and emotional clues of the news that was shared on Facebook | More than 80 % of correct predictions can be made using the configurations that perform the best. |
| Wang et al. [80] | LIAR dataset short statements | Major baseline, regularized LR, SVM, Bi-LSTMs, and CNNs used on LIAR surface-level patterns. | CNNs beat all models on the holdout test set, achieving 0.270 accuracies. |
| Shu et al. [81] | Fake-News net articles | The TriFN framework was created to detect false news by modelling the interaction between news outlets and their readers in real-time. | TriFN has a 4.72 (BuzzFeed) and 5.84 (PolitiFact) accuracy advantage over LIWC and Castillo. |
| Lu et al. [82] | The Cascades shared from the Twitter15 and Twitter16 datasets | Developed a neural network-based Graph-aware Co Attention Networks (GCAN) model to detect bogus tweets by using the original tweet and the matching retweet sequence without any comments as inputs. | Outperformed all baselines with 87.67% accuracy on Twitter15 vs. 90.84 on the Twitter16 dataset. |
| Wang | Kaggle COVID-19 short sentences | Combining the Bidirectional Encoder Representations from | 93.47% accuracy |

| | | | |
|---|---|---|---|
| et al. [83] | | Transformers (BERT) model with BiLSTM and CNN layers on top, using both frozen and non-frozen parameters, we were able to detect a wide range of object types. Misinformation about COVID-19 | while testing BERT-related models |
| Jarrahi et al. [84] | Articles of news taken from the Fake News net dataset | UPFD framework to recognize false news combining content and graph modelling to collect user preferences | Accomplished a level of accuracy of 90.6% on the PolitiFact dataset and 97.8% on the Gossip cop dataset. |
| Sahoo et al. [85] | News articles from Fake News Net dataset | Using LSTM, Chrome can recognize bogus news on Facebook by evaluating the account's behavior. | An accuracy of 99.4 percent was reached by the deep learning algorithm that was proposed. |
| Seddari et al. [86] | Expressions took from BuzzFeed News's databases | A hybrid approach, which combines knowledge-based tactics with language-based ones, was proposed as a means of identifying bogus news. | A combined 94.4% accuracy is achieved by the suggested system when using both feature types. |

This is an area that is anticipated to require a significant amount of effort.

- Fake news is a global problem. Each country has social, political, and technical problems to counter fake news, which must be explored. A benchmark dataset addressing local or national instances should be created for fake news empirical research.
- Most existing datasets identify information as true or untrue. A more precise categorization of data by intent may help distinguish fake news from opinion and parody news.
- Machine learning and deep learning-based fake news detection approaches should be supplemented by other credible methods.
- The fake news detection study applies to rumor identification, spammer and bot detection, stance detection, clickbait detection, and truth discovery. These sections require future Work to fix concerns.
- Big false news data and reducing feature vector size are major subjects.
- There is no reliable work in visual feature extraction from photographs and videos and constructing a fake news identification system.

- The attention mechanism improves bogus news detection models. This area needs more development.

# 6. Conclusion

This Work comprehensively discusses the fundamental notion of spotting fake news, as well as its various varieties, features, and characteristics. Furthermore, a taxonomy for the false news detection model is offered here. Several approaches for detecting false information have been developed in order to identify user behaviour that contributes to the spread of rumours or disinformation. On three datasets: liars, fake news, and the corpus, standard machine learning approaches and deep learning algorithms were compared and contrasted. The findings of this study suggest that deep learning methods outperform typical machine learning methods. The findings of this study will be useful for future research into establishing new models or techniques for recognizing and detecting fake news. Another beneficial outcome is that the police department's cyber unit can utilize this research to urge the adoption of proper procedures and methods for dealing with bogus data, thus enhancing society. The analysis's most obvious weakness is that it only considers textual data. However, it can be developed in the future to examine both text and image data, yielding analytical findings for significantly larger and more diverse datasets.

# Declarations

### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the Work reported in this paper.


### Acknowledgments

The authors thank the anonymous reviewers for their helpful comments and suggestions.

### Funding

Not applicable


### Ethical approval

Not applicable.

### Consent to participate

Not applicable.

### Consent to publish

Not applicable.

### Conflict of interest

The authors declare no competing interests.

### Data Availability

No data availability

## References


1. Gaozhao, Dongfeng. "Flagging fake news on social media: An experimental study of media consumers' identification of fake news." Government Information Quarterly 38, no. 3 (2021): 101591.
2. Hopp, Toby. "Fake news self-efficacy, fake news identification, and content sharing on Facebook." Journal of Information Technology & Politics 19, no. 2 (2022): 229-252.
3. Deepak, S., and Bhadrachalam Chitturi. "Deep neural approach to Fake-News identification." Procedia Computer Science 167 (2020): 2236-2243.
4. Barakat, Karine Aoun, Amal Dabbous, and Abbas Tarhini. "An empirical approach to understanding users' fake news identification on social media." Online Information Review (2021).
5. Jadhav, Shrutika S., and Sudeep D. Thepade. "Fake news identification and classification using DSSM and improved recurrent neural network classifier." Applied Artificial Intelligence 33, no. 12 (2019): 1058-1068.
6. Bhatt, Gaurav, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. "On the benefit of combining neural, statistical and external features for fake news identification." arXiv preprint arXiv:1712.03935 (2017).
7. Anoop, K., P. Deepak, and V. L. Lajish. "Emotion cognizance improves health fake news identification." In IDEAS, vol. 2020, p. 24th. 2020.
8. Van Duyn, Emily, and Jessica Collier. "Priming and fake news: The effects of elite discourse on evaluations of news media." Mass Communication and Society 22, no. 1 (2019): 29-48.
9. de Oliveira, Nicollas R., Pedro S. Pisa, Martin Andreoni Lopez, Dianne Scherly V. de Medeiros, and Diogo MF Mattos. "Identifying fake news on social networks based on natural language processing: trends and challenges." Information 12, no. 1 (2021): 38.
10. Mrowca, Damian, Elias Wang, and Atli Kosson. "Stance detection for fake news identification." Eliaswang. Com (2017).
11. Trueman, Tina Esther, Ashok Kumar, P. Narayanasamy, and J. Vidya. "Attention-based C-BiLSTM for fake news detection." Applied Soft Computing 110 (2021): 107600.
12. Koloski, Boshko, Timen Stepišnik-Perdih, Senja Pollak, and Blaž Škrlj. "Identification of COVID-19 related fake news via neural stacking." In International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation, pp. 177-188. Springer, Cham, 2021.

13. Datta, Anisha, and Shukrity Si. "A supervised machine learning approach to fake news identification." In International Conference on Intelligent Data Communication Technologies and Internet of Things, pp. 197-204. Springer, Cham, 2019.
14. Paschen, Jeannette. "Investigating the emotional appeal of fake news using artificial intelligence and human contributions." Journal of Product & Brand Management (2019).
15. Anoop, K., P. Deepak, and V. Lajish. "Emotion cognizance improves fake news identification." arXiv preprint arXiv:1906.10365 (2019).
16. Paschalides, Demetris, Chorioallantois Christodoulou, Kalia Orphaned, Rafael Andreou, Alexandros Kornilakis, George Pallis, Marios D. Dimakakos, and Evangels Markets. "Check-It: A plugin for detecting fake news on the web." Online Social Networks and Media 25 (2021): 100156.
17. Albahar, Marwan. "A hybrid model for fake news detection: Leveraging news content and user comments in fake news." IET Information Security 15, no. 2 (2021): 169-177.
18. Park, Ho Young, Chong Hyun Suh, Sungmin Woo, Pyeong Hwa Kim, and Kyung Won Kim. "Quality reporting of systematic review and meta-analysis according to PRISMA 2020 guidelines: results from recently published papers in the Korean Journal of Radiology." Korean journal of radiology 23, no. 3 (2022): 355.
19. Awan, Tahir Mumtaz, Mahroz Aziz, Aruba Sharif, Tehreem Raza Ch, Taha Jasam, and Yusra Alvi. "Fake news during the pandemic times: A Systematic Literature Review using PRISMA." Open Information Science 6, no. 1 (2022): 49-60.
20. Johore, Russel, and Marita Turpin. "How to Identify Fake News on Social Media: A Systematic Literature Review." In World Conference on Information Systems and Technologies, pp. 102-111. Springer, Cham, 2022.
21. Hosseini, Mahamat Ali, Deji Chen, and Xiao Yang. "The Application of Blockchain in Social Media: A Systematic Literature Review." Applied Sciences 12, no. 13 (2022): 6567.
22. Mahmoud, Haitham, Wenyan Wu, and Mohamed Medhat Gaber. "A Time-Series Self-Supervised Learning Approach to Detection of Cyber-physical Attacks in Water Distribution Systems." Energies 15, no. 3 (2022): 914.
23. Patil, Dharmaraj R. "Fake News Detection Using Majority Voting Technique." arXiv preprint arXiv:2203.09936 (2022).
24. K. M. Yazdi, A. M. Yazdi, S. Khadafi, J. Hou, W. Zhou, and. Sandy, "Improving fake news detection using k-means and support vector machine approaches," Int. J. Electron. Commun. Eng., vol. 14, o. 2, pp. 38–42, 2020.
25. A. B. Presidio, R. R. Isnanto, D. Eridian, Y. A. A. Soetrisno, M. Arfan, and A. Sofwan, "Hoax detection system on Indonesian news sites-based text classification using SVM and SGD," in Proc. 4th Int. Conf. Inf. Technol., Compute., Electra. Eng. (ICITACEE), Oct. 2017, pp. 45–49.
26. H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news sing n-gram analysis and machine learning techniques," in Proc. Int. onf. Intell., Secure, Dependable Syst. Distrib. Cloud Environ. Cham, Switzerland: Springer, 2017, pp. 127–138.

27. S. B. Deokate, "Fake news detection using support vector machine learning algorithm," Int. J. for Res. Appl. Sci. Eng. Technol., vol. 7, no. 7, pp. 438–444, Jul. 2019.
28. H. Ahmed, I. Traore, and S. Saad, "Detecting opinion spams and fake news using text classification," Secure. Privacy, vol. 1, no. 1, p. e9, Jan. 2018, DOI: 10.1002/spy2.9.
29. V. Agarwal, H. P. Sultana, S. Malhotra, and A. Sarkar, "Analysis of classifiers for fake news detection," Proc. Compute. Sci., vol. 165, pp.377–383, Jan. 2019.
30. Ahmed, Sajjad, Knut Winkelmann, and Flavio Corradino. "Development of fake news model using machine learning through natural language processing." arXiv preprint arXiv:2201.07489 (2022).
31. Elyassami, Sanaa, Safa Alseiari, Maryam AL Zaabi, Anwar Hashem, and Nouf Aljahoori. "Fake News Detection Using Ensemble Learning and Machine Learning Algorithms." In Combating Fake News with Computational Intelligence Techniques, pp. 149-162. Springer, Cham, 2022.
32. Asha, J., and A. Meenakowshalya. "Fake News Detection Using N-Gram Analysis and Machine Learning Algorithms." Journal of Mobile Computing, Communications & Mobile Networks 8, no. 1 (2021): 33-43p.
33. Goldani, Mohammad Hadi, Reza Safa bakhsh, and Saeedeh Momtazi. "Convolutional neural network with margin loss for fake news detection." Information Processing & Management 58, no. 1 (2021): 102418.
34. Song, Chenggang, Nansen Ning, Yunlei Zhang, and Bin Wu. "A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks." Information Processing & Management 58, no. 1 (2021): 102437.
35. Saleh, Hager, Abdullah Alharbi, and Saeed Hamood Alsamhi. "OPCNN-FAKE: optimized convolutional neural network for fake news detection." IEEE Access 9 (2021): 129471-129489.
36. Islam, Noman, Asadullah Shaikh, Asma Qaiser, Yousef Asiri, Sultan Almakdi, Adel Sulaiman, Verdah Moazzam, and Syeda Aiman Babar. "Ternion: An Autonomous Model for Fake News Detection." Applied Sciences 11, no. 19 (2021): 9292.
37. Lyu, Shikun, and Dan Chia-Tien Lo. "Fake news detection by a decision tree." In 2020 Southeast on, pp. 1-2. IEEE, 2020.
38. Krishna, N. Leela Siva Rama, and M. Ademola. "Fake News Detection System using Logistic Regression and Compare Textual Property with Support Vector Machine Algorithm." In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), pp. 48-53. IEEE, 2022.
39. Rao, V. Chandra Shekhar, Pulyala Radhika, Niranjan Polala, and Siripuri Kiran. "Logistic Regression versus XGBoost: Machine Learning for Counterfeit News Detection." In 2021 Second International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), pp. 1-6. IEEE, 2021.
40. Julio CS is, et al. "Supervised learning for fake news detection." EEE Intelligent Systems 34.2 (2019): 76-81.

41. Abdullah-All-Tanvir, Ehesas Mia Maher, Saima Akhter, and Mohammad ethanol Huq "Detecting Fake News using Machine Learning and Deep Learning Algorithms" 2019 7th International Conference on Smart Computing and Communications (ICSCC)
42. Supanya Aphiwongsophon, and Prabhas Chongstitvatana. "Detecting fake news with machine learning methods." 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications & Information Technology (ECTI-CON). IEEE, 2018.
43. Jain, Anjali, et al. "A smart System for Fake News Detection Using Machine Learning." 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT). Vol. 1. IEEE, 2019.
44. Agarwal, Arush, and Akhil Dixit. "Fake News Detection: An Ensemble Learning Approach." 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2020.
45. Ahmed, Hadeer, Issa Traore, Sherif Saad. "Detection of online fake news using n-gram analysis and machine learning techniques." International conference on intelligent, secure, dependable systems in distributed & cloud environments. Springer, Cham, 2017.
46. Mugdha, Shafayat Bin Shabbir, Marian Binte Mohammed Moinuddin Kuddus, Lubaba Salsabil, Afra Anika, Piya Prue Marma, Zahid Hossain, and Swakkhar Shatabdi. "A gaussian naive Bayesian classifier for fake news detection in Bengali." Emerging Technologies in Data Mining and Information Security, pp. 283-291. Springer, Singapore, 2021.
47. Parita Jain Swati Sharma, and Puneet Kumar Aggarwal. "Classifying Fake News Detection Using SVM, Naive Bayes, and LSTM." In 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 460-464. IEEE, 2022.
48. Kaliyar, Rohit Kumar, Anurag Goswami, and Pratik Narang. "Echo Faked: improving fake news detection in social media with an efficient deep neural network." Neural computing and applications 33, no. 14 (2021): 8597-8613.
49. Mouratidis, Despoina, Maria Nefeli Nikiforuk, and Katia Lada Koumanidis. "Deep learning for fake news detection in a pairwise textual input schema." Computation 9, no. 2 (2021): 20.
50. Goldani, Mohammad Hadi, Saeedeh Mumtaz, and Reza Safa bakhsh. "Detecting fake news with capsule neural networks." Applied Soft Computing 101 (2021): 106991.
51. Ni, Shiwen, Jiawen Li, and Hung-Yu Kao. "MVAN: Multi-View Attention Networks for Fake News Detection on Social Media." IEEE Access 9 (2021): 106907-106917.
52. Jain, Vidit, Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Harshvardhan Sharma. "AENeT: an attention-enabled neural architecture for fake news detection using contextual features." Neural Computing and Applications 34, no. 1 (2022): 771-782.
53. F. C. Fernández-Reyes and S. Shinde, "Evaluating deep neural networks for automatic fake news detection in the political domain," in Proc. Ibero-Amer. Conf. Artif. Intell. Cham, Switzerland: Springer, 2018, pp. 206–216
54. Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "EANN: Event adversarial neural networks for multimodal fake news detection," in Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery

Data Mining, Jul. 2018, pp. 849–857.

55. J. Zhang, B. Dong, and P. S. Yu, "Fake Detector: Effective fake news detection with a deep diffusive neural network," in Proc. IEEE 36th Int. Conf. Data Eng. (ICDE), Apr. 2020, pp. 1826–1829.
56. S. Singhania, N. Fernandez, and S. Rao, "3HAN: A deep neural network for fake news detection," in Proc. Int. Conf. Neural Inf. Process. Cham, Switzerland: Springer, 2017, pp. 572–581.
57. N. Ruchansky, S. Seo, and Y. Liu, "CSI: A deep hybrid model for fake news detection," in Proc. ACM Conf. Inf. Knowl. Manage., Nov. 2017, pp. 797–806
58. S. Chopra, S. Jain, and J. M. Scholar, "Towards automatic identification of fake news: Headline-article stance detection with LSTM attention models," in Proc. Stanford CS224d Deep Learn. NLP Final Project, 2017, pp. 1–15.
59. Nasir, Jamal Abdul, Osama Suhani Khan, and Irakli's Varlas's. "Fake news detection: A hybrid CNN-RNN based deep learning approach." International Journal of Information Management Data Insights 1, no. 1 (2021): 100007.
60. Samadi, Mohammadreza, Maryam Mousavi an, and Saeedeh Momtazi. "Deep contextualized text representation and learning for fake news detection." Information Processing & Management 58, no. 6 (2021): 102723.
61. Song, Chenggang, Kai Shu, and Bin Wu. "Temporally evolving graph neural network for fake news detection." Information Processing & Management 58, no. 6 (2021): 102712.
62. Yuan, Hua, Jie Zheng, Qiongwei Ye, Yu Qian, and Yan Zhang. "Improving fake news detection with a domain-adversarial and graph-attention neural network." Decision Support Systems 151 (2021): 113633.
63. Bahad, Pritika, Preeti Saxena, and Raj Kamal. "Fake news detection using bi-directional LSTM-recurrent neural network." Procedia Computer Science 165 (2019): 74-82.
64. Jadhav, Shrutika S., and Sudeep D. Thepade. "Fake news identification and classification using DSSM and improved recurrent neural network classifier." Applied Artificial Intelligence 33, no. 12 (2019): 1058-1068.
65. Zhang, Tong, Di Wang, Huanhuan Chen, Zhiwei Zeng, Wei Guo, Chunyan Miao, and Lizhen Cui. "BDANN: BERT-based domain adaptation neural network for multimodal fake news detection." In 2020 joint international conference on neural networks (IJCNN), pp. 1-8. IEEE, 2020.
66. Kaliyar, Rohit Kumar, Anurag Goswami, Pratik Narang, and Soumendu Sinha. "FNDNet–a deep convolutional neural network for fake news detection." Cognitive Systems Research 61 (2020): 32-44.
67. Huang, Yin-Fu, and Po-Hong Chen. "Fake news detection using an ensemble learning model based on self-adaptive harmony search algorithms." Expert Systems with Applications 159 (2020): 113584.
68. Tida, Vijay Srinivas, Dr. Hsu, and Dr. Hei. "Unified fake news detection using transfer learning of bidirectional encoder representation from transformers model." arXiv preprint arXiv:2202.01907 (2022).

69. Kaliyar, Rohit Kumar, Anurag Goswami, and Pratik Narang. "DeepFakE: improving fake news detection using tensor decomposition-based deep neural network." The Journal of Supercomputing 77, no. 2 (2021): 1015-1037.
70. Umer, Muhammad, Zainab Imtiaz, Saleem Ullah, Arif Mehmood, Gyu Sang Choi, and Byung-Won On. "Fake news stance detection using deep learning architecture (CNN-LSTM)." IEEE Access 8 (2020): 156695-156706.
71. Kumar, Sachin, Rohan Asthana, Shashwat Upadhyay, Nidhi Upreti, and Mohammad Akbar. "Fake news detection using deep learning models: A novel approach." Transactions on Emerging Telecommunications Technologies 31, no. 2 (2020): e3767.
72. https://paperswithcode.com/dataset/liar
73. https://www.kaggle.com/c/fake-news
74. https://github.com/several27/FakeNewsCorpus/releases/tag/v1.0
75. https://www.kaggle.com/code/tarzon/custom-word-cloud/data.
76. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th International Conference on World Wide Web, pp. 675–684 (2011)
77. Rubin, V.L., Conroy, N.J., Chen, Y.: Towards news verification: Decep detection methods for news discourse. In: Hawaii International Conference on System Sciences, pp. 5–8 (2015)
78. Ferreira, W., Vlachos, A.: Emergent: a novel data-set for stance classi- fiction. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2016). ACL
79. Janze, C., Risius, M.: Automatic detection of fake news on social media platforms (2017)
80. Wang, W.Y.:" liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648 (2017)
81. Shu, K., Wang, S., Liu, H.: Exploiting tri-relationship for fake news detection. arXiv preprint arXiv:1712.07709 8 (2017)
82. Lu, Y.-J., Li, C.-T.: Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. arXiv preprint arXiv:2004.11648 (2020)
83. Wang, Y., Zhang, Y., Li, X., Yu, X.: Covid-19 fake news detection using bidirectional encoder representations from transformers-based models. arXiv preprint arXiv:2109.14816 (2021)
84. Jarrahi, A., Safari, L.: Fr-detect: A multi-modal framework for early fake news detection on social media using publishers' features. arXiv preprint arXiv:2109.04835 (2021).
85. Sahoo, S.R., Gupta, B.B.: Multiple features-based approach for auto-matic fake news detection on social networks using deep learning. Applied Soft Computing 100, 106983 (2021).
86. Seddari, N., Derhab, A., Belaoued, M., Halboob, W., Al-Muhtadi, J., Bouras, A.: A hybrid linguistic and knowledge-based analysis approach for fake news detection on social media. IEEE Access (2022).
87. Zhou, X., Zafarani, R.: A survey of fake news: Fundamental theo- ries, detection methods, and opportunities. ACM Computing Surveys (CSUR) 53(5), 1–40 (2020)

# Figures

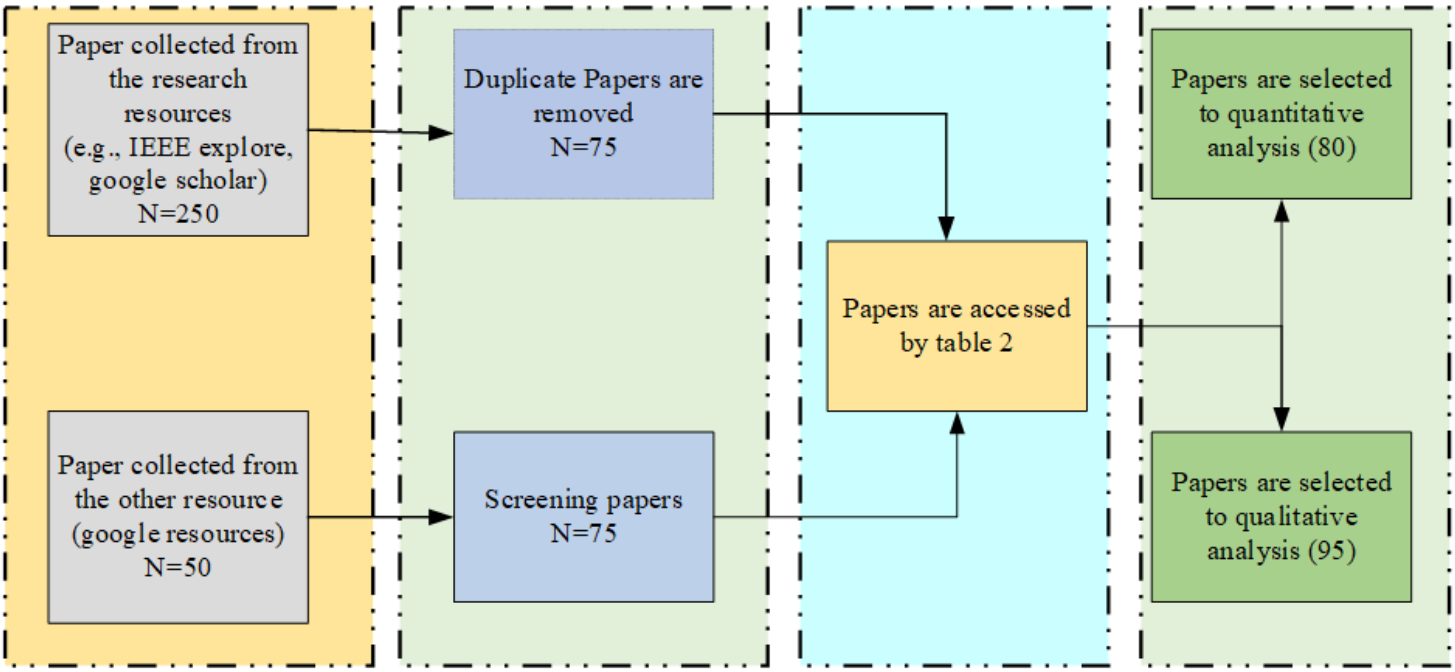



**Figure 1**

PRISMA flow chart for fake news prediction

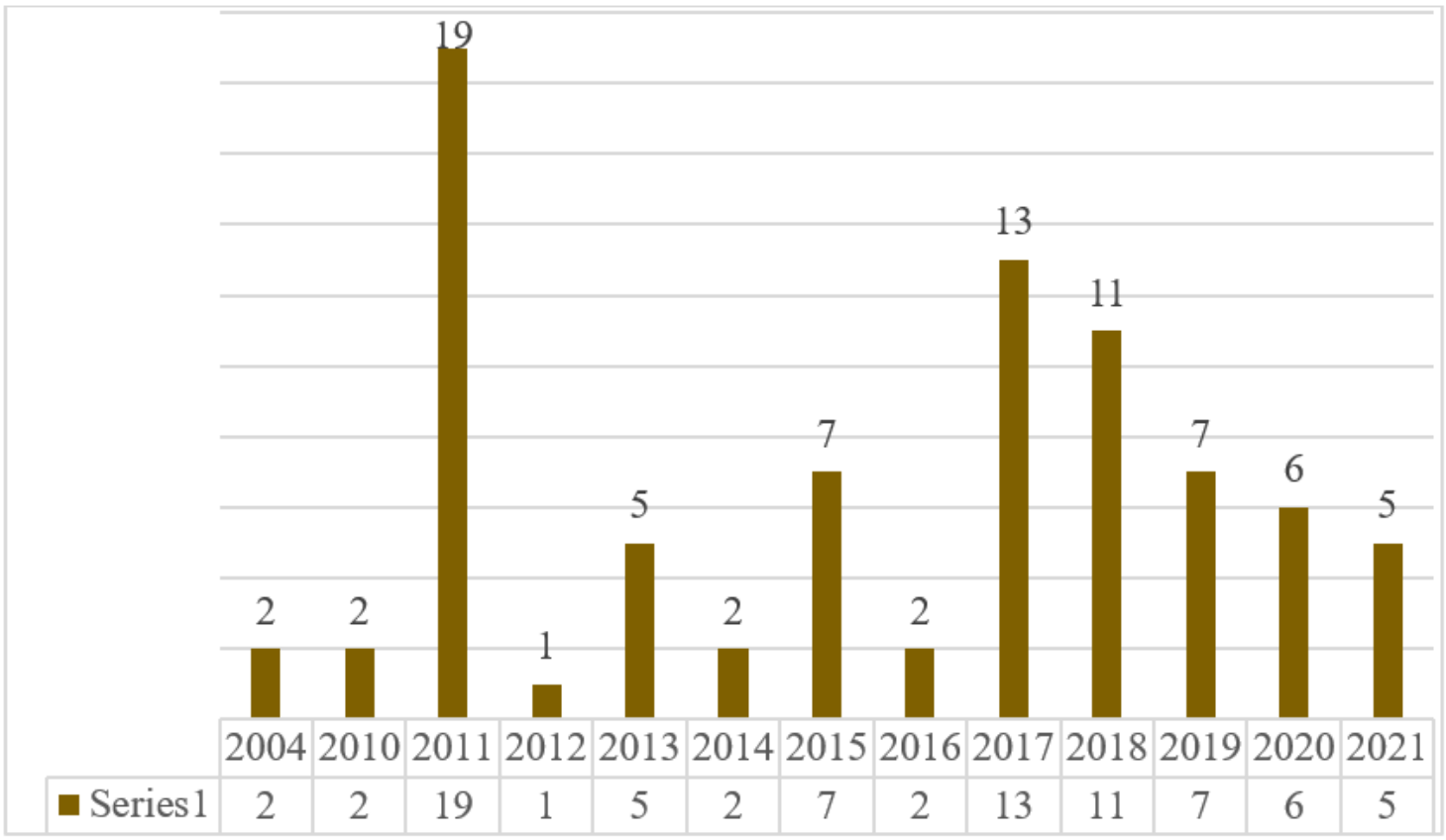
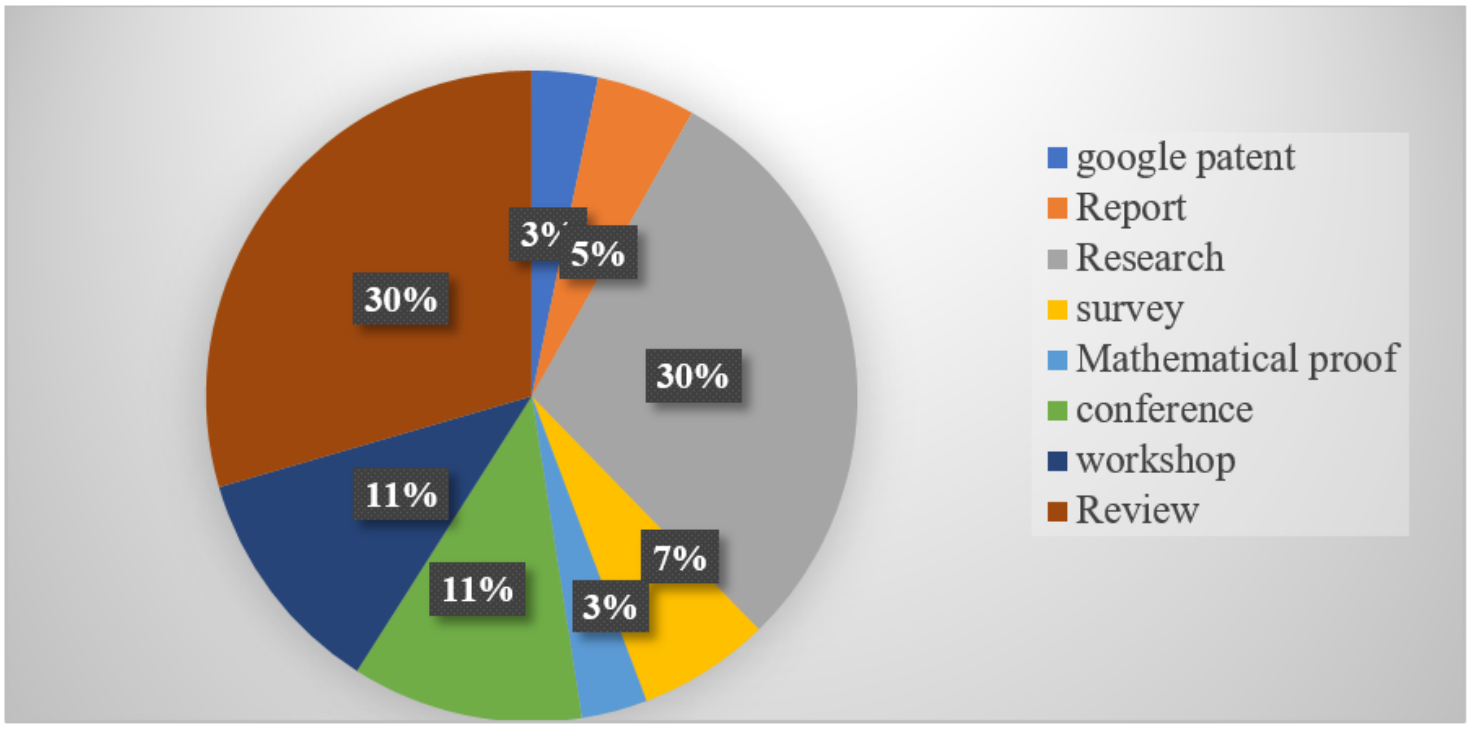


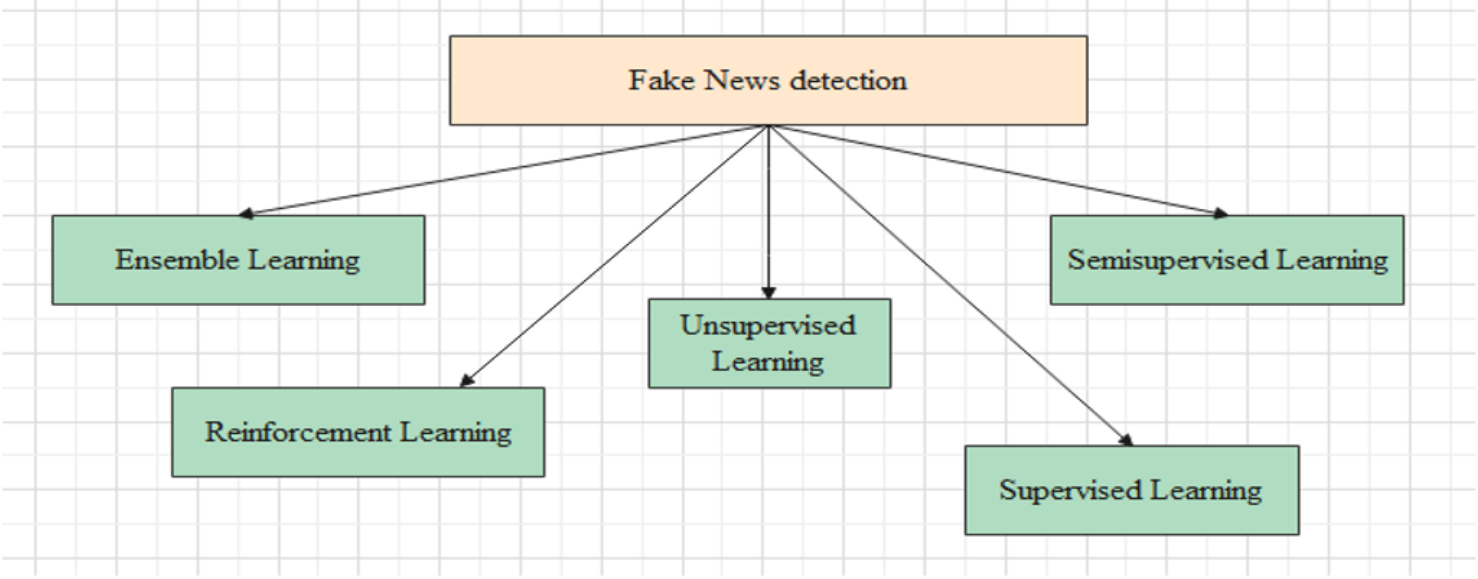

**Figure 2**

Papers for the published year



## Figure 3

Analytical study of papers




## Figure 4

Types of fake news detection

**Figure 5**

Process of fake news detection

**Figure 6**

Analysis of Supervised Learning

**Figure 7**

Analysis of Unsupervised Learning

## Figure 8

Analysis of Reinforcement Learning

## Figure 9

Analysis of ensemble Learning

## Figure 10

Analysis of (a) word cloud of fake news and (b) word cloud of real news [75]

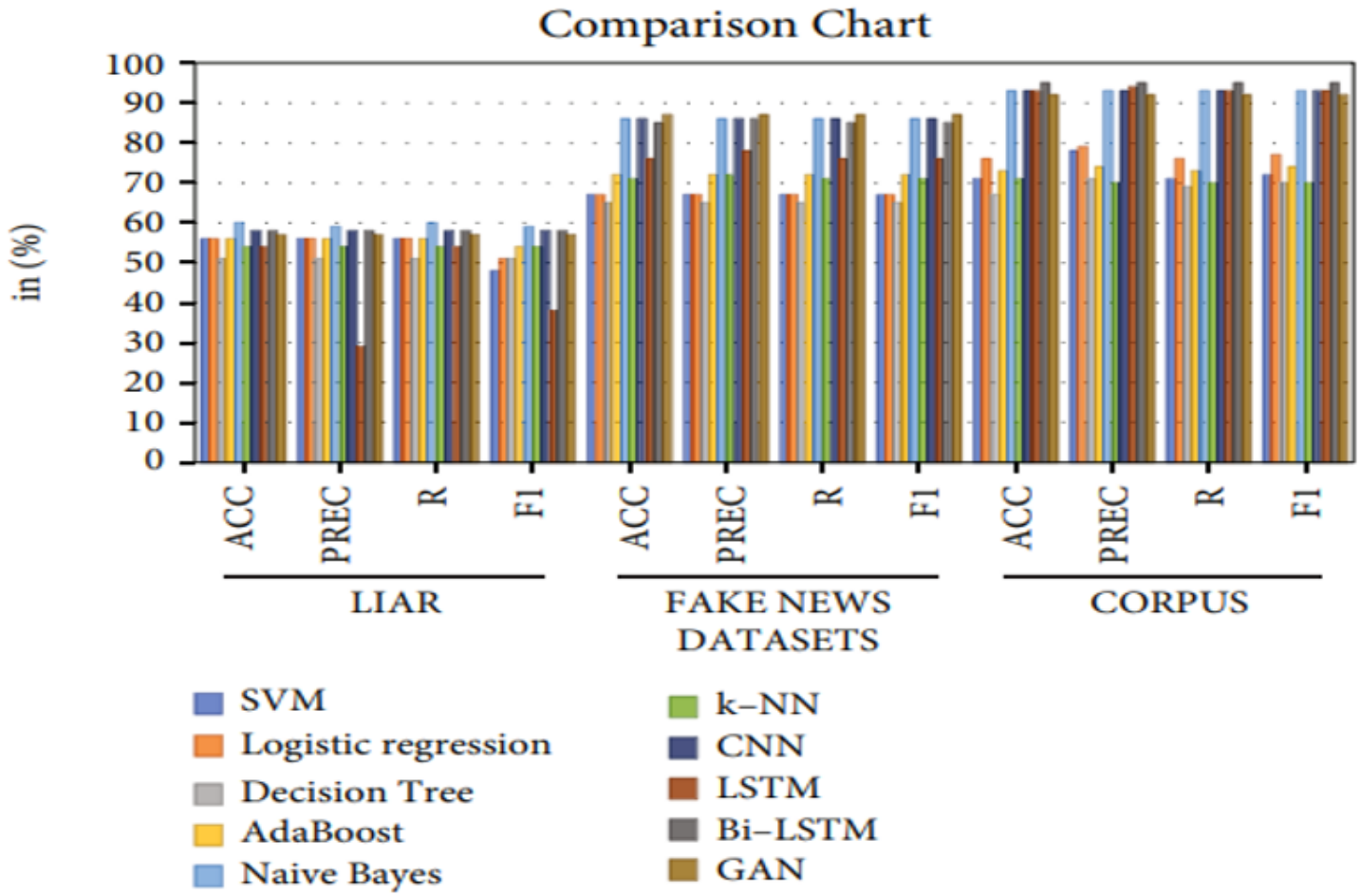



**Figure 11**

Accuracy evaluation of the different learning techniques for various databases

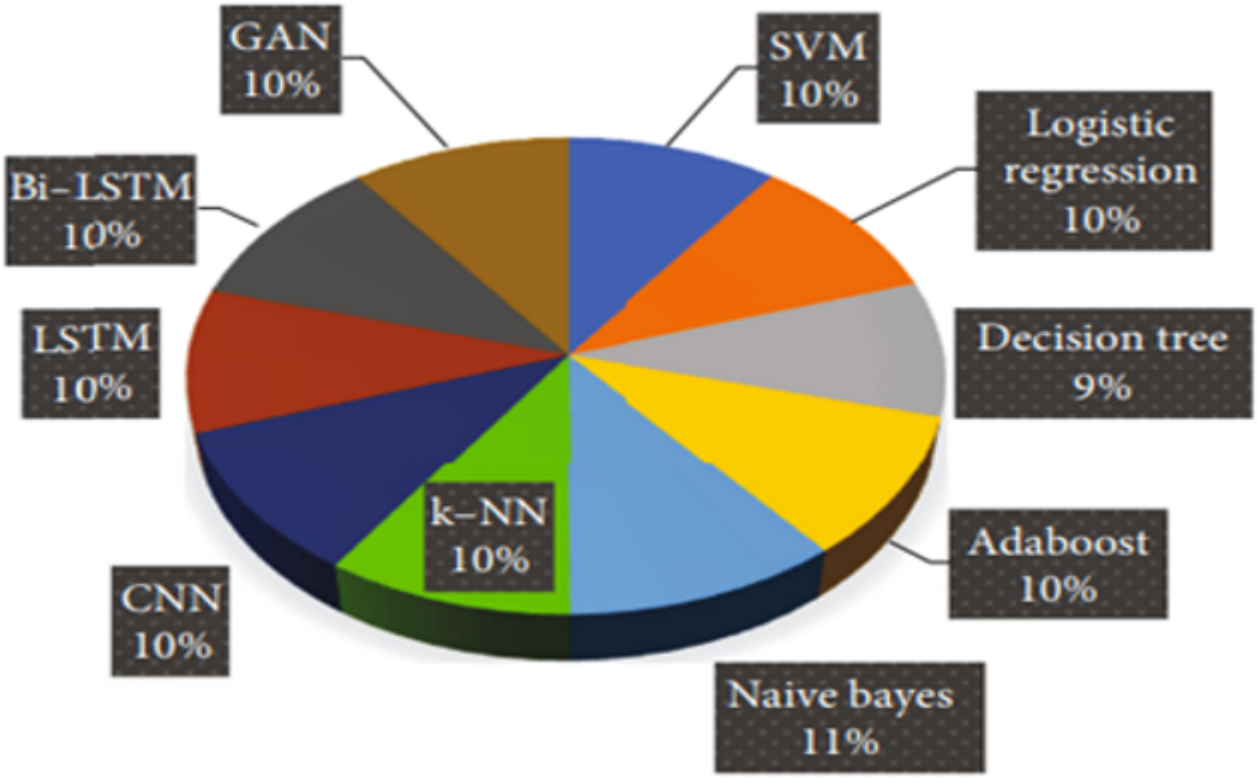



**Figure 12**

Analysis of liar dataset with the Proportion of different methods used

**Figure 13**

Analysis of corpus dataset with the Proportion of different methods used

**Figure 14**

Analysis of fake news dataset with the Proportion of different methods used

Check for updates

*Literature Review*

# Fifty Years of Research in Deposit Insurance: A Bibliometric Analysis and Review

FIIB Business Review
1–21
© 2022 Fortune Institute of International Business
Reprints and permissions:
in.sagepub.com/journals-permissions-india
DOI: 10.1177/23197145221116455
journals.sagepub.com/home/fib



**Varda Sardana[1] and Shubham Singhania[2]**

## Abstract

Deposit insurance is a tool to enhance banking stability in a country, by augmenting the confidence of the bank depositors and reducing their incentive to run on banks. Over the years, this instrument of bank safety net has garnered considerable attention, owing to the advent of multiple banking shocks and financial crises. This is a first-of-its-kind study that aims to trace the development of deposit insurance research over the past 50 years, from 1971 to 2021, through bibliometric analysis and network visualization. A total of 1322 research articles and reviews on deposit insurance, drawn from the Scopus database, have been examined and synthesized through volume analysis, citation analysis, co-authorship analysis, co-citation analysis and keyword co-occurrence analysis. Two well-known bibliometric tools, VOSviewer and Biblioshiny, have been used to create and analyze various bibliometric maps, networks and tables. This study uncovers the temporal and spatial trends of publications on deposit insurance, the most influential documents, authors, journals and countries in this domain, the status of collaborations and the major themes that have been studied in the literature. The paper also highlights new concepts that have emerged over the past five decades and suggests future avenues for research in this area.

## Keywords

Deposit Insurance, bibliometric analysis, network visualization, Scopus, banking

## Introduction

The banking system of a country plays a significant role in financial intermediation and acts as the nodal point for an economy (Bhatia & Gupta, 2020). Any negative blow to the banking sector can lead to a meltdown in the savings, investments and payment system of a nation. To ensure the smooth functioning of the financial system and the economy as a whole, it is imperative to provide protection to this sector. Deposit insurance is one such safeguard (Gatti & Oliviero, 2018).

Deposit insurance is a system under which deposit holders of specified financial institutions are offered a guarantee as to the full or partial protection of the principal amount and accrued interest held by them in their deposit accounts (Garcia, 1996; Putri et al., 2021). Such protection may be explicitly given by law or may be implicit from the past actions of the authorities (International Association of Deposit Insurers [IADI], 2021). The term 'specified financial institutions' mainly refers to banks. However, in certain countries, the scope has been extended to include other deposit-taking institutions as well, such as insurance companies (Australia, Belgium, Ecuador and Malaysia) and securities companies (Denmark, Korea and Sweden; IADI, 2021). The primary functions performed by banks make them vulnerable to bank runs and, hence, entitled to protection in the form of deposit insurance (Calomiris, 1990). Banks carry liabilities (demand deposits) that are liquid in nature, and these liquid liabilities are invested in illiquid assets in the form of loans and advances. This imbalance on their balance sheets exposes them to liquidity risk, that is, the risk of running out of demand deposits if a large number of depositors want to withdraw their money from their bank accounts at the same time (Jackson & Kotlikoff, 2021; Sardana & Shukla, 2020). Added to this is the issue of 'sequential service constraint' as per which banks promise the depositors to fulfil their withdrawal

[1] Shri Ram College of Commerce, University of Delhi, India.
[2] University School of Management and Entrepreneurship, Delhi Technological University, Delhi, India.

**Corresponding author:**
Varda Sardana, Shri Ram College of Commerce, University of Delhi, Maurice Nagar, Roop Nagar, Delhi 110007, India.
E-mail: varda.sardana@gmail.com

needs on a first-come-first-serve basis (Diamond & Dybvig, 1983). In such a situation, if the depositors believe that a bank is in trouble, all of them would run to convert their account balances into cash before other customers (Calomiris, 1999), and this bank run in itself would lead to a bank failure (Dubois, 2021).

A system of deposit insurance controls a widespread bank failure and instability by assuring the depositors that their deposits are safe in the bank even if the bank is in trouble (Putri et al., 2021). This enhances the confidence of the depositors (Atanasova et al., 2019) and reduces the probability of panic-stricken bank runs (Wu et al., 2020). Apart from deposit insurance, the tools that form part of the bank safety net include regulation and supervision of the banking system, and the facility of lender of last resort (Calomiris, 1999). These three systems complement each other in safeguarding banking system stability.

The current research article aims to trace the development of deposit insurance research and assess the paradigm of studies that have accumulated through time. To establish a better understanding of our field of study and discover developing research areas, bibliometric analysis has been employed. Bibliometric analysis is a method to examine a field's underlying scope, recent status and advancements (Castriotta et al., 2019; Tunger & Eulerich, 2018). It is now widely accepted as a reliable tool for evaluating research in a plethora of subjects and domains (Ellegaard & Wallin, 2015). Through this study, we answer the following research questions:

RQ1: How is the volume of publications in the deposit insurance research distributed over time and geographies over the past 50 years?
RQ2: Which publications, authors, journals and countries have impacted the research on deposit insurance?
RQ3: What is the social structure among authors and countries that have come together to study the academic construct of deposit insurance?
RQ4: What are the various thematic sets formed in our field of study?
RQ5: What specific subjects and concepts have emerged through the trajectory of our field?

Although deposit insurance, as a research topic, has been extensively studied through qualitative literature reviews and conceptual reviews, this is the first study that employs bibliometric analysis to this domain of banking and financial regulation. The bibliometric review in this article helps uncover the temporal and spatial trend of publications on deposit insurance; the most influential documents, authors, journals and countries in this domain; the status of collaborations; and the major themes that have been studied in the literature. By providing a clearer picture of the literature and its evolution, this study holds relevance for researchers as well as academicians. The findings of this study would also have practical utility for deposit insurers, deposit-taking institutions and policymakers, as it would provide them a direction for decision-making in terms of fulfilling their claim obligations as well as the deposit insurance laws.

## Literature Review

There is no dearth of literature on deposit insurance. The research in this area has focused on multiple aspects of deposit protection, including the design features of the system, the ability of the system to fulfil its objectives, the associated externalities—both negative and positive—and evaluation of the deposit insurance system in specific countries, as explained further.

A deposit insurance system should, preferably, be designed in a country as per its own institutional and supervisory environment (Alam et al., 2021; Barth et al., 2004), since adopting a pre-designed insurance system can cause more harm than good (Demirgüç-Kunt & Kane, 2002; Garcia, 1996). Where on one hand, a government-managed deposit insurance system can make the system credible and better controlled (Diamond & Dybvig, 1983), on the other hand, a public–private partnership can help to reduce the burden of depositor claims on the government-managed insurer (Stone, 2021). The coverage level of insurance is an important decision, and one of the most challenging (Chen et al., 2021). As per Lee and Kwok (2000) and Chiaramonte et al. (2020), a full coverage or a blanket guarantee promotes banking stability by allowing some extra time for the banks to revive themselves in the absence of constant pressure from depositors. Simultaneously, there is ample evidence to support the claim that a higher level of coverage invites higher moral hazard and lowers market discipline (Demirgüç-Kunt & Detragiache, 2002; Demirgüç-Kunt & Huizinga, 2004; Ngalawa et al., 2016). Additionally, a deposit insurance with partial coverage can motivate the depositors to divide their funds and deposit them in accounts with multiple banks, which further leads to the erosion of competition among banks (Mohammad & Nishiyama, 2019; Shy et al., 2016).

While a voluntary membership of the deposit insurance scheme may be offered by the insurer, such a system would lead to a problem of adverse selection (Lee & Kwok, 2000) and a further increase in risk-taking by the insured banks (Wheelock & Kumbhakar, 1995). The premium charged from insured institutions can be either at a flat rate or variable rate. While a flat rate premium provides the convenience of ease of calculation, it exacerbates the issue of moral hazard (Kane, 1986; Keeley, 1990). A risk-based premium, on the other hand, reduces this problem to a great extent (Chang & Ho, 2017; Merton, 1977; Wu et al., 2020).

As far as the evidence of financial stability is concerned, on one hand, Prean and Stix (2011) suggest that adoption of deposit insurance during a banking crisis enhances the confidence, trust and credibility among the depositors,

whereas, on the other hand, Boyle et al. (2015) emphasize that such an adoption may only be partially successful in preventing panic-based withdrawals from banks, as compared to a pre-existing insurance scheme. This has been substantiated by Demirgüç-Kunt et al. (2014), who point out that those countries that had an explicit system of deposit insurance in place before the global financial crisis faced fewer runs on banks. Recently, Igwe and Toby (2021) as well as Alam et al. (2021) have found a positive impact of deposit insurance on banking stability, given the presence of other favourable economic factors.

However, the adoption of an explicit deposit insurance does not necessarily solve all the problems. Such a system may, in fact, increase the likelihood of bank failures (Demirgüç-Kunt & Detragiache, 2002), destabilize the banking system in good times (Anginer et al., 2014) and hamper the functioning of financial markets (Kozińska, 2021). The provision of an insurance facility is often accompanied by certain externalities. In the case of deposit insurance, the negative externalities are in the form of a possible deterioration in the market discipline (Calomiris & Jaremski, 2019; Demirgüç-Kunt & Huizinga, 2004; Vernikov, 2020) and a possible aggravation of moral hazard (Chi & Binh, 2020; Gupta & Sardana, 2021; Kusairi et al., 2018; Zhang & Wu, 2020), both of which have been discussed extensively by researchers.

As far as positive externalities are concerned, in case the law-and-order conditions are strong, deposit protection gives a push to the development of the banking sector in the long run (Bergbrant et al., 2016). Furthermore, external deposits increase in banks following the adoption of explicit deposit insurance (Altunbaş & Thornton, 2013). As per Huizinga and Nicodeme (2006), there is an improvement in bank deposits held by non-residents in developed countries after the adoption of deposit insurance if the system is credible and ensures additional protection for international depositors. The extent of deposit insurance coverage also has an impact on the effectiveness of monetary policy (Lin, 2015), and it contributes to better financial intermediation (Jameaba, 2018).

Although the literature review underlines the various themes that have been studied, there is still a need to synthesize the literature and bibliometric data in order to accentuate the gaps, recognize the domains that remain partially or fully unexplored and determine the developments as well as scope in this field. The present study tries to accomplish the same through bibliometric analysis.

## Research Methodology

The study employs bibliometric analysis to analyse and review documents in our field of research. This method equips us to deal with a huge pool of data through a quantitative structure and to discern some of the evolving trends in the research domain through the years (Xu et al., 2018). Scopus database has been used to source relevant documents, as it is the world's largest research documents database for peer-reviewed papers, books and conference articles (Singh et al., 2021). It offers a broad coverage in the highest-level academic disciplines such as social sciences, life sciences, physical sciences and health sciences (Mongeon & Paul-Hus, 2016).

Figure 1 lays down the step-by-step process followed for screening the documents used for the analysis. Primarily, a search was initiated in the main collection of the Scopus database using a string of keywords, after a thorough review of literature in the field, to ensure that alternative forms of usage of the word 'deposit insurance' are captured in the title, abstract or keywords. The keywords employed were: 'deposit* insur*' OR 'deposit* guarantee*' OR 'deposit* protect*'. The document titles were screened discipline-wise, and only those disciplines wherein the title of documents matched the definition of deposit insurance in context of banking were included. Furthermore, to ensure the quality of the peer-reviewed process, only research articles and reviews were considered, thereby excluding any other type of grey literature. The past 50 years have shown a sustained and secular flow of research work in deposit insurance, making it an area of interest for bibliometric mapping. A total of 1,322 documents were finalized for the analysis, after carefully going through the abstracts of each document to avoid duplication or unsuitability.

Ultimately, our bibliometric sample was gathered, with bibliographic details such as document title, authors' names, affiliations, publication source and year, as well as the synopsis, key terms and references of each publication. This dataset was used to create and analyse various bibliometric maps, tables and graphs through Microsoft Excel, VOSviewer (Van Eck & Waltman, 2010) and Biblioshiny, which runs on the Bibliometric package in R (Aria & Cuccurullo, 2017). VOSviewer, an open-source software, offers multiple user-friendly functionalities and interactive mapping features (Goel et al., 2021; Khanra et al., 2020) and allows a high level of flexibility as well as elaborate visualization (Tomar et al., 2021). Biblioshiny is also a widely accepted tool that can perform multiple analyses, which are unique vis-à-vis other tools (Moral-Muñoz et al., 2020). These software have enabled the network visualization of various scientific mapping techniques explored in our article. This analysis will not only synthesize and draw a structure to understand past research but also set up future implications for researchers (Sánchez-Teba et al., 2021).

## Research Analysis and Findings

The use of bibliometric analysis has assisted in discovering relationships and linkages between documents, authors, countries and journals through bibliometric maps and thematic clusters (Xu et al., 2018). It has enabled the

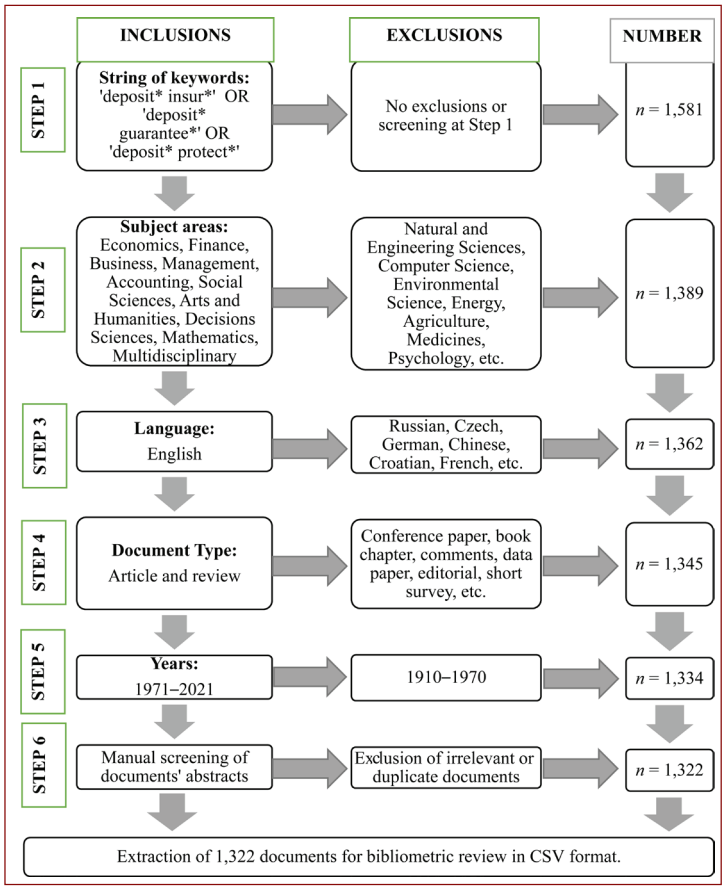



**Figure 1.** Step-by-step Inclusion and Exclusion Criteria for Documents for Bibliometric Analysis

**Source:** The authors.

examination of the evolution and academic advancement of our research field. The following segments describe five distinct kinds of analysis: volume, citation, co-authorship, co-citation and keyword co-occurrence analysis, which have been undertaken to address our research questions.

## *Volume Analysis of Publications*

The volume of documents published on deposit insurance has an upward trajectory (Figure 2). It has witnessed two sustained peaks: one in the 1990s and another immediately after the Global Financial Crisis (GFC), that is, in 2009.

The 1990s saw the deliberation and adoption of explicit deposit guarantees by multiple countries such as Argentina, Brazil, Colombia, Croatia, Greece, Hungary, Korea, Poland, Portugal and Sweden, thereby leading to a substantial publication of discussion papers and policy research papers. The next jump came after the GFC of 2007–2008, which had a massive impact on the banking sector, the savings and investments of households and corporates. The crisis brought back the focus of policymakers and researchers on the various safety nets, one of which is deposit insurance, as well as the behaviour of depositors and banks during a banking crisis. The global crisis even led to hasty adoption of deposit insurance, or increase in insurance coverage limits, such as in the United States of America (USA) and European countries, or substantial changes in insurance design features by affected countries (Atmaca et al., 2021).

The year 2019 has set the record for the highest number of publications, thereby highlighting the fact that countries, central banks, regulators and governments are continuously recognizing the importance of deposit insurance and, hence, the researchers (including regulatory organizations) and academicians are keen to study its impact and relations with different dimensions of the financial system.

As far as the spatial distribution of publications is concerned, Figure 3 provides an overview of the same for the 20 countries with the highest number of publications over the past 50 years, out of a total of 77 countries publishing in the domain. The size of the bubble is representative of the volume of publications in the particular country.

The USA tops the charts, with 554 published documents, and it remains unrivalled in terms of the number of publications. This is understandable because the USA has the oldest deposit insurance system in the world, aged 88 years. Furthermore, the deposit insurance system in the

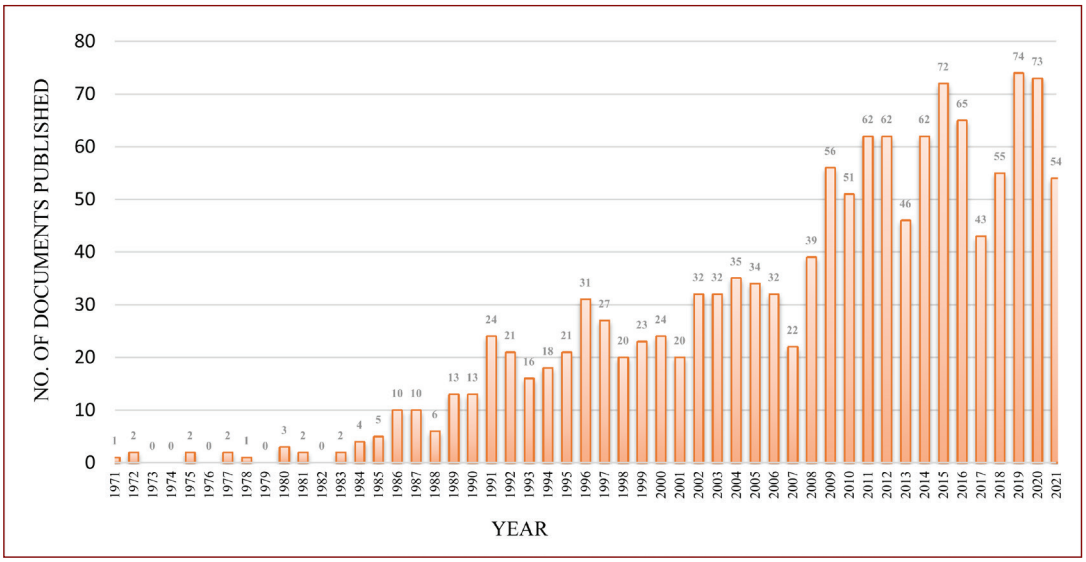



**Figure 2.** Volume of Documents Published over the Past 50 Years

**Source:** The authors' creation based on Scopus database.

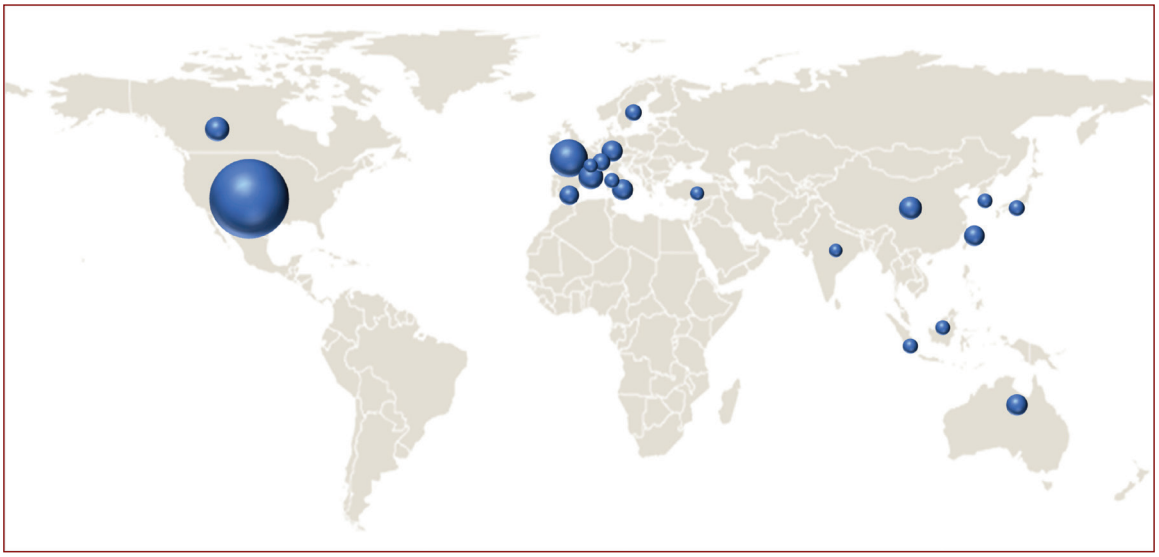


**Figure 3.** Geographical Distribution of Publications for the Top 20 Countries by Volume

**Source:** The authors' creation based on Scopus database.

USA has undergone many regulatory modifications, and the insurance fund has been utilized extensively due to the frequent banking failures witnessed by the economy, especially during the GFC (Maura-Pérez & Romero-Perez, 2021). India, the twentieth country in the list by volume of publications, has about 16 publications only, though the system here is 60 years old and the second oldest in the world. Figure 2 also makes it clear that the European countries have a major share in publications as well, that too in the countries that are a part of the European Union (EU), that is, France, Germany, Italy, Netherlands, Finland and Spain. This is majorly in response to the directives adopted by the European Commission on the mandatory establishment of deposit guarantees and harmonization of protection schemes across the EU nations (see Dubois, 2021). Overall, 46.2% of the research contributions are from North America, 27.5% from Europe, 18.6% from Asia, 3.02% from Australia, 2.8% from Africa and the remaining from South America. The lower share of publications in the continents of Asia, Australia and Africa is due to the relatively lesser macro-level disruptions in the banking arena of the countries that span these continents.

## Citation Analysis of Documents, Authors, Journals and Countries

One of the tactics of scientific mapping is citation analysis, which works on the postulation that citations exhibit academic connections between papers when they are cited (Sánchez-Teba et al., 2021). This analysis is one of the most straightforward methods out of all, and it works by defining the influence of a research paper by the volume of citations that it receives (Stremersch et al., 2007). It assists in revealing documents that have impacted researchers strongly as compared to the ones that are cited less frequently in a particular field of research (Donthu et al., 2021; Fabregat-Aibar et al., 2019).

Table 1 presents the top 20 most-cited documents published over the past five decades, along with the total citations (TC) received by them. Diamond and Dybvig (1983) is a seminal paper that studies the relation between bank runs and deposit protection, and that advocates the case for government administered insurance. Merton (1977) is the first work in the field to derive the formula of the cost of deposit insurance using option pricing theory.

**Table 1.** The 20 Most Influential Deposit Insurance Documents Ranked by Citations

| Rank | Document Title | Author(s) | Source | Year | TC | ACPY |
|---|---|---|---|---|---|---|
| 1 | Bank Runs, Deposit Insurance and Liquidity | D. W. Diamond and P. H. Dybvig | *Journal of Political Economy* | 1983 | 3183 | 83.76 |
| 2 | Bank Governance, Regulation and Risk Taking | L. Laeven and R. Levine | *Journal of Financial Economics* | 2009 | 1282 | 106.83 |
| 3 | Deposit Insurance, Risk and Market Power in Banking | M. C. Keeley | *American Economic Review* | 1990 | 1151 | 37.13 |
| 4 | Bank Regulation and Supervision: What Works Best? | J. R. Barth, G. Caprio Jr. and R. Levine | *Journal of Financial Intermediation* | 2004 | 1002 | 58.94 |
| 5 | An Analytic Derivation of the Cost of Deposit Insurance and Loan Guarantees: An Application of Modern Option Pricing Theory | R. C. Merton | *Journal of Banking and Finance* | 1977 | 844 | 1.91 |
| 6 | Determinants of Commercial Bank Interest Margins and Profitability: Some International Evidence | A. Demirgüç-Kunt and H. Huizinga | *World Bank Economic Review* | 1999 | 703 | 31.95 |
| 7 | The Determinants of Banking Crises in Developing and Developed Countries | A. Demirgüç-Kunt and E. Detragiache | *IMF Staff Papers* | 1998 | 684 | 29.74 |
| 8 | Does Deposit Insurance Increase Banking System Stability? An Empirical Investigation | A. Demirgüç-Kunt and E. Detragiache | *Journal of Monetary Economics* | 2002 | 537 | 28.26 |
| 9 | A Theory of Bank Capital | D. W. Diamond and R. G. Rajan | *Journal of Finance* | 2000 | 483 | 23 |
| 10 | Risk in Banking and Capital Regulation | D. Kim and A. M. Santomero | *Journal of Finance* | 1988 | 431 | 13.06 |
| 11 | Bank Competition and Stability: Cross-Country Heterogeneity | T. Beck, O. De Jonghe and G. Schepens | *Journal of Financial Intermediation* | 2013 | 391 | 48.87 |
| 12 | Pricing Risk-Adjusted Deposit Insurance: An Option-Based Model | E. I. Ronn and A. K. Verma | *Journal of Finance* | 1986 | 356 | 10.17 |
| 13 | An International Comparison of Capital Structure and Debt Maturity Choices | J. P. H. Fan, S. Titman and G. Twite | *Journal of Financial and Quantitative Analysis* | 2012 | 352 | 39.11 |
| 14 | A Model of Reserves, Bank Runs and Deposit Insurance | J. Bryant | *Journal of Banking and Finance* | 1980 | 347 | 8.46 |
| 15 | Size and Performance of Banking Firms: Testing the Predictions of Theory | J. H. Boyd and D. E. Runkle | *Journal of Monetary Economics* | 1993 | 318 | 11.35 |
| 16 | Financial Intermediaries and Liquidity Creation | G. Gorton and G. Pennacchi | *Journal of Finance* | 1990 | 310 | 10 |
| 17 | Technical, Scale and Allocative Efficiencies in US Banking: An Empirical Investigation | H. Y. Aly, R. Grabowski, C. Pasurka and N. Rangan | *Review of Economics & Statistics* | 1990 | 303 | 9.77 |
| 18 | Deregulation and Bank Financial Policy | A. J. Marcus | *Journal of Banking and Finance* | 1984 | 292 | 7.89 |
| 19 | Deposit Insurance and Wealth Effects: The Value of Being `Too Big to Fail' | M. O'hara and W. Shaw | *Journal of Finance* | 1990 | 275 | 8.87 |
| 20 | Capital Regulation and Bank Risk-Taking | F. T. Furlong and M. C. Keeley | *Journal of Banking and Finance* | 1989 | 270 | 8.43 |

**Source:** The authors' analysis based on Scopus database.
**Note:** TC = Total citations (total citations within the database); ACPY = Average citations per year.

The average citations per year (ACPY) is highest for Laeven and Levine (2009), which though published only 12 years ago, has been highly influential by virtue of its novelty in empirical assessment of bank risk-taking theories and its linkages with deposit insurance, capital regulation and bank ownership structure. On similar lines, Beck et al. (2013) has a reasonably high ACPY, though it has completed less than a decade of its publication. Table 1 is indicative of the continuously developing curiosity and research in the domain of deposit insurance amongst scholars, since the most influential publications go back to as early as 1977 to as recent as 2013.

Table 2 lists the authors who have made the most impact in our field, ranked by the number of citations on their publications. It also lays down the h-index, which is indicative of the authors' productivity (Pandey & Joshi, 2021). A comparison of Table 1 and Table 2 reveals that even though the paper written by Diamond and Dybvig (1983) is the most influential in terms of citations as well as ACPY, Asli Demirgüç-Kunt has made most substantial contributions to the field of deposit insurance, followed by Luc Laeven, both of whom are affiliated to prominent global organizations. Edward J. Kane and James R. Barth, both affiliated to institutions in the USA, are also among the most substantial contributors, in terms of impact as well as number of documents (NOD). Gary Gorton also deserves a special mention since despite having only six Scopus-listed publications in the past five decades, his studies have been able to create a rippling effect, making him an author with one of the highest TC and average citations per document (ACPD). It is also to be noted that the majority of the influential authors are affiliated to institutions in the USA, which corroborates our findings of volume analysis from Figure 2, while some of them are affiliated to the central banks of their respective countries or states.

The list of top 20 influential journals publishing documents on deposit insurance (Table 3) discloses that although the *Journal of Banking and Finance* has the greatest number of citations, owing to its large number of publications in the domain, the *Journal of Financial Economics* and the *Journal of Finance* are relatively more influential in terms of ACPD, thereby leading to a higher CiteScore as well. It is also clear that the avenues of publications in deposit insurance can be categorized into three major groups: journals on finance and economics (including banking), review-based journals and journals on economic history. Additionally, the publisher for more than 50% of these influential journals is Elsevier, followed by Springer Nature and Wiley-Blackwell. The range of noteworthy journals is quite expansive as it covers journals which had their origin in the early 1940s, to the journals which were listed in Scopus in 2000s.

**Table 2.** The 20 Most Influential Authors Publishing Deposit Insurance Articles Ranked by Citation

| Rank | Author Name | Author Affiliation | TC | NOD | ACPD | H-index |
|---|---|---|---|---|---|---|
| 1 | Asli Demirgüç-Kunt | World Bank | 2298 | 9 | 255.33 | 126 |
| 2 | Luc Laeven | European Central Bank | 1638 | 8 | 204.75 | 91 |
| 3 | Edward J. Kane | Boston University, USA | 1075 | 23 | 46.74 | 61 |
| 4 | James R. Barth | Auburn University, USA | 1067 | 14 | 76.21 | 60 |
| 5 | Gary Gorton | Yale University, USA | 883 | 6 | 147.16 | 70 |
| 6 | Charles W. Calomiris | Columbia University, USA | 567 | 9 | 63 | 66 |
| 7 | J. H. Boyd | University of Minnesota, USA | 467 | 6 | 77.83 | Not available |
| 8 | George Pennacchi | University of Illinois, USA | 458 | 7 | 65.43 | 37 |
| 9 | F. González | University of Oviedo, Spain | 357 | 5 | 71.4 | 68 |
| 10 | David C. Wheelock | Federal Reserve Bank of St. Louis, USA | 336 | 9 | 37.33 | 45 |
| 11 | B. D. Smith | University of Texas-Austin, USA | 205 | 7 | 29.28 | Not available |
| 12 | Jin Chuan Duan | National University of Singapore, Singapore | 188 | 5 | 37.6 | 39 |
| 13 | C. W. Sealey | University of North Carolina at Charlotte, USA | 152 | 6 | 25.33 | 12 |
| 14 | Joao A. C. Santos | Federal Reserve Bank of New York, USA | 115 | 5 | 23 | 33 |
| 15 | M. Kabir Hassan | University of New Orleans, USA | 111 | 5 | 22.2 | 72 |
| 16 | J. P. Niinimäki | BOFIT, Bank of Finland, Finland | 86 | 7 | 12.28 | Not available |
| 17 | Min-Teh Yu | Providence University and National Tsing Hua University, Taiwan | 75 | 6 | 12.5 | 22 |
| 18 | D. G. Mayes | University of Auckland, New Zealand and University of Oslo, Norway | 62 | 8 | 7.75 | Not available |
| 19 | Sumon C. Mazumdar | LECG Corporation, California, USA | 42 | 7 | 6 | 4 |
| 20 | George G. Kaufman | Loyola University and Federal Reserve Bank of Chicago, USA | 38 | 7 | 5.43 | 58 |

**Source:** The authors' analysis based on Scopus database.
**Notes:** TC = Total citations; NOD = Number of documents; ACPD = Average citations per document.
H-index of author is as per Google Scholar profile.

**Table 3.** The 20 Most Active Journals Publishing Deposit Insurance Articles Ranked by Citations

| Rank | Journal Name | Publisher | Coverage in Scopus | TC | NOD | ACPD | Cite Score (2020) |
|---|---|---|---|---|---|---|---|
| 1 | *Journal of Banking and Finance* | Elsevier | 1977–ongoing | 5,750 | 114 | 50.49 | 4.4 |
| 2 | *Journal of Finance* | Wiley-Blackwell | 1946–ongoing | 3,549 | 22 | 161.32 | 11.2 |
| 3 | *Journal of Financial Intermediation* | Elsevier | 1990–ongoing | 2,229 | 23 | 96.91 | 6.3 |
| 4 | *Journal of Financial Economics* | Elsevier | 1974–ongoing | 1,514 | 8 | 189.25 | 9.6 |
| 5 | *Journal of Monetary Economics* | Elsevier | 1975–ongoing | 1,492 | 22 | 67.82 | 4.3 |
| 6 | *Journal of Financial Services Research* | Springer Nature | 1987–ongoing | 806 | 46 | 17.52 | 3.3 |
| 7 | *Journal of Money, Credit and Banking* | Wiley-Blackwell | 1996–ongoing | 659 | 21 | 31.38 | 2.2 |
| 8 | *Journal of Financial and Quantitative Analysis* | Cambridge University Press | 1966–ongoing | 529 | 7 | 75.57 | 5.3 |
| 9 | *Journal of Financial Stability* | Elsevier | 2005–ongoing | 481 | 28 | 17.17 | 6.3 |
| 10 | *Review of Finance* | Oxford University Press | 2004–ongoing | 413 | 5 | 82.6 | 6.6 |
| 11 | *Journal of Economic History* | Cambridge University Press | 1941–ongoing | 382 | 6 | 63.66 | 4.1 |
| 12 | *European Economic Review* | Elsevier | 1969–ongoing | 333 | 5 | 66.6 | 3.0 |
| 13 | *Review of Quantitative Finance and Accounting* | Springer Nature | 1991–ongoing | 237 | 6 | 39.5 | 2.6 |
| 14 | *Journal of International Money and Finance* | Elsevier | 1982–ongoing | 185 | 8 | 23.12 | 3.9 |
| 15 | *Quarterly Review of Economics and Finance* | Elsevier | 1992–ongoing | 179 | 18 | 9.94 | 3.2 |
| 16 | *Financial Markets, Institutions and Instruments* | Wiley-Blackwell | 2003–ongoing | 135 | 9 | 15 | 0.9 |
| 17 | *International Review of Economics and Finance* | Elsevier | 1992–ongoing | 129 | 11 | 11.73 | 3.6 |
| 18 | *Journal of Banking Regulation* | Springer Nature | 2008–ongoing | 127 | 25 | 5.08 | 0.8 |
| 19 | *Journal of International Financial Markets, Institutions and Money* | Elsevier | 1997–ongoing | 121 | 9 | 13.44 | 5.2 |
| 20 | *Explorations in Economic History* | Elsevier | 1969–ongoing | 118 | 5 | 23.6 | 2.9 |

**Source:** Authors' analysis based on Scopus database.
**Note:** TC = Total citations; NOD = Number of documents; ACPD = Average citations per document.

Coming to the top 20 countries with the highest citations and research influence in deposit insurance (Table 4), the USA is ranked 1 since it accounts for about 42% of the total publications in the domain, followed by the United Kingdom (UK). While Canada, France and China have a relatively higher number of publications, the European countries of Belgium, Netherlands and Spain are way ahead in terms of impactful publications measured through TC, irrespective of research quantity.

Out of the 77 countries publishing in the area of deposit insurance, these 20 most impactful countries account for about 90% of the publications. A careful observation seems to point towards a possible tandem in the movement of the age of the deposit insurance scheme and average publication year of these countries. This is logical since a regulatory introduction or adoption is often followed by its academic as well as research evaluations. However, at the same time, one can observe a clustering of average publication years around 2010 for many countries, because of a boom in the number of publications after the GFC, making the years 2009–2011 as the epicentre for many financial and banking research developments. This is also indicative of the frequent policy amendments and upgradations undertaken by countries in their deposit insurance schemes, keeping in mind the changing banking dynamics.

Table 4 also shows that China and Poland are the countries where deposit insurance research has recently started picking pace (average publication year: 2016), along with some other developing countries such as Saudi Arabia, Malaysia, Ukraine, Pakistan, Tunisia and Lithuania. As an exceptional case, Israel, a country with no explicit deposit insurance scheme, features among the top 20 influential countries because of the interest of scholars and policymakers in its successfully implemented implicit insurance scheme.

## Co-authorship Analysis

Co-authorship analysis helps to interpret the collaborations and intellectual networks formed amongst authors of

**Table 4.** The 20 Most Active Countries Publishing Deposit Insurance Articles Ranked by Citations

| Rank | Country | TC | NOD | Year of First Established Explicit Deposit Insurance Scheme | Average Publication Year | Percentage of Total Publications |
|---|---|---|---|---|---|---|
| 1 | USA | 20,273 | 554 | 1933 | 2005.3 | 41.9 |
| 2 | United Kingdom | 3,209 | 126 | 2000 | 2009.9 | 9.53 |
| 3 | Belgium | 2,177 | 18 | 1974 | 2011.2 | 1.36 |
| 4 | Netherlands | 2,124 | 28 | 1978 | 2011.4 | 2.12 |
| 5 | Spain | 990 | 33 | 1977 | 2009.8 | 2.5 |
| 6 | Australia | 749 | 40 | 2008 | 2010.8 | 3.02 |
| 7 | Canada | 734 | 52 | 1967 | 2009 | 3.93 |
| 8 | Hong Kong | 663 | 15 | 2004 | 2010.4 | 1.13 |
| 9 | Germany | 565 | 39 | 1976 | 2014.2 | 2.95 |
| 10 | France | 559 | 46 | 1980 | 2010.5 | 3.48 |
| 11 | Italy | 479 | 39 | 1987 | 2013.9 | 2.95 |
| 12 | Israel | 323 | 12 | No explicit deposit insurance | 2010.5 | 0.9 |
| 13 | Finland | 306 | 24 | 1970 | 2010 | 1.82 |
| 14 | Taiwan | 305 | 37 | 1983 | 2010 | 2.8 |
| 15 | Switzerland | 298 | 20 | 2005 | 2011.8 | 1.51 |
| 16 | China | 221 | 46 | 2015 | 2016.2 | 3.48 |
| 17 | Indonesia | 174 | 20 | 2004 | 2015.9 | 1.51 |
| 18 | Poland | 159 | 13 | 1995 | 2016.4 | 0.98 |
| 19 | Turkey | 151 | 17 | 1983 | 2010.7 | 1.28 |
| 20 | South Korea | 147 | 19 | 1986 | 2010 | 1.43 |

**Source:** The authors' analysis based on Scopus database.
**Note:** TC = Total citations; NOD = Number of documents.

numerous countries, origins and organizations in a field of study (Cisneros et al., 2018). As the literary curiosity for research is progressing around the globe, interactions among scholars are ubiquitous (Donthu et al., 2021). It has created spaces for them to exchange ideas, insights and theories correlating to advances in research. Through co-authorship analysis, we have examined the collaborative efforts of authors through the years explicating the path of academic development and highlighting the social structure in the said field of research.

As far as co-authorship among authors is concerned, out of 1963 authors, 88 have at least one collaboration with another/other author/s. The co-authorship network map (Figure 4) not only shows the prominent authors in terms of collaborations but also brings out the clusters or connections to which the authors belong.

James R. Barth from Auburn University, USA, has the most collaborations, while being an author/co-author of the second highest number of publications. Edward J. Kane, affiliated with Boston University, USA, has been engaged in the second highest number of collaborations while publishing the highest NOD (as depicted by the size of the node in Figure 4 as well as Table 2). Asli Demirgüç-Kunt and Luc Laeven follow suit in terms of collaborative efforts. A closer look at the clusters reveals that collaborations are somewhat restricted among the seasoned authors. For example, three of the most impactful authors, Asli Demirgüç-Kunt, Luc Laeven and Edward J. Kane, are part of a close group of collaborators (indicated by purple colour). At the same time, while we see multiple clusters portraying groups of collaborations, majority of the authors among these groups belong to the same country or to countries in close proximity, which limits the knowledge sharing to some extent.

Figure 5 provides an overview of the collaboration structure among academicians across the countries, with the thickness of the lines indicating the volume of collaborations and darkness of country's colour indicating the relative frequency of inter-country collaborations.

It sheds light on the fact that the USA is at the forefront not only in terms of quantity and quality of publications but also in building upon social structure and intellectual knowledge through collaborations outside the USA. Scholars from the USA have engaged in inter-country collaborations for 137 publications, with authors from 26 countries, both developed as well as developing. The highest concentration for collaboration is among the USA and Canada because of proximity and similar financial and banking structures, and the USA and UK due to the similarities of the banking shocks and crises faced by these two economies. Although the collaboration patterns seem to be spread across the globe, a closer look at the numbers reveal that out of top 20 collaborating countries, more than 90% are among developed nations.

As we have progressed from 2010 to 2018, researchers from countries like Poland and Saudi Arabia are making efforts to collaborate with other countries. This indicates the growing ambit of intellectual evolution, research

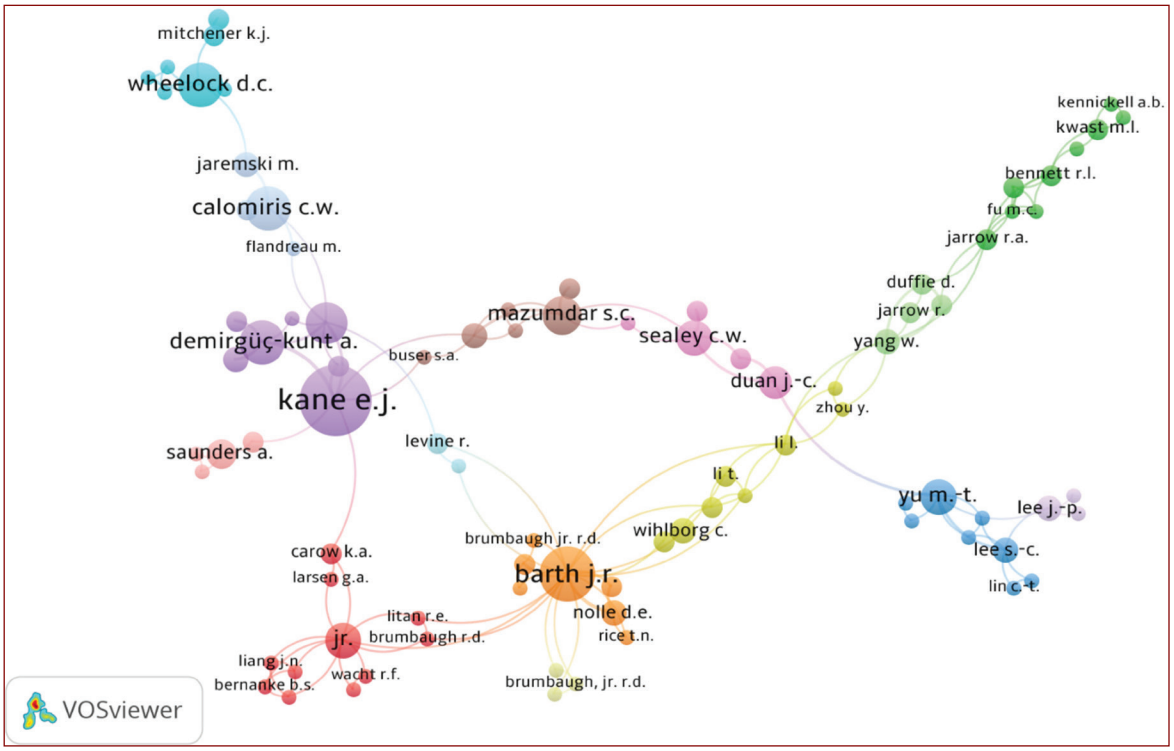



**Figure 4.** Co-authorship Network Among the Authors Publishing on Deposit Insurance

**Source:** The authors (through VOSviewer).

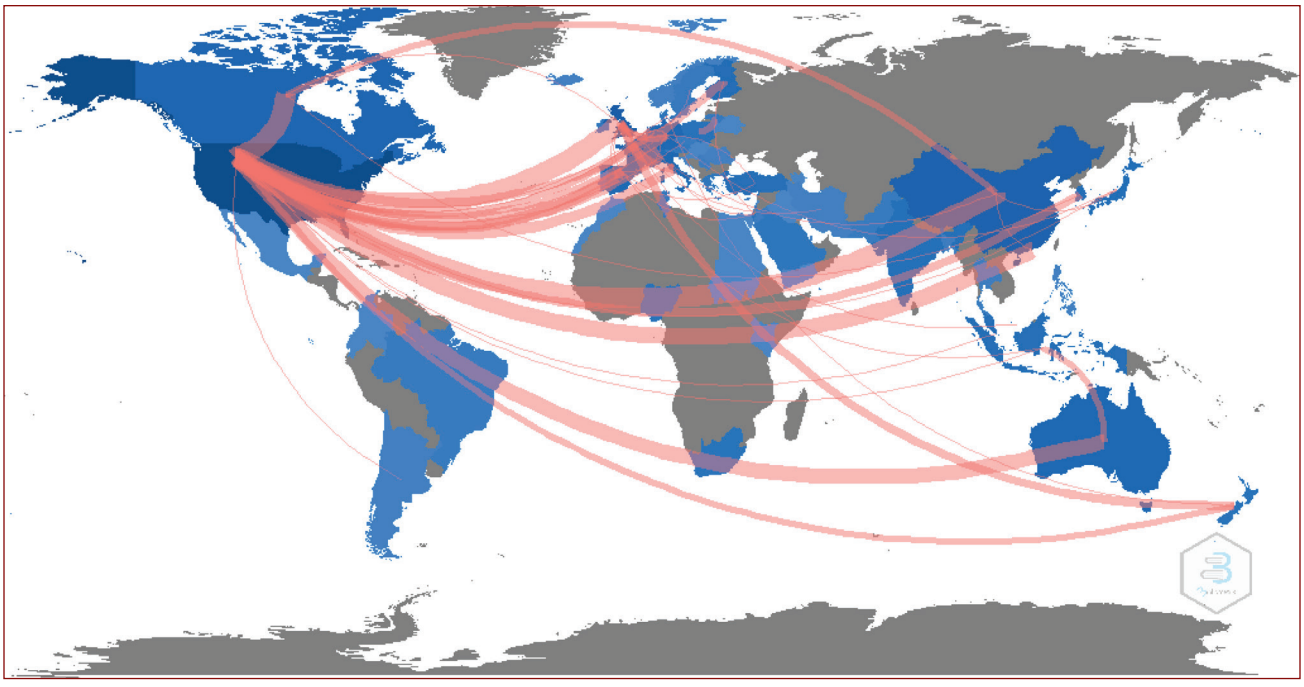


**Figure 5.** Overview of Co-authorship Among Countries Publishing on Deposit Insurance

**Source:** The authors (through Biblioshiny).

collaboration and the attention to deposit insurance across countries through the years. There is a need for academicians from developed countries to explore the banking dynamics and deposit insurance design features of developing countries as well, in order to establish a wider network of teamwork amongst countries, while equipping prospective researchers to have a connection with established and trending scholars in our research field.

## *Co-citation Analysis*

Co-citation analysis is employed to uncover the intellectual framework of a field of research and discern the various emerging themes within it (Donthu et al., 2021; Rossetto et al., 2018). One of the advantages of using this analysis is that it aids researchers to identify inter-relationships (Bhaiswar et al., 2021) and thematic clusters based on the referenced documents. It also assists scholars to determine the most impactful research papers, original publications and academic foundations (Donthu et al., 2021).

In the network map of Figure 6, the nodes depict a cited reference, and the size of the nodes indicate the NOD in which they have been co-cited. Out of 41,055 cited references, 185 meet the threshold of co-cited references with at least five citations each. The linking lines between various nodes represent a co-citation.

We can recognize five distinct clusters on the network map, and a content analysis of the documents from each cluster reveals the following themes represented by the clusters:

### *Cluster 1: Rational and Design Features of Deposit Insurance*

Cluster 1, represented by red colour, is the largest cluster that focuses on the rational and need for establishing a system of deposit insurance. The banking panics, triggered by recession (Gorton, 1988), rumours or other external factors have a detrimental impact on the bank liquidity. The 'sequential service constraint' faced by banks (Diamond & Dybvig, 1983) intensifies the bank runs in such situations, leading to self-fulfilling insolvency issues (Calomiris, 1990; Peck & Shell, 2003). The linkages between banks can have a snowball effect because of contagion, which can ultimately turn into a banking and financial crisis (Allen & Gale, 1998, 2000; Saunders & Wilson, 1996). This is where deposit insurance steps in. An assurance to the depositors that their deposits are safe reduces their incentives to run on banks (Gatti & Oliviero, 2018).

The articles in this cluster also discuss the design features of deposit insurance that can minimize the problems associated with such a scheme (Bryant, 1980). This includes the private administration of deposit insurance schemes vis-à-vis federal administration (Stone, 2021), compulsory or voluntary membership to the scheme (Chiaramonte et al., 2020; Lee & Kwok, 2000), extent of coverage (Sardana & Shukla, 2020; Shy et al., 2016), the broadness of the insurer's mandate (IADI, 2021) and the combination of deposit insurance with reinsurance (Britz et al., 2021). Chang et al. (2018) finds that the political systems, the law and statutes, bank structure and macroeconomic development in a country are the major determinants of the extent of optimal insurance coverage.

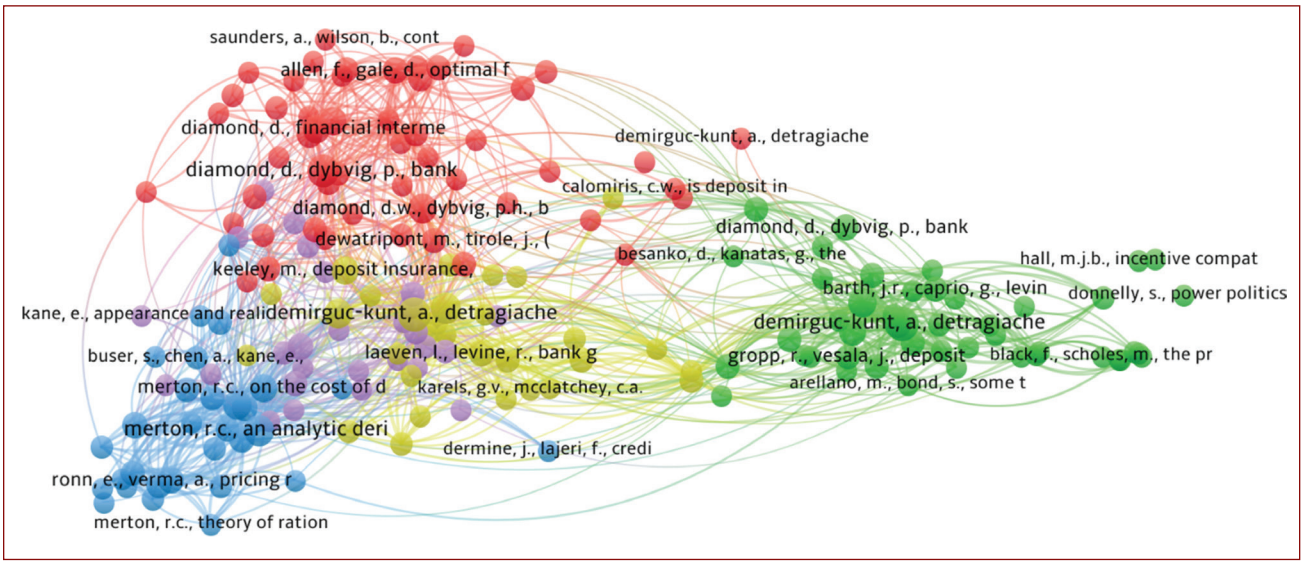



**Figure 6.** Co-citation Network Based on Cited References with Minimum Five Citations

**Source:** The authors (through VOSviewer).

### *Cluster 2: Deposit Insurance in the Context of Bank Risk and Financial Stability*

The second largest cluster (green colour) has an extensive number of empirical articles that determine the regulatory implications of deposit insurance on bank risk-taking and financial stability. The cluster consists of both cross-country studies and country-wise case studies, with mixed evidences.

In normal times, a generous deposit insurance may make the banking system fragile, thereby enhancing bank risk, while promoting stability during a crisis period (Anginer et al., 2014). Furthermore, financial integration among countries, such as the Eurozone, can intensify the spillover of deposit insurance linked banking instability (Dubois, 2021). In such a situation, harmonization of deposit insurance features at regional level can promote both stability and welfare (Bodellini, 2020; Chan et al., 2018), but at the same time can lead to cross-subsidization among the stronger and weaker nations (Kuznichenko et al., 2021).

Cross-country evidence also suggests that in the presence of government-administered deposit protection, the likelihood of a banking crisis increases (Demirgüç-Kunt & Detragiache, 2002). Laeven and Levine (2009) find that the impact of deposit insurance on bank risk is also dependent on banks' governance and ownership structure. It has been emphasized that adopting credible coverage limits, co-insurance and risk-based premium minimizes the destabilizing effect of deposit insurance (Barth et al., 2004; Demirgüç-Kunt & Kane, 2002). Additionally, the presence of a strong regulatory and supervisory environment in a country can enhance the stabilizing functions of an explicit deposit insurance (Anginer et al., 2014; Demirgüç-Kunt et al., 2008).

### *Cluster 3: Pricing of Deposit Insurance*

The blue cluster, depicting cluster 3, is made up of research papers that focus on valuation and pricing of deposit insurance. Merton (1977) laid the foundation by deriving the value of insurance using the option pricing theory, which has been extended to multiple deposit insurers such as the Federal Deposit Insurance Corporation (Marcus & Shaked, 1984). This has been further built upon by modelling deposit insurance as a callable put option (Allen & Saunders, 1993), as an Israeli put option (Wong et al., 2020), and empirical estimation of the insurance premium using market data (Ronn & Verma, 1986). Duffie et al. (2003) suggests the use of a risk-neutral deposit insurance valuation model for calculation of fair premia, using the debt process and probability of bank failure. In light of the various valuation models, under-pricing and over-pricing of a country's deposit insurance premium can be revisited and examined (Pennacchi, 1987).

Putri et al. (2021) suggests the adoption of multi-state regime-switching volatility in the insurance premium pricing, as it would account for the dynamism and volatility in the banks' assets. A risk-based structure for pricing of deposit insurance premium, as adopted by multiple countries, can be employed to calculate fair premium on the basis of credit derivatives (Wu et al., 2020). Various deposit insurance valuation models have been derived that take into account the interest rate spread of banks, risk of bankruptcy (Chiang & Tsai, 2020), asymmetric bankruptcy costs (Wong, 2020) and contingent capital bonds (Chang et al., 2022).

### *Cluster 4: Market Discipline and Moral Hazard in the Presence of Deposit Insurance*

Cluster 4 (yellow colour) consists of articles that provide empirical evidence for the presence or absence of market monitoring and moral hazard following the adoption of deposit insurance. The two concepts are the consequences associated with any insurance contract, but they tend to magnify in deposit insurance because of the presence of three parties: banks, depositors and insurers (Gupta & Sardana, 2021). There is ample evidence on the negative effect of deposit insurance on the discipline exercised by the market as the depositors become complacent (Calomiris & Jaremski, 2019; Demirgüç-Kunt & Huizinga, 2004; Vernikov, 2020). At the same time, studies from certain countries suggest that deposit insurance does not undermine market discipline because of strict and credible insurance limits (Imai, 2006; Martinez-Peria & Schmukler, 2001; Trinugroho et al., 2020). Moral hazard is an issue that leads to an increased risk taking on part of bank managers (Laeven & Levine, 2009; Sahadewo et al., 2018). Deposit insurance incentivizes the bankers to indulge in risky activities in hope for better returns (Kusairi et al, 2018). This problem intensifies in the presence of a fixed rate of premium (Keeley, 1990), instead of a risk-based premium structure.

The problem of moral hazard, due to the establishment of deposit insurance scheme, may not hold true for credit unions with already lenient capital regulations (Karels & Mcclatchey, 1999). Assa and Okhrati (2018) suggest that the issue of moral hazard in deposit insurance can be resolved by developing a multi-layered optimal policy. Another means to mitigate this issue is by extending the liability of default to bank shareholders rather than relying on external restrictions (Salter et al., 2017).

### *Cluster 5: Deposit Insurance and Other Safety-nets*

The smallest cluster, depicted by purple colour, is cluster 5. It consists of articles that discuss other types of bank safety nets, in addition to deposit insurance. Prudential regulation in the form of bank capital requirement (Ashraf et al., 2020; Kim & Santomero, 1988) and banking supervision (Mumtaz et al., 2019) can be one way to control the bank risk in the presence of deposit insurance. However, in certain cases, it can also add to the financial instability (Gennotte & Pyle, 1991; Besanko & Kanatas, 1996) and can lead to pareto-inefficient outcomes (Hellmann et al., 2000). Furthermore, when bank examination procedures

and capital adequacy regulations come into play, insured banks may change their preferences from risky assets to relatively safe assets, even though they may still opt for risky portfolio returns (Flannery, 1989).

While deposit insurance acts as a good safety net in a less competitive banking environment, regulations on activity restrictions are more effective for a highly competitive environment (Noman et al., 2018). Nayak (2021) compared different regulatory guidelines among countries and found that stringent capital requirements, along with entry barriers and external monitoring, can strengthen the risk management system of insured banks.

## Keyword Co-occurrence Analysis

The keyword co-occurrence analysis postulates that two terms appearing in a paper together often have some conceptual relation to one another, form thematic clusters and indicate the temporal developments (Kumar et al., 2021). Academicians also use this method of scientific mapping to predict forthcoming concepts in their field of research, which helps in formulating various theoretical implications for the future (Donthu et al., 2021). Therefore, the co-occurrence of keywords provides us with a prescience into the upcoming intellectual path of a field.

To conduct a keyword co-occurrence analysis, we first prepared a thesaurus for VOSviewer to ensure that plurals of same words and words with different spellings are clubbed together to avoid duplication. To illustrate a bibliometric map for co-occurrence, the minimum frequency of a keyword was set to five, which led to 89 keywords that met the threshold out of a total of 2,257. The co-occurrence network map in Figure 7 portrays that 'deposit insurance' is the highest occurring keyword in our 1,322 documents, with a frequency of 305.

Although various alternative versions of this safety net scheme are used across countries, 'deposit insurance' is the most common because it is portrayed as an explicit insurance provided by the government or a statutory body to its depositors. The words 'bank regulation', 'banking' and 'bank' follow suit, since deposit protection is meant primarily for bank depositors and has now become an integral part of the banking regulation, though some countries have extended this scheme to securities firms, financial cooperatives and other deposit-taking institutions as well (IADI, 2021). One of the main objectives of such a protection is to provide a safeguard to the monies of depositors through a system of insurance, so as to reduce the probability of bank runs and the resultant financial shocks (Gatti & Oliviero, 2018; Putri et al., 2021). Hence, 'financial crisis', 'insurance system' and 'bank runs' are also among the top 10 frequently employed keywords in the literature. 'Moral hazard' and 'market discipline' are the two concepts that have been extensively researched

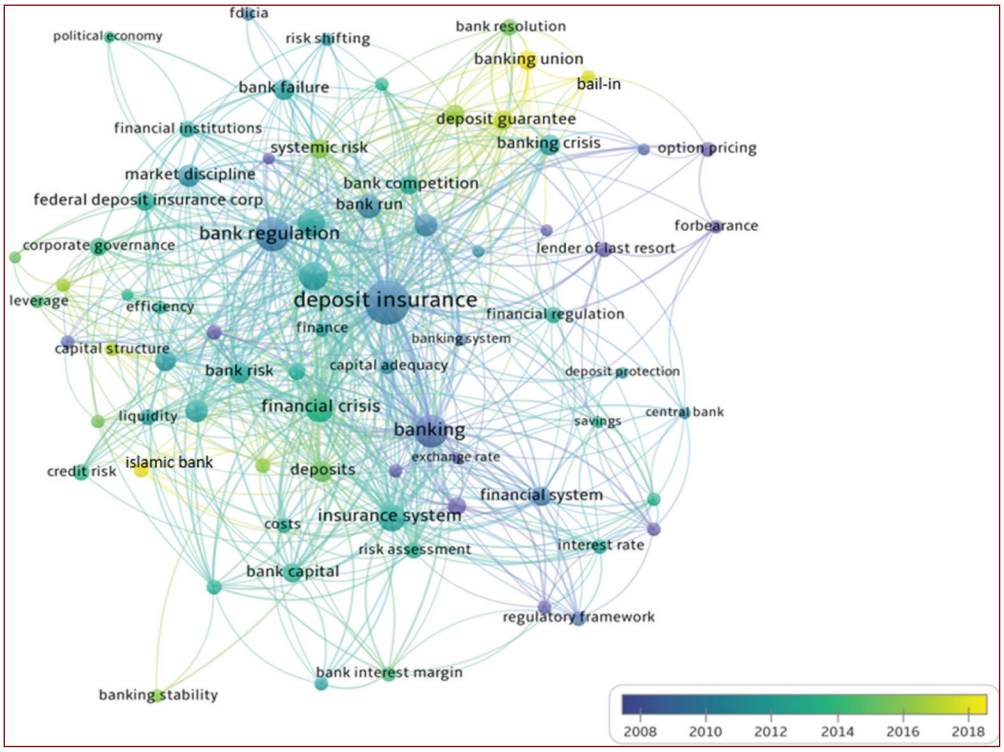



**Figure 7.** Keyword Co-occurrence Network on Deposit Insurance (Threshold of Five Occurrences)

**Source:** The authors (through VOSviewer).

in the deposit insurance literature, since these are the common fallouts of an insurance contract (Chi & Binh, 2020; Demirgüç-Kunt & Huizinga, 2004; Kusairi et al., 2018; Ngalawa et al., 2016; Zhang & Wu, 2020).

Figure 7 also provides a bird's eye view of the evolution of the prominent keywords after the emergence of the GFC. A shift in focus from foundational terms to concepts related to banking crisis, supervision and revival solutions has been witnessed in the couple of years following the GFC. Usage of keywords such as 'bank supervision', 'regulation', 'prompt corrective action', 'banking crisis', 'lender of last resort', 'risk-taking', 'bank failure' and 'FDICIA'[1] is centred around the periods immediately following the crisis (Figure 8), as researchers tried to capture the implications of the global meltdown of markets and institutions on safety nets and vice versa.

As research in this domain has progressed, new concepts and topics are being explored by academicians (Figure 8). Some of the trending topics in the domain are 'credit risk' and 'bank risk-taking', both of which have been growing in the banking domain and can call for trouble. The extension of deposit insurance to Islamic banks by some countries such as the UK, Afghanistan, Bahrain, Bangladesh, Iran, Indonesia, Malaysia and Philippines (IADI, 2021) has received much attention and has led to the creation of specific guidelines such as the 'Core Principles for Effective Islamic Deposit Insurance System' by the International Association of Deposit Insurers (IADI) and Islamic Financial Services Board (IFSB; IADI & IFSB, 2021). Hence, 'Islamic banks' has been a widely used term in the past couple of years. With many countries still relying on implicit assurances and guarantee of the government for the safety of their deposits, rather than having an explicit insurance, the word 'deposit guarantee' has also been in circulation. Further, the concepts of 'banking union', 'capital structure' and 'bail-in' are being tapped in the area, apart from the evergreen terms 'banking stability' and 'financial stability'. While the presence of banking unions can have an impact on the perception about deposit insurance, capital structure norms and bail-in clause have ramifications for the resolution of troubled banks. These regulations are being promoted by the Financial Stability Board (FSB) to minimize the loss to taxpayers and depositors (FSB, 2017; Lewrick et al., 2019). Figures 7 and 8, thus, highlight certain concepts that have been linked to deposit insurance in recent years and, therefore, have limited research addressing them.

## Discussions and Future Scope for Research

Through the bibliometric analysis, this study answers the research questions posed and provides an overview of the research in the field of deposit insurance. The volume analysis over the years shows that research publications have a secular upward trend ($RQ_1$). Since a deposit insurance

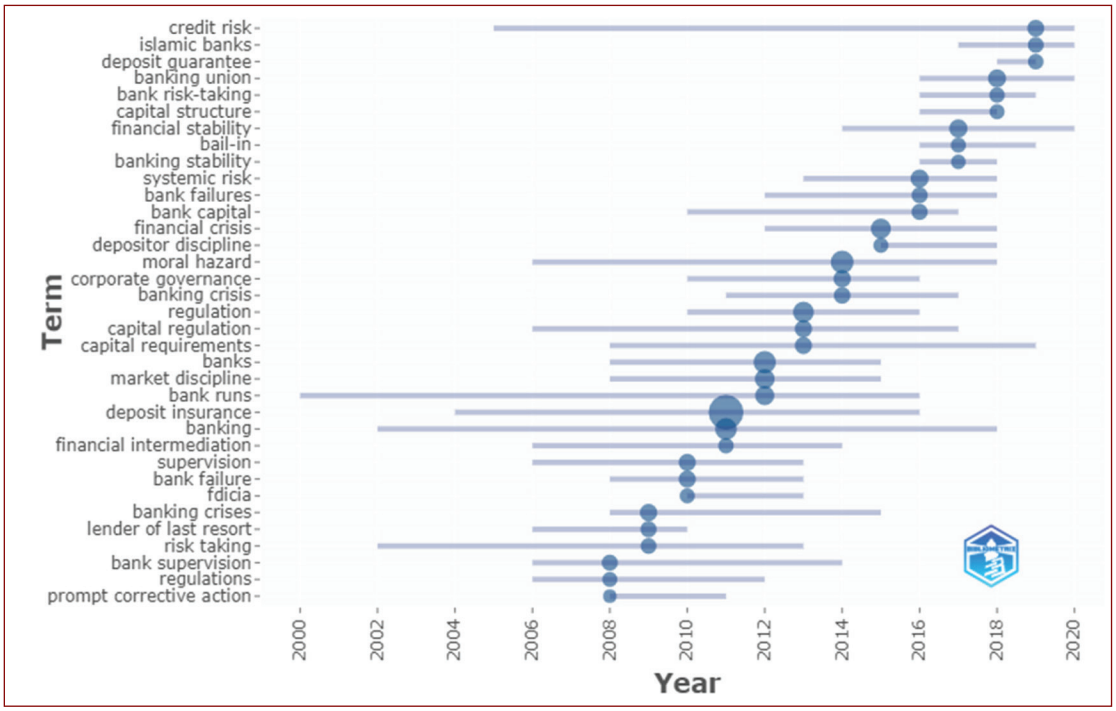



**Figure 8.** Trend Topics over the Years (Minimum Term Frequency: 5; No. of Words Per Year: 3)

**Source:** The authors (through Biblioshiny).

scheme aims to ensure banking stability, any external financial shock acts as a cue for scholars to study the impact of these shocks and policy changes in the presence of such insurance, both within and across countries. Although the concept of deposit insurance is quite old, the research output in this area has not yet reached saturation, and it hit a new high in 2019. With the world facing a health crisis since more than 2 years, the banking activities have been severely affected, and this provides new opportunities for research in the field. In terms of geographical distribution of research output, the USA is leading in both the quantity and quality of publications, with the UK and European Union following suit ($RQ_1$). These countries have witnessed multiple banking crises over the years, which have provided a fertile ground for researchers to study how the financial institutions and depositors react to various banking regulations. These studies have also set the stage for similar research in other countries. Some Asian countries have also fared well in research on deposit insurance, especially since their exposure to the Asian Financial Crisis of late 1990s.

As far as most influential research in the area is concerned ($RQ_2$), the 20 most cited papers are from as early as 1977 to as recent as 2013, and they bring to light certain seminal works that have laid the foundation of various concepts in the field of deposit insurance. These influential papers are mostly conceptual studies or empirical studies in the international context, which have helped to shape the present research. The most seasoned and influential authors in this domain are affiliated with global organizations, such as the World Bank and the European Central Bank, and other institutions of repute. Such affiliations allow access to otherwise unavailable data, as well as knowledge-sharing with heads of deposit insurers across countries. The contributions of such authors have become highly valuable in building multiple concepts and enabling robust cross-country studies on deposit insurance. Furthermore, the majority of authors with substantial research contributions are from the USA, which supports our findings from the volume analysis across nations. As deposit insurance is a niche area of banking, the research studies in this domain have a narrow set of discipline-specific avenue of publications, spanning to finance, economics, banking and economic history. Also, since a major portion of publications come from developed countries, the citations from these countries are high, making them the most influential nations in the deposit insurance literature. A few developing countries like China and Indonesia also feature among the top 20 influential countries due to the meaningful contributions of their authors.

The social structure among authors ($RQ_3$) shows an emphasis on intra-country collaborations. Researchers from different institutions in same country have led the pattern of collaborations, followed by inter-country collaborations among authors from developed countries such as the USA, the UK and Canada. There is a tendency of authors to form groups based on similarity of school of thought, and to collaborate within those groups. Furthermore, hardly any cross-country collaborations have been observed among the countries of South America and Africa. Among the developing nations, the cross-country collaborations are restricted to China, Indonesia and Lithuania to some extent. There are five thematic clusters that are brought to light through the co-citation analysis ($RQ_4$), and they have been reviewed through content analysis. These themes, which have built the literature of deposit insurance, extend to the rationale and design of deposit insurance, bank runs and financial stability, pricing of deposit insurance schemes, role of market discipline and moral hazard, and safety nets in addition to deposit guarantees. The keyword co-occurrence analysis further highlights the prominent keywords that are consistent with our thematic sets derived through co-citation analysis ($RQ_5$). The temporal trend reveals that fundamental terms such as 'deposit insurance', 'bank regulation' and 'financial crisis' have evolved over a period of time, and new concepts have emerged, especially after the global financial crisis, focusing on the revival of the banking system through regulatory modifications. In the last five years, certain novel topics have come into picture, such as the extension of deposit insurance to Islamic banks, introduction of bail-in clause and optimal capital structure, and the presence of banking unions—all of which impact or may be impacted by deposit insurance and breathe fresh ideas into the literature.

## Future Research Avenues

The bibliometric analysis and review has brought forward certain gaps and opportunities for contributing to the literature, which can be addressed by researchers in the future. These are as follows:

### *Contributions from Developing Countries*

More than 70% of research studies in the field of deposit insurance are from North America and Europe, especially from developed nations. While some of the research implications from the studies of developed countries can be useful for emerging economies, the regulations and strengths of the banking system as well the resilience of the economy varies a lot across developing countries (Demirgüç-Kunt & Kane, 2002; Barth et al., 2004). Hence, there is a need to undertake research on deposit insurance systems, their design, implications and consequences in more developing countries, taking into consideration their institutional and supervisory structures. Collaborations among authors from developed and developing nations are also limited. Such collaborations, if undertaken, would allow the extension of ideas (Tahamtan et al., 2016) and the building of globally accepted best practices of deposit guarantees as well.

### The Facility of Reinsurance

Past literature on the design features of deposit insurance have studied important characteristics like the coverage limit of guarantee (Chen et al., 2021; Chiaramonte et al., 2020; Lee & Kwok, 2000), type of premium charged (Chang & Ho, 2017; Keeley, 1990; Wu et al., 2020), the method of funding (Liu et al., 2016; Lazarov, 2017), membership of the insurance scheme (Lee & Kwok, 2000) and benefits of co-insurance (Boyle et al., 2015; Stern, 1998). However, reinsurance as a policy initiative in the context of deposit insurance has not received much attention. While Britz et al. (2021) advocate deposit insurance with the facility of reinsurance, and Kane (2006) suggests the issuance of reinsurance derivatives for enhancing a country's insurance capabilities, the extension of the reinsurance feature, its feasibility, rationale and its forms need to be explored to highlight the policy implications and possible benefits.

### Studies Inculcating Depositors' Behaviour

A handful of studies have used primary data to study the element of human behaviour in the context of deposit insurance (Bijlsma & Van Der Wiel, 2015; Boyle et al., 2015). The perception of depositors, their awareness, past experiences with bank failures and their faith levels play a major role in the success of such a depositor protection scheme. Distrust in the system on part of the beneficiaries may make the entire system useless (Prean & Stix, 2011), no matter how generous the insurance coverage limits are. Hence, researchers may need to redirect efforts towards studies that take the human element of depositors into account.

### Issue of Cross-subsidization

The counteractive consequences of adopting deposit insurance, such as moral hazard (Gupta & Sardana, 2021; Kusairi et al., 2018; Zhang & Wu, 2020), declining depositor's discipline (Calomiris & Jaremski, 2019; Vernikov, 2020) and banking instability (Anginer et al., 2014; Demirgüç-Kunt et al., 2014), have been studied extensively. But one consequence that has received little attention is the issue of cross-subsidization. Subsidization of insurance facilities from strong countries to weak countries has been highlighted by Carmassi et al. (2021) apropos European Union, while Freixas and Rochet (1998) elaborated on the subsidization of less efficient banks by the most efficient ones. However, the study of deposit-insurance-based cross-subsidization within countries, along with its implications, is missing, and needs to be factored in by researchers.

### Deposit Insurance Fund

The deposit insurance fund, which is created for paying the claims of insured depositors in case of a bank failure, is an area that demands in-depth research. The past studies focus on the choice between ex-ante, ex-post and hybrid source of funding (Chang & Ho, 2017; Garcia, 1996; Mohammad & Nishiyama, 2019), the administration of the fund (Kleftouri, 2015; Morrison & White, 2011) and the target ratios for such a fund (IADI, 2018). However, research on the adequacy and solvency of the fund as well the fund modelling across countries for future viability is required to understand it better.

### Impact of New Banking Norms

The effectiveness of a deposit insurance system is influenced by the regulations, laws and statutes put in place by a country. Though few studies have attempted to capture the impact of prudential regulations such as bank supervision, external monitoring and capital requirements (Ashraf et al., 2020; Kim & Santomero, 1988; Mumtaz et al., 2019), there is a need to explore this area in detail, especially with the onset of new types of bank control measures such as bail-in clause, capital structuring norms and banking unions.

### Studies Using Bank Micro-level Data

There is a paucity of bank-level studies that aim to understand the reaction of bankers and depositors during times of bank stability and bank instability or during single or multiple banks' insolvency. Though efforts have been made to track the depositors' run on banks during a solvency risk (Iyer & Puri, 2012; Iyer et al., 2016), the unavailability or inaccessibility of bank-level data acts as an impediment to such studies. Research organizations, central banks or regulators in possession of such data can provide great insights into this dimension of deposit insurance, which would have further implications on the usefulness of such a safety net when most needed.

## Academic and Practical Implications

This study has academic implications for researchers and practical implications for deposit insurers, banks and policymakers. For research scholars and academicians, the current study would help to understand the development of literature, the spatial and temporal trend, the evolution of essential themes and future research avenues in the field of deposit insurance. The study of the various clusters highlighted by the study can enable them to narrow down their research objectives, while the visualization through network diagrams would allow the identification of seminal works and studies that have made enduring contributions to the literature. An in-depth study of these documents would help in the recognition of the origins of the various themes in this research area, along with their theoretical foundations (Zheng & Kouwenberg, 2019). For deposit

insurers and banks, the study throws light on what can be expected next—from extending deposit guarantees to diverse categories of banks, to dealing with the bail-in or capital structure requirements. These evolving research trends would provide a direction to these two institutions and help in better planning of their insurance claim obligations. While the past literature has focused on establishing and extending deposit insurance to deal with bank runs and crises, as well as understanding its implications and externalities, the recent literature places emphasis on improving the robustness of the insurance scheme and adopting best practices to fine-tune the system. This would help policymakers and regulators foresee the changing scenario of the deposit insurance laws.

## Conclusion

Research contributions on the topic of deposit insurance have grown exponentially. This study attempts to capture the dimensions of this growth and makes multiple contributions to the literature. First, it tracks the growth of research and interest in this topic over half a century, across time and geographies. Second, it brings the most influential academicians, countries, research articles and sources to the forefront, while highlighting the seminal works in this area. Third, it reveals the author collaborations and relational ties that have been built across countries, as well as within countries, by bringing researchers together in this area of research. Fourth, it helps the budding researchers or academicians to gain an insight into the various concepts and themes that have emerged and taken shape over a period of time, as well as the new facets of sub-topics that are opening up for further study in the field of deposit insurance. By drawing the overall patterns of research in this topic, our article builds the trajectory and path that this field has taken up and underlines its importance in the broader discipline of banking and finance. The findings from our study also suggest that the research in the field of deposit insurance has not yet reached saturation and that there is scope for exploring multiple dimensions of the same.

Our study has a few limitations. The results of the bibliometric analysis in our study are based on the Scopus database only. In the future, a wider analysis, by including other databases such as Web of Science (WoS), can be undertaken. Further, only three keywords, which are the alternative versions of the main keyword, have been used to retrieve documents from the selected database. Other variations of this keyword or new keywords that relate to this field can be employed in the future. Additionally, other bibliometric tools, such as BibExcel and Gephi, may be employed in future to allow the results to be compatible across other software packages (Persson et al., 2009) as well as data formats (Fahimnia et al., 2015). Nonetheless, being the first of its kind, this study can be regarded as the stepping-stone for understanding the overall intellectual framework of research in deposit insurance and to enable the identification of future research opportunities as well as challenges.

## Declaration of Conflicting Interests

The authors declare no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

## Funding

The authors received no financial support for the research, authorship and/or publication of this article.

## ORCID iD

Varda Sardana https://orcid.org/0000-0003-1326-6273

## Note

1. FDICIA, which stands for Federal Deposit Insurance Corporation Improvement Act, sought to amend and improve the functioning of the deposit insurer of the USA in the light of the global crisis.

## References

Alam, N., Sivarajah, G., & Bhatti, M. I. (2021). Do deposit insurance systems promote banking stability? *International Journal of Financial Studies*, *9*(3), 52.

Allen, F., & Gale, D. (1998). Optimal financial crises. *The Journal of Finance*, *53*(4), 1245–1284.

Allen, F., & Gale, D. (2000). Financial contagion. *Journal of Political Economy*, *108*(1), 1–33.

Allen, L., & Saunders, A. (1993). Forbearance and valuation of deposit insurance as a callable put. *Journal of Banking and Finance*, *17*(4), 629–643.

Altunbaş, Y., & Thornton, J. (2013). Deposit insurance and private capital inflows: Further evidence. *Journal of International Financial Markets*, *Institutions and Money*, *23*, 243–247.

Aly, H. Y., Grabowski, R., Pasurka, C., & Rangan, N. (1990). Technical, scale, and allocative efficiencies in US banking: An empirical investigation. *The Review of Economics and Statistics*, *72*(2), 211–218.

Anginer, D., Demirgüç-Kunt, A., & Zhu, M. (2014). How does deposit insurance affect bank risk? Evidence from the recent crisis. *Journal of Banking and Finance*, *48*, 312–321.

Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975.

Ashraf, B. N., Zheng, C., Jiang, C., & Qian, N. (2020). Capital regulation, deposit insurance and bank risk: International evidence from normal and crisis periods. *Research in International Business and Finance, 52,* 101188.

Assa, H., & Okhrati, R. (2018). Designing sound deposit insurances. *Journal of Computational and Applied Mathematics*, *327*, 226–242.

Atanasova, C., Li, M., Mugerman, Y., & Rastan, M. (2019). Government guarantees and the risk-taking of financial institutions: Evidence from a regulatory experiment. *Journal of Asset Management*, *20*(6), 476–492.

Atmaca, S., Kirschenmann, K., Ongena, S., & Schoors, K. (2021). *Increasing deposit insurance and nationalisation in*

*times of crisis*. https://voxeu.org/article/increasing-deposit-insurance-and-nationalisation-times-crisis

Barth, J. R., Caprio Jr, G., & Levine, R. (2004). Bank regulation and supervision: What works best? *Journal of Financial Intermediation*, *13*(2), 205–248.

Beck, T., De Jonghe, O., & Schepens, G. (2013). Bank competition and stability: Cross-country heterogeneity. *Journal of Financial Intermediation*, *22*(2), 218–244.

Bergbrant, M. C., Campbell, K. T., Hunter, D. M., & Owers, J. E. (2016). Does deposit insurance retard the development of non-bank financial markets? *Journal of Banking and Finance*, *66*, 102–125.

Besanko, D., & Kanatas, G. (1996). The regulation of bank capital: Do capital standards promote bank safety? *Journal of Financial Intermediation*, *5*(2), 160–183.

Bhaiswar, R., Meenakshi, N., & Chawla, D. (2021). Evolution of electronic word of mouth: A systematic literature review using bibliometric analysis of 20 years (2000–2020). *FIIB Business Review*, *10*(3), 215–231.

Bhatia, P., & Gupta, P. (2020). Sub-prime crisis or COVID-19: A comparative analysis of volatility in Indian banking sectoral indices. *FIIB Business Review*, *9*(4), 286–299.

Bijlsma, M., & Van Der Wiel, K. (2015). Consumer perception of deposit insurance: Little awareness, limited effectiveness? *Applied Economics*, *47*(32), 3439–3461.

Bodellini, M. (2020). Alternative forms of deposit insurance and the quest for European harmonized deposit guarantee scheme-centred special administrative regimes to handle troubled banks. *Uniform Law Review*, *25*(2–3), 212–255.

Boyd, J. H., & Runkle, D. E. (1993). Size and performance of banking firms: Testing the predictions of theory. *Journal of Monetary Economics*, *31*(1), 47–67.

Boyle, G., Stover, R., Tiwana, A., & Zhylyevskyy, O. (2015). The impact of deposit insurance on depositor behavior during a crisis: A conjoint analysis approach. *Journal of Financial Intermediation*, *24*(4), 590–601.

Britz, V., Gersbach, H., & Haller, H. (2021). Deposit insurance and reinsurance. *Annals of Finance*, *17*(4), 425–470.

Bryant, J. (1980). A model of reserves, bank runs, and deposit insurance. *Journal of Banking and Finance*, *4*(4), 335–344.

Calomiris, C. W. (1990). Is deposit insurance necessary? A historical perspective. *The Journal of Economic History*, *50*(2), 283–295.

Calomiris, C. W. (1999). Building an incentive-compatible safety net. *Journal of Banking and Finance*, *23*(10), 1499–1519.

Calomiris, C. W., & Jaremski, M. (2019). Stealing deposits: Deposit insurance, risk-taking, and the removal of market discipline in early 20th-century banks. *The Journal of Finance*, *74*(2), 711–754.

Carmassi, J., Dobkowitz, S., Evrard, J., Parisi, L., Silva, A. F., & Wedow, M. (2020). Completing the Banking Union with a European deposit insurance scheme: Who is afraid of cross-subsidization? *Economic Policy*, *35*(101), 41–95.

Castriotta, M., Loi, M., Marku, E., & Naitana, L. (2019). What's in a name? Exploring the conceptual structure of emerging organizations. *Scientometrics*, *118*(2), 407–437.

Chan, A., Godwin, A., & Ramsay, I. (2020). Depositor preference and deposit insurance schemes–challenges for regulatory convergence and regulatory coordination in Asia. In *Research handbook on Asian financial law*. Edward Elgar Publishing.

Chang, C. C., & Ho, R. J. (2017). Risk shifting behaviour at commercial banks with different deposit insurance assessments: Further evidence from US market. *Journal of Financial Research*, *40*(1), 55–80.

Chang, C. C., Chung, S. L., Ho, R. J., & Hsiao, Y. J. (2022). Revisiting the valuation of deposit insurance. *Journal of Futures Markets*, *42*(1), 77–103.

Chang, Y., Zhao, S., Yang, H., He, J., & Hu, F. (2018). Determinants of deposit insurance coverage. *Prague Economic Papers*, *27*(5), 588–605.

Chen, W., Zhang, Z., Hamori, S., & Kinkyo, T. (2021). Not all bank systemic risks are alike: Deposit insurance and bank risk revisited. *International Review of Financial Analysis*, *77*, https://doi.org/10.1016/j.irfa.2021.101855

Chi, P. T., & Binh, D. T. (2020). Deposit insurance and bank risks in Vietnam. *Business and Economic Research*, *10*(2), 46–63.

Chiang, S. L., & Tsai, M. S. (2020). The valuation of deposit insurance allowing for the interest rate spread and early-bankruptcy risk. *The Quarterly Review of Economics and Finance*, *76*, 345–356.

Chiaramonte, L., Girardone, C., Migliavacca, M., & Poli, F. (2020). Deposit insurance schemes and bank stability in Europe: How much does design matter? *The European Journal of Finance*, *26*(7–8), 589–615.

Cisneros, L., Ibanescu, M., Keen, C., Lobato-Calleros, O., & Niebla-Zatarain, J. (2018). Bibliometric study of family business succession between 1939 and 2017: Mapping and analyzing authors' networks. *Scientometrics*, *117*(2), 919–951.

Demirgüç-Kunt, A., & Detragiache, E. (1998). The determinants of banking crises in developing and developed countries. *Staff Papers*, *45*(1), 81–109.

Demirgüç-Kunt, A., & Detragiache, E. (2002). Does deposit insurance increase banking system stability? An empirical investigation. *Journal of Monetary Economics*, *49*(7), 1373–1406.

Demirgüç-Kunt, A., & Huizinga, H. (2004). Market discipline and deposit insurance. *Journal of Monetary Economics*, *51*(2), 375–399.

Demirgüç-Kunt, A., & Kane, E. J. (2002). Deposit insurance around the globe: Where does it work? *Journal of Economic Perspectives*, *16*(2), 175–195.

Demirgüç-Kunt, A., Kane, E. J., & Laeven, L. (2008). Determinants of deposit-insurance adoption and design. *Journal of Financial Intermediation*, *17*(3), 407–438.

Demirgüç-Kunt, A., Kane, E. J., & Laeven, L. (2014). *Deposit insurance database* (Policy Research Working Paper No. 6934). The World Bank.

Diamond, D. W., & Dybvig, P. H. (1983). Bank runs, deposit insurance, and liquidity. *Journal of Political Economy*, *91*(3), 401–419.

Diamond, D. W., & Rajan, R. G. (2000). A theory of bank capital. *The Journal of Finance*, *55*(6), 2431–2465.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296.

Dubois, C. (2021). Deposit insurance and financial integration in the Eurozone: A DSGE model. *Economics Letters*, *207*. https://doi.org/10.1016/j.econlet.2021.110032

Duffie, D., Jarrow, R., Purnanandam, A., & Yang, W. (2003). Market pricing of deposit insurance. *Journal of Financial Services Research*, *24*(2), 93–119.

Ellegaard, O., & Wallin, J. A. (2015). The bibliometric analysis of scholarly production: How great is the impact? *Scientometrics*, *105*(3), 1809–1831.

Fabregat-Aibar, L., Barberà-Mariné, M. G., Terceño, A., & Pié, L. (2019). A bibliometric and visualization analysis of socially responsible funds. *Sustainability*, *11*(9), 2526.

Fahimnia, B., Sarkis, J., & Davarzani, H. (2015). Green supply chain management: A review and bibliometric analysis. *International Journal of Production Economics*, *162*, 101–114.

Fan, J. P., Titman, S., & Twite, G. (2012). An international comparison of capital structure and debt maturity choices. *Journal of Financial and Quantitative Analysis*, *47*(1), 23–56.

Flannery, M. J. (1989). Capital regulation and insured banks choice of individual loan default risks. *Journal of Monetary Economics*, *24*(2), 235–258.

Freixas, X., & Rochet, J. C. (1998). Fair pricing of deposit insurance: Is it possible? Yes. Is it desirable? No. *Research in Economics*, *52*(3), 217–232.

Financial Stability Board. (2017). *Principles on bail-in execution* (FSB Consultative Document). https://www.fsb.org/wp-content/uploads/P301117-1.pdf

Furlong, F. T., & Keeley, M. C. (1989). Capital regulation and bank risk-taking: A note. *Journal of Banking and Finance*, *13*(6), 883–891.

Garcia, G. (1996). *Deposit insurance: Obtaining the benefits and avoiding the pitfalls* (IMF Working Paper No. 96/83). International Monetary Fund. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=882979

Gatti, M., & Oliviero, T. (2018). *Deposit insurance and banks' deposit rates: Evidence from the 2009 EU Policy* (University of Milan Bicocca Department of Economics, Management and Statistics Working Paper No. 419). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3452664

Gennotte, G., & Pyle, D. (1991). Capital controls and bank risk. *Journal of Banking and Finance*, *15*(4–5), 805–824.

Goel, P., Garg, A., Walia, N., Kaur, R., Jain, M., & Singh, S. (2021). Contagious diseases and tourism: A systematic review based on bibliometric and content analysis methods. *Quality and Quantity*, 1–26.

Gorton, G. (1988). Banking panics and business cycles. *Oxford Economic Papers*, *40*(4), 751–781.

Gorton, G., & Pennacchi, G. (1990). Financial intermediaries and liquidity creation. *The Journal of Finance*, *45*(1), 49–71.

Gupta, J., & Sardana, V. (2021). Deposit Insurance and Banking Risk in India: Empirical Evidence on the Role of Moral Hazard. *Mudra: Journal of Finance and Accounting*, *8*(2), 79–94.

Hellmann, T. F., Murdock, K. C., & Stiglitz, J. E. (2000). Liberalization, moral hazard in banking, and prudential regulation: Are capital requirements enough? *American Economic Review*, *90*(1), 147–165.

Huizinga, H., & Nicodème, G. (2006). Deposit insurance and international bank liabilities. *Journal of Banking and Finance*, *30*(3), 965–987.

International Association of Deposit Insurers. (2018). *Deposit insurance fund target ratio*. https://www.iadi.org/en/assets/File/IADI_Research_Paper_Deposit_Insurance_Fund_Target_Ratio_July2018.pdf

International Association of Deposit Insurers. (2021). *Deposit insurance systems worldwide*. https: //www.iadi.org/en/about-iadi/deposit-insurance-systems/dis-worldwide/

International Association of Deposit Insurers & Islamic Financial Services Board. (2021). *Core principles for effective Islamic deposit insurance system*. https://www.iadi.org/en/assets/File/Core%20Principles/IADI-IFSB%20CPIDIS_IFSB%20Approved_05%20July%202021%20Clean.pdf

Igwe, E. I., & Toby, A. J. (2021). Deposit insurance and credit risk of Nigeria banking system: A time series analysis. *Journal of Global Economics and Business*, *2*, 1–30.

Imai, M. (2006). Market discipline and deposit insurance reform in Japan. *Journal of Banking and Finance*, *30*(12), 3433–3452.

Iyer, R., & Puri, M. (2012). Understanding bank runs: The importance of depositor-bank relationships and networks. *American Economic Review*, *102*(4), 1414–1445.

Iyer, R., Puri, M., & Ryan, N. (2016). A tale of two runs: Depositor responses to bank solvency risk. *The Journal of Finance*, *71*(6), 2687–2726.

Jackson, T., & Kotlikoff, L. J. (2021). Banks as potentially crooked secret keepers. *Journal of Money*, *Credit and Banking*, *53*(7), 1593–1628.

Jameaba, M. S. (2018). Deposit insurance and financial intermediation: The case of Indonesia Deposit Insurance Corporation. *Cogent Economics and Finance*, *6*(1). https://doi.org/10.1080/23322039.2018.1468231

Kane, E. J. (1986). Appearance and reality in deposit insurance: The case for reform. *Journal of Banking and Finance*, *10*(2), 175–188.

Kane, E. J. (2006). Can the European community afford to neglect the need for more accountable safety-net management? *Atlantic Economic Journal*, *34*(2), 127–144.

Karels, G. V., & McClatchey, C. A. (1999). Deposit insurance and risk-taking behavior in the credit union industry. *Journal of Banking and Finance*, *23*(1), 105–134.

Keeley, M. C. (1990). Deposit insurance, risk, and market power in banking. *The American Economic Review*, *80*(5), 1183–1200.

Khanra, S., Dhir, A., & Mäntymäki, M. (2020). Big data analytics and enterprises: A bibliometric synthesis of the literature. *Enterprise Information Systems*, *14*, 737–768.

Kim, D., & Santomero, A. M. (1988). Risk in banking and capital regulation. *The Journal of Finance*, *43*(5), 1219–1233.

Kleftouri, N. (2015). Deposit protection system's design and effectiveness. *Journal of Banking Regulation*, *16*(1), 51–63.

Kozińska, M. (2021). Role of deposit insurance schemes on financial markets. *International Journal of Monetary Economics and Finance*, *14*(1), 91–104.

Kumar, S., Lim, W.M., Pandey, N., & Westland, J.C. (2021). 20 years of electronic commerce research. *Electronic Commerce Research*, *21*, 1–40.

Kusairi, S., Sanusi, N. A., & Ismail, A. G. (2018). Dilemma of deposit insurance policy in ASEAN countries: Does it promote banking industry stability or moral hazard? *Borsa Istanbul Review*, *18*(1), 33–40.

Kuznichenko, P., Frolov, S., Orlov, V., & Boiko, O. (2021). European deposit insurance scheme implementation: Pros and cons. *Banks and Bank Systems*, *16*(1), 116–126.

Laeven, L., & Levine, R. (2009). Bank governance, regulation and risk taking. *Journal of Financial Economics*, *93*(2), 259–275.

Lazarov, I. (2017). Deposit insurance in the EU: Repetitive failures and lessons from across the Atlantic. *Common Market Law Review*, *54*(6).

Lee, W. S., & Kwok, C. C. (2000). Domestic and international practice of deposit insurance: A survey. *Journal of Multinational Financial Management*, *10*(1), 29–62.

Lewrick, U., Serena, J. M., & Turner, G. (2019). *Believing in bail-in? Market discipline and the pricing of bail-in bonds* (Bank for International Settlements Working Paper No. 831). https://www.bis.org/publ/work831.pdf

Lin, M.Y. (2015). Deposit insurance and effectiveness of monetary policy. *Applied Economics Letters*, *22*(18), 1443–1449.

Liu, L., Zhang, G., & Fang, Y. (2016). Bank credit default swaps and deposit insurance around the world. *Journal of International Money and Finance*, *69*, 339–363.

Marcus, A. J. (1984). Deregulation and bank financial policy. *Journal of Banking and Finance*, *8*(4), 557–565.

Marcus, A. J., & Shaked, I. (1984). The valuation of FDIC deposit insurance using option-pricing estimates. *Journal of Money, Credit and Banking*, *16*(4), 446–460.

Martinez Peria, M. S., & Schmukler, S. L. (2001). Do depositors punish banks for bad behavior? Market discipline, deposit insurance, and banking crises. *The Journal of Finance*, *56*(3), 1029–1051.

Maura-Pérez, M. J., & Romero-Perez, H. (2021). Factors related to the failure of FDIC-insured US banks. *Journal of Financial Regulation and Compliance*, *30*(1), 82–106.

Merton, R. C. (1977). An analytic derivation of the cost of deposit insurance and loan guarantees an application of modern option pricing theory. *Journal of Banking and Finance*, *1*(1), 3–11.

Mohammad, K. U., & Nishiyama, S. I. (2019). An empirical investigation into the effect of explicit deposit insurance and design on the liability structure of banks. *Journal of Central Banking Theory and Practice*, *8*(3), 179–206.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*, 213–228.

Moral-Muñoz, J. A., Herrera-Viedma, E., Santisteban-Espejo, A., & Cobo, M. J. (2020). Software tools for conducting bibliometric analysis in science: An up-to-date review. *El Profesional de la Información*, *29*(1), e290103.

Morrison, A. D., & White, L. (2011). Deposit insurance and subsidized recapitalizations. *Journal of Banking and Finance*, *35*(12), 3400–3416.

Mumtaz, R., Jadoon, I. A., & Sohail, N. (2019). Explicit deposit insurance and bank risk-taking: Does banking supervision matter? *Pakistan Journal of Commerce and Social Sciences (PJCSS)*, *13*(3), 727–748.

Nayak, R. (2021). Banking regulations: Do they matter for performance? *Journal of Banking Regulation*, *22*(4), 261–274.

Ngalawa, H., Tchana, F. T., & Veigi, N. (2016). Banking instability and deposit insurance: The role of moral hazard. *Journal of Applied Economics*, *19*(2), 323–350.

Noman, A. H. M., Gee, C. S., & Isa, C. R. (2018). Does bank regulation matter on the relationship between competition and financial stability? Evidence from Southeast Asian countries. *Pacific-Basin Finance Journal*, *48*, 144–161.

O'hara, M., & Shaw, W. (1990). Deposit insurance and wealth effects: The value of being 'too big to fail'. *The Journal of Finance*, *45*(5), 1587–1600.

Pandey, K., & Joshi, Sujata. (2021). Trends in destination choice in tourism research: A 25-year bibliometric review. *FIIB Business Review*, *10*(4), 371–392.

Peck, J., & Shell, K. (2003). Equilibrium bank runs. *Journal of Political Economy*, *111*(1), 103–123.

Pennacchi, G. G. (1987). A reexamination of the over-(or under-) pricing of deposit insurance. *Journal of Money, Credit and Banking*, *19*(3), 340–360.

Persson, O., Danell, R., & Schneider, J. W. (2009). How to use Bibexcel for various types of bibliometric analysis. *Celebrating Scholarly Communication Studies*, 9–24.

Prean, N., & Stix, H. (2011). The effect of raising deposit insurance coverage in times of financial crisis–Evidence from Croatian microdata. *Economic Systems*, *35*(4), 496–511.

Putri, E. R., Tjahjono, V. R., & Imron, C. (2021). A deposit insurance pricing with a multi-state regime-switching volatility. *International Journal of Applied and Computational Mathematics*, *7*(6), 1–20.

Ronn, E. I., & Verma, A. K. (1986). Pricing risk-adjusted deposit insurance: An option-based model. *The Journal of Finance*, *41*(4), 871–895.

Rossetto, D. E., Bernardes, R. C., Borini, F. M., & Gattaz, C. C. (2018). Structure and evolution of innovation research in the last 60 years: Review and future trends in the field of business through the citations and co-citations analysis. *Scientometrics*, *115*(3), 1329–1363.

Sahadewo, G. A., Purwanto, B. M., & Pradiptyo, R. (2018). Does a deposit insurance scheme induce moral hazard among bankers? Evidence from an experiment with bankers. *Gadjah Mada International Journal of Business*, *20*(3), 355–737.

Salter, A. W., Veetil, V., & White, L. H. (2017). Extended shareholder liability as a means to constrain moral hazard in insured banks. *The Quarterly Review of Economics and Finance*, *63*, 153–160.

Sánchez-Teba, E. M., Benítez-Márquez, M. D., & Porras-Alcalá, P. (2021). Gender diversity in boards of directors: A bibliometric mapping. *Journal of Open Innovation: Technology, Market, and Complexity*, *7*(1), 12.

Sardana, V., & Shukla, A. (2020). Deposit insurance coverage limit: How much is enough? evidence from India. *Indian Journal of Finance and Banking*, *4*(4), 66–83.

Saunders, A., & Wilson, B. (1996). Contagious bank runs: Evidence from the 1929–1933 period. *Journal of Financial Intermediation*, *5*(4), 409–423.

Shy, O., Stenbacka, R., & Yankov, V. (2016). Limited deposit insurance coverage and bank competition. *Journal of Banking and Finance*, *71*, 95–108.

Singh, J., Singhania, S., & Aggrawal, D. (2021). Gender diversity on corporate boards: Review and future research agenda through bibliometric mapping. *Corporate Governance and Sustainability Review*, *5*(3), 57–72.

Stern, G. H. (1998). Government safety nets, banking system stability, and economic development. *Journal of Asian Economics*, *9*(1), 21–29.

Stone, A. L. (2021). Double the insurance, double the funds? *Journal of Banking and Finance*, *133*, 106284.

Stremersch, S., Verniers, I., & Verhoef, P. C. (2007). The quest for citations: Drivers of article impact. *Journal of Marketing*, *71*(3), 171–193.

Tahamtan, I., Safipour Afshar, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics*, *107*(3), 1195–1225.

Tomar, S., Kumar, S., & Sureka, R. (2021). Financial planning for retirement: Bibliometric analysis and future research directions. *Journal of Financial Counseling and Planning*, *32*(2), 1–19.

Trinugroho, I., Pamungkas, P., Ariefianto, M. D., & Tarazi, A. (2020). Deposit structure, market discipline, and ownership type: Evidence from Indonesia. *Economic Systems*, *44*(2), 100758.

Tunger, D., & Eulerich, M. (2018). Bibliometric analysis of corporate governance research in German-speaking countries: Applying bibliometrics to business research using a custom-made database. *Scientometrics*, *117*(3), 2041–2059.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.

Vernikov, A. (2020). Deposit protection in a paternalistic state: The Russian case. *Terra Economicus*, *18*(1), 28–42.

Wheelock, D. C., & Kumbhakar, S. C. (1995). Which banks choose deposit insurance? Evidence of adverse selection and moral hazard in a voluntary insurance system. *Journal of Money, Credit and Banking*, *27*(1), 186–201.

Wong, T. W. (2020). Game theoretic valuation of deposit insurance under jump risk: From too small to survive to too big to fail. *Mathematics and Financial Economics*, *14*(1), 67–95.

Wong, T. W., Fung, K. W. T., & Leung, K. S. (2020). Strategic bank closure and deposit insurance valuation. *European Journal of Operational Research*, *285*(1), 96–105.

Wu, Y. C., Chen, T. F., & Lin, S. K. (2020). Risk management of deposit insurance corporations with risk-based premiums and credit default swaps. *Quantitative Finance*, *20*(7), 1085–1100.

Xu, X., Chen, X., Jia, F., Brown, S., Gong, Y., & Xu, Y. (2018). Supply chain finance: A systematic literature review and bibliometric analysis. *International Journal of Production Economics*, *204*, 160–173.

Zhang, Z., & Wu, F. (2020). Moral hazard, external governance and risk-taking: Evidence from commercial banks in China. *Finance Research Letters*, *37*. https://doi.org/10.1016/j.frl.2019.101383

Zheng, C., & Kouwenberg, R. (2019). A bibliometric review of global research on corporate governance and board attributes. *Sustainability*, *11*(12), 3428.

## About the Authors

**Varda Sardana** is an Assistant Professor at Shri Ram College of Commerce, University of Delhi. She is also a doctoral researcher at the Department of Commerce, University of Delhi. Her academic and research interests include banking, finance, corporate governance and education. She can be reached at varda.sardana@gmail.com

**Shubham Singhania** is a PhD Research Scholar at the University School of Management and Entrepreneurship, Delhi Technological University, India. He has attained JRF in commerce and management and NET in HRM. He has published research papers in Scopus-indexed and ABDC-listed journals, and he also serves as a reviewer for many editorials. His area of interests includes education, corporate governance and sustainability. He can be reached at shubhamsinghania1@gmail.com

Journal of Electronic Materials
https://doi.org/10.1007/s11664-022-09901-3

ORIGINAL RESEARCH ARTICLE

# First-Principles Study of Enhanced Absorption in Van der Waals Heterostructure of $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ in the Visible Region

**Anurag Chauhan**[1] · **Prateek Tomar**[1]

Received: 16 February 2022 / Accepted: 23 August 2022
© The Minerals, Metals & Materials Society 2022

## Abstract

To explore the optoelectronic characteristics of the quaternary compound, cadmium zinc telluride selenium ($Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$), and monolayer $MoS_2$, DFT simulations are performed on the van der Waals (vdW) heterostructure. $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ and monolayer $MoS_2$ are investigated and found to exhibit absorption in the ranges of ~180 to 350 nm and ~150 to 300 nm, respectively, and the $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ heterostructure showed in a redshift phenomenon with low absorption outside the visible region. The absorption coefficient ($\alpha_\alpha$) shifted towards a higher wavelength with selective absorption in the range ~450 to 640 nm with negligible interference from the UV and infrared regions. This consistent absorption in the visible spectrum demonstrates its uses in optoelectronics and suggests its possibility as a suitable material for solar cells. These findings have been corroborated by the results obtained for the refractive index ($\eta$) and dielectric constant ($\varepsilon$).

**Keywords** Optoelectronics · density functional theory (DFT) · transition metal dichalcogenides (TMD) · redshift

## Introduction

The possibility of making atomically thin electronic devices became a reality after the discovery of graphene. The zero-band gap of graphene and its limited applications in digital design systems forced researchers to look beyond gapless materials.[1–3] The alternative came in the form of transition metal dichalcogenides (TMDs), which can replace graphene.[2–4] The van der Waals (vdW) bounded layers in these TMDs can be easily separated by mechanical exfoliation into stable units of atomic thickness from the bulk, which shows remarkable optical and mechanical properties. Moreover, these properties are widely tunable by external fields, doping, strain, and atomic layer thickness. The formation of heterostructures by vertically stacking 2D flexible membranes has attracted tremendous technological interest.[5,6] The TMD structure can be represented by the strongly bonded in-plane atoms with weakly attracted off-plane atoms through vdW forces. These weakly attracted interlayers can be transformed into a few or a single layer of atoms that help tun the photonic properties.[7–9] The weak interlayer vdW interactions between stacked layers enable wide applications in flexible electronics such as plastic and paper-based devices.[7] $MoS_2$, a member of the $MX_2$ family (where M is the transition metals and X is the chalcogens), has good light-matter interaction and outstanding absorption capacity near the UV region of the visible light spectrum, producing impressive photovoltaic applications such as in photodetectors, electrocatalysts for lithium batteries, and solar cells. Because of the widespread presence of molybdenite, $MoS_2$ is one of the most researched monolayer transition metal dichalcogenides (TMDs). Monolayer $MoS_2$ is a direct band gap semiconductor with a band gap of 1.79 eV, while bulk $MoS_2$ is an indirect band gap semiconductor, and this tunable band gap property of $MoS_2$ is motivating because it largely compensates for the weakness of gapless graphene and enables the use of 2D materials in the next-generation switching and optoelectronics field.[10–12] $MoS_2$ atomic layers are isolated and have dangling bonds; elimination of these dangling bonds is beneficial for the heterostructure as it helps make the structure stable.[6] Since the early 1970s, cadmium telluride (CdTe) has been considered a potential semiconducting material for detecting strong X-rays and gamma rays. These detectors have many applications ranging from imaging, non-proliferation and national safety

✉ Anurag Chauhan
anuragchauhan@dtu.ac.in

Prateek Tomar
prateektomar.08@gmail.com

[1] Delhi Technological University, New Delhi, India

Published online: 10 September 2022

to astrophysics.[13,14] Meanwhile, silicon (Si) and germanium (Ge) detectors have great energy resolution and charge transport features but have poor stopping power for high-energy photons, limiting their IR detection applications. Compounds with high atomic numbers are more suitable for IR and X-ray detector applications due to their high quantum efficiency.[15–17] The large band gap of CdTe allows its compounds to operate at room temperature. With the advancement of technology, the problem of fabricating high-resolution detectors was solved as it is now possible to produce high-quality single crystals of CdTe and cadmium zinc telluride (CdTeZn). The introduction of Se in CdZnTe resulted in a quaternary compound that helped minimize compositional inhomogeneity and sub-grain boundary networks present in $Cd_{0.90}Zn_{0.10}Te$ and reduced its production cost.[18] $Cd_{0.90}Zn_{0.10}Te$ CZT (cadmium zinc telluride) has proved to be the most promising material that can be used commercially.[19] Quaternary compound $Cd_{1-x}$ $Zn_x$ $Te_{1-y}$ $Se_y$ was suggested to eliminate all those defects present in CdTeZn for a fixed percentage of Zn, 10%. By decreasing the percentage of Se from 7 to 2%, it was observed that the uniform composition of Te and Se along the crystal decreases (Figs. 1 and 2). Moreover, mobility and lifetime product was reported to be maximum with 7% selenium; hence, $Cd_{0.90}$ $Zn_{0.10}$ $Te_{0.93}$ $Se_{0.07}$ was suggested as the best compositional material for radiation detectors.[18] The band gap of $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ is measured as 1.501 eV in Fig. 3(b), which is consistent with previously published work.[19,20] The nature of 2D materials ranges from semiconducting ($MoS_2$, $WS_2$) to semi-metallic ($WTe_2$), metallic ($TiS_2$), superconductors ($TaS_2$) and insulators (hBN). More interestingly, vertically stacking these materials over one another forms a heterostructure that exhibits totally new characteristics at the interface and enables unique properties.[21] The neighbouring layers of these materials are not connected by any chemical bonds but interact by van der Waals forces. After detailed analysis, $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ is chosen with proper percentages of Zn and Se. These two materials have been investigated and explored to understand the unique physics at the interface. So in this paper, optical and electronic properties of heterostructure of $MoS_2$ and $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ have been investigated accordingly.

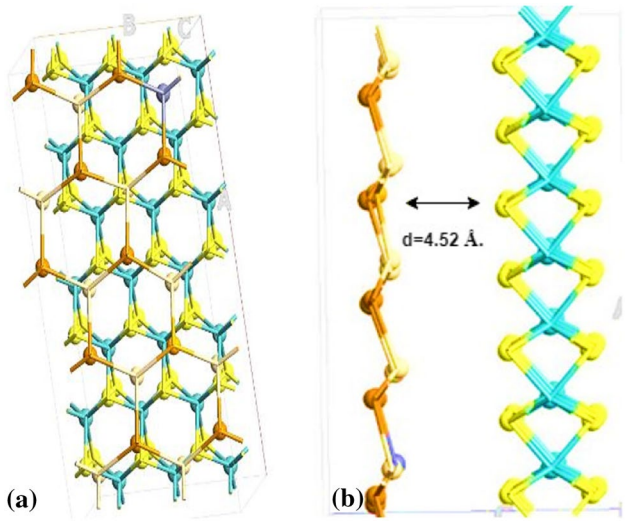



**Fig. 1** Relaxed geometry (a) Heterostructure of $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ (b) Interlayer distance between $^{Cd}0.90^{Zn}$ $0.10^{Te}0.93^{Se}0.07$ and $^{MoS}2$.

## Computational Details

The optoelectronic properties of $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ heterostructure were investigated on a Quantumwise Atomistix Tool Kit (ATK) using density functional theory (DFT)-based simulations. A limited-memory BFGS (LM-BFGS) algorithm is used to relax all the atomic positions with an applied force tolerance of 0.05 eV/Å. Generalized gradient approximation (GGA) with Perdew-Burke-Ernzerhof[22,23] (PBE) functional is used to calculate the exchange–correlation energy.[24] Fig. 1 shows the optimized geometry with k-points sampling of $9\times9\times1$ in all three directions, x, y, and z. The Monkhorst–Pack Scheme is used with 150 Ry mesh cut-off energy.[25,26]

A vertically stacked monolayer of $MoS_2$ on $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ is shown in Fig. 1(a). Both the geometries are repeated in the x and y plane, separated by an interlayer distance of 4.52 Å on the z-axis. The interaction of vdW forces cannot be described entirely by using the PBE functional alone; therefore, semi-empirical DFT-D2 Grimme was used. DFT:LCAO and MGGA exchange–correlation along with TB09-LD were selected for the calculation of optical properties. A symmetry sampling of $15\times15\times1$ was taken into account by setting the broadening value to 0.01 eV.[25]

The Kubo-Greenwood equation is used to determine susceptibility, as shown by the equation below.[27]

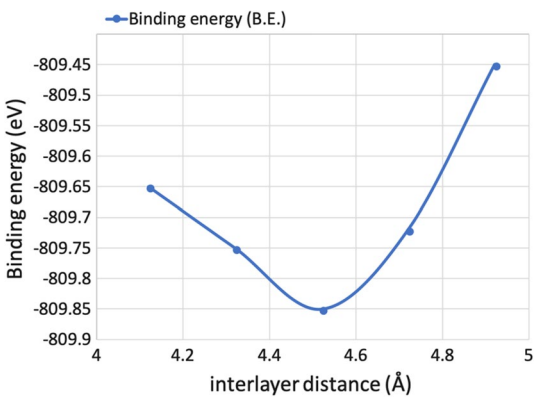



**Fig. 2** Interlayer distance vs. Binding energy.

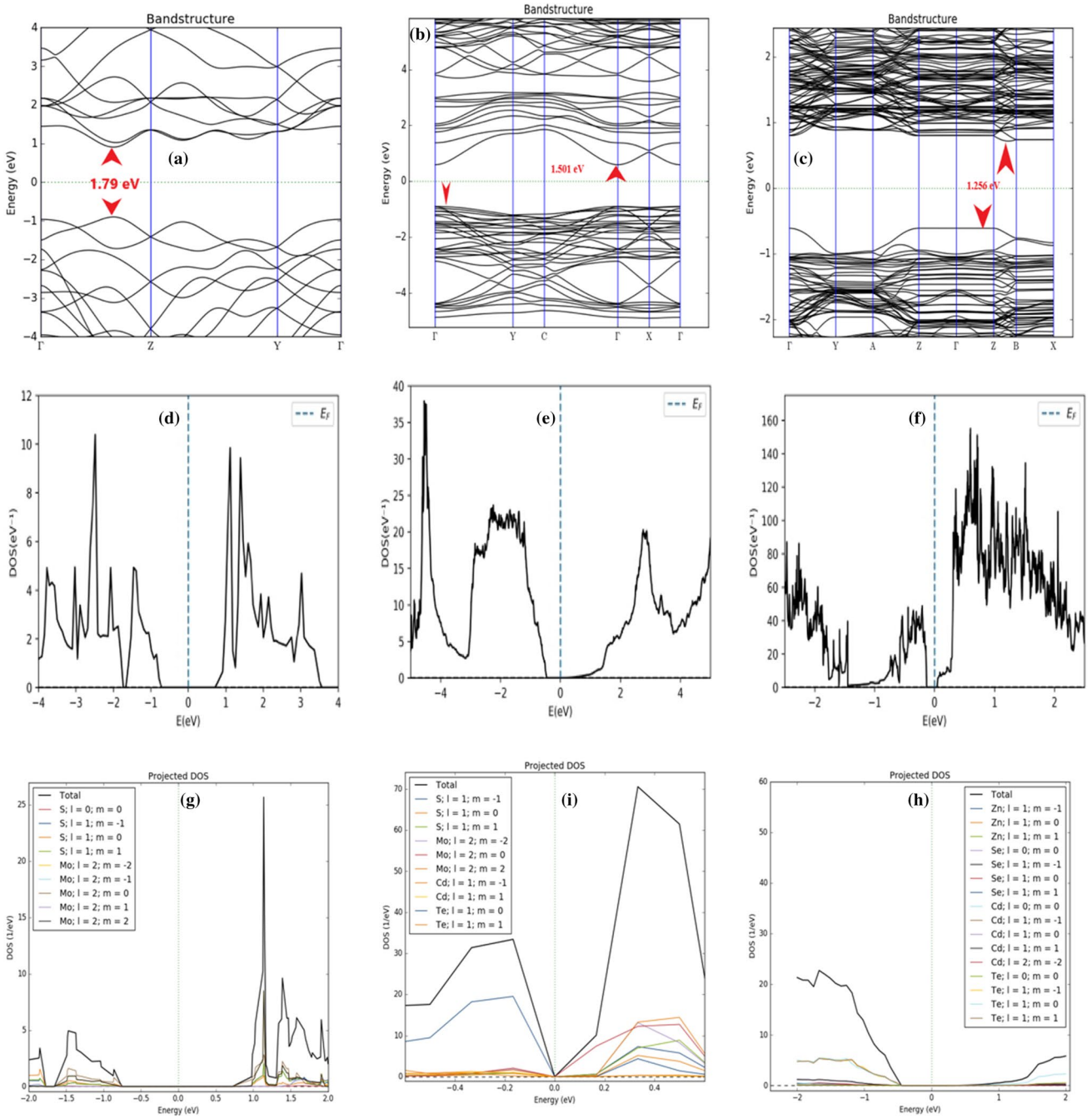



**Fig. 3** Band structure for (a) $MoS_2$ (b) $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ (c) $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ heterostructure. DOS for (d) $MoS_2$ (e) $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ (f) $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ heterostructure. PDOS for (g) $MoS_2$ (h) $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ (i) $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ heterostructure.

$$\chi_{i,j}(\omega) = -\frac{e^2h^4}{\varepsilon_0 m^2\omega^2 V}\sum_{nm}\frac{f(E_m) - f(E_n)}{E_{nm} - h\omega - i\Gamma}\pi^i_{nm}\pi^j_{nm}$$

where $f$ is the Fermi function, $\pi^i_{nm}$ is the *ith* component of the dipole matrix element between state $n$ and $m$, $\Gamma$ is the broadening and $V$ is the volume. For the calculation of susceptibility, relative dielectric constant '$\varepsilon_r$', optical conductivity '$\sigma$', and polarizability '$\alpha$' are used.[28]

$$\varepsilon_r = 1 + \chi(\omega)$$

$$\sigma(\omega) = -\mathrm{Img}\left(\omega\varepsilon_0\chi(\omega)\right)$$

$$\alpha(\omega) = V\varepsilon_0\chi(\omega)$$

Refractive index (η) is related to both $\kappa$ and $\varepsilon_r$ as:

$$\eta + i\kappa = \sqrt{\varepsilon_r}$$

$$\varepsilon(\omega) = (\eta(\omega) + i\kappa(\omega))^2$$

$\kappa$ represents the extinction coefficient. Real $(\varepsilon_1)$ and complex $(\varepsilon_2)$ parts of $\varepsilon_r$, $\kappa$ and refractive index $\eta$ are given below.

$$\eta = \sqrt{\frac{\sqrt{\varepsilon_1^2 + \varepsilon_2^2} + \varepsilon_1}{2}} \quad \kappa = \sqrt{\frac{\sqrt{\varepsilon_1^2 + \varepsilon_2^2} - \varepsilon_1}{2}}$$

The relation between optical absorption coefficient and the extinction coefficient is given as[29]

$$\alpha_\alpha = 2\frac{\omega}{c}\kappa$$

## Results and Analysis

### Geometric Structure and Stability

Vertically stacking flexible 2D membrane of one material over another resulted in the formation of the heterostructure. Figure 1(a) shows $MoS_2$ over $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ vertically stacked in such a way that the lattice mismatch subsides. Only 2.5% of lattice mismatch was reported in the supercell, which ensures a negligible effect on its electronic structure. The $MoS_2$ monolayer has a hexagonal structure and calculated lattice parameters have γ = 120°, α and β of 90° with lattice constant, a = 3.16 Å which is confirmed with previous studies.[30] Atomic layers of Mo are sandwiched between two atomic S layers with Mo-S-Mo atomic bond length of 2.42 Å and bond angle of 81.2. $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ has a zigzag structure along the y-axis with a bond length of 2.73 Å, 2.71 Å and 2.82 Å for Zn-Te, Se-Cd and Cd-Te, respectively. The $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ supercell is obtained by its repetition (2 × 3 × 1) along the lattice plane, the supercell with lattice parameters a = 22.903 Å, b = 9.1641 Å, c = 4.58205 Å is obtained by substituting a percentage of specified indices with the new element using normal distribution. The wide range of doping strategies and sources presently available offers intriguing possibilities for introducing novel characteristics and functionalities, such as substitutional, exterior charge transfer and electrostatic doping.[31]

Interface binding energy that gives the stable configuration of the heterostructure is defined by $E_B = \frac{E-(E_1+E_2)}{N}$, where $E_1$ and $E_2$ represent the total energy of free-standing $MoS_2$ and $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$, respectively.[22] E gives the total energy of the $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ heterostructure and N gives the total number of free electrons in the heterostructure.[9,32] Figure 2 provides the graph of interlayer distance vs. binding energy of the heterostructure. Binding

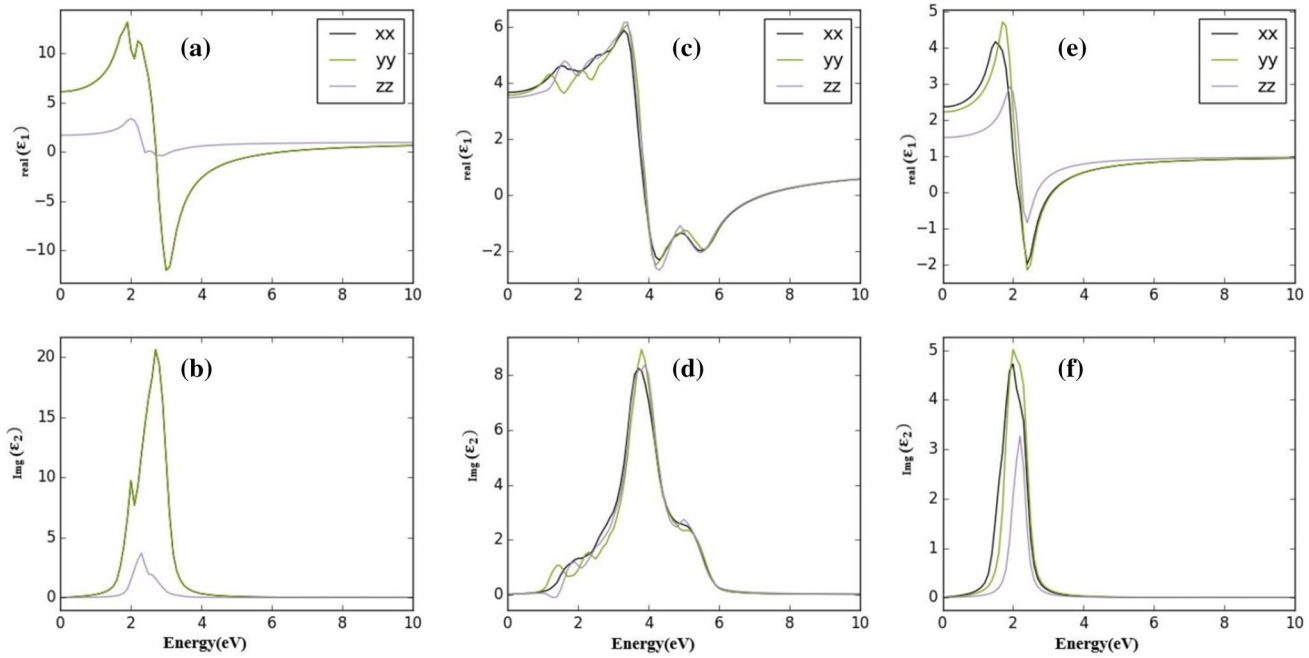



**Fig. 4** Real part ($\varepsilon_1$) and imaginary part ($\varepsilon_2$) of the dielectric constant($\varepsilon$) versus energy for (a) and (b) $MoS_2$ (c)and (d) $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ (e) and (f) Heterostructure of $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ in xx, yy and zz tensors.

energy is calculated at different interlayer distances, and most negative energy is preferred as this provides the most stable structure. This confirms the strong vdW interaction between stacked layers. In our calculations, the most negative energy per atom is 809.85 eV found at a distance of 4.52 Å.

## Electronic Structure

Density of states (DOS) and band structure of $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$, monolayer $MoS_2$ and the heterostructure of $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ have also been calculated to investigate the electronic properties. From Fig. 3(b), the calculated indirect band gap for $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ is 1.501 eV. Interestingly, for the $MoS_2$ monolayer, valence band maximum (VCM) and conduction band minimum (CBM) are found to be at K symmetry points. As can be seen in Fig. 3(a), monolayer $MoS_2$ resulted in a direct band gap semiconductor with a band gap of 1.79 eV, which is in confirmation with the previous results.[33,34] Exciton binding energy is directly related to electron–hole interaction; traditionally used semiconductors such as Si and Ge have low binding energy, due to which electrical and optical band gaps are often at the same value. But in the case of inorganic semiconductors such as $MoS_2$ , binding energy is higher as compared to Si and Ge; therefore, optical and electrical band gap cannot be the same.[33,35,36] Understanding the excitonic nature of excited states is a measure of charge transport characteristics in optoelectronic applications. Hence, we calculate the exciton binding energy for $MoS_2$, which is 690 meV. Observing the band structure and DOS in Fig. 3(c) and (f), respectively, the heterostructure resulted in a semiconducting nature with an indirect band gap of 1.256 eV. The projected density of states (PDOS) have also been plotted as shown in Fig. 3(g), (h), (i). As can be seen in Fig. 3(g), electronic states in the $MoS_2$ monolayer are mainly contributed by the d orbital of Mo. In $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$, the majority of conducting states are contributed by the p ($p_x$, $p_y$ *and* $p_z$) orbital in Te, Cd and Se, which can be observed in Fig. 3(h). The PDOS of heterostructure shown in Fig. 3(h) near the Fermi level is mainly dominated by the *p* ($p_x$, $p_y$ *and* $p_z$) orbitals of Cd and Te and *d* orbitals of Mo.

## Optical Properties

In this section, a comparative analysis of the individual material and heterostructure is done to infer the results of the dielectric constant and absorption coefficient. The interaction of incident photons and propagation of electron–hole pair through the material is best explained by the energy-dependent dielectric constant ($\varepsilon$).[9,37] The dielectric constant ($\varepsilon$) consists of both the real part ($\varepsilon_1$) and imaginary part ($\varepsilon_2$), polarization and anomalous dispersion effects are given by ($\varepsilon_1$), and the imaginary component ($\varepsilon_2$) provides information about the absorption of energy and the excitonic energy that drives excitons in the system. The computed real and imaginary parts of the dielectric constant versus photon energy is shown in Fig. 4 along the xx, yy, and zz tensors for all the structures. Optical absorption of the materials can be directly calculated from the imaginary part ($\varepsilon_2$) as it provides the information about energy absorption in the material; hence, the curve of ($\varepsilon_2$) along the y-axis is in Fig. 4(b), (d, (f).[35] For $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07,}$ peaks of $\varepsilon_2$ are observed in the range of (~3–5) eV, with the highest peak at ~3.8 eV in Fig. 4(d). For the $MoS_2$ monolayer, the peak can be seen at ~2.7 eV in Fig. 4(b).

For the $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ heterostructure, only one peak can be observed with a maximum value at ~2 eV in Fig. 4(f) that has shifted to much lower energy as compared to $MoS_2$ and $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$. This indicates that excitonic energy for the heterostructure is much lower than its constituents with an improved absorption coefficient of the incident energy. The above results are consistent with the absorption coefficient vs. energy results in Fig. 5, as the dielectric constant is directly related to the absorption coefficient. Figure 5 shows the comparative view of the energy-dependent absorption coefficient where $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ has two peaks at ~3.8 eV and ~5.8 eV, respectively. Also, the resulting heterostructure has stable and selective absorption in the visible region and it has shifted towards the higher wavelength of the visible spectrum or the lower energy as compared to monolayer

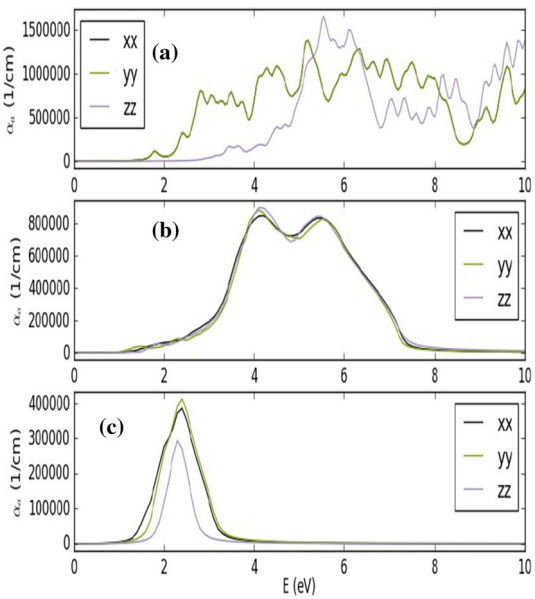



**Fig. 5** Absorption coefficient as function of energy for (a) $MoS_2$ (b) $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ (c) Heterostructure of $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ in xx, yy and zz tensors.

$MoS_2$ and $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ individual material as shown in Figs. 5(c) and 6(c). This resulted in a redshift phenomenon with broad absorption range and stable absorption for visible light as compared to its constituent materials. Incident photons energy is strongly absorbed in the visible spectrum as $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ heterostructure

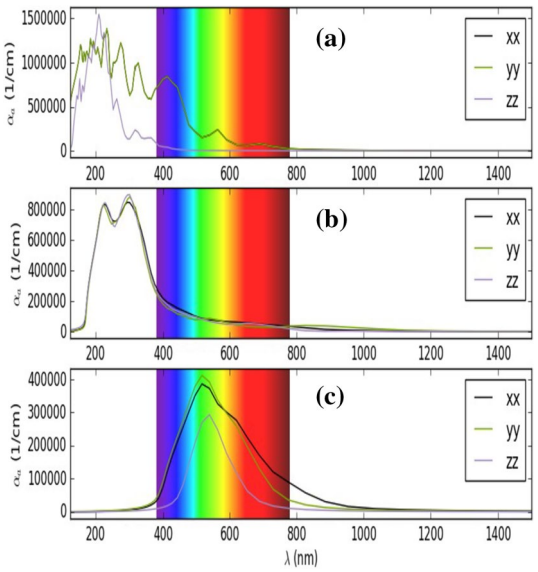



**Fig. 6** Absorption coefficient as a function of wavelength for (a) $MoS_2$ (b) $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ (c) Heterostructure of $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ in xx, yy and zz tensors.

provides a lower band gap in comparison with $MoS_2$ monolayer and $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$. Therefore, the photogenerated electrons are excited directly from $MoS_2$ valence band to $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ conduction band, resulting in stable and selective absorption in the visible region. This can also be confirmed from Fig. 7(c), which has shifted peaks of refractive index ($\eta$) towards lower energy values with more selectivity. The heterostructure resulted in selective and stable absorption of visible light with an increased rate and range than that of its constituents, proving its application in optoelectronics, solar cell absorbers, LEDs and optical communication.[38–40] In optoelectronics application, absorption in UV and IR is not suitable as it appears in losses due to the device heating which decreases the efficiency of the system. Figure 6(c) absorption coefficient clearly shows the negligible interference of UV and infrared regions. Therefore, the photon energy that occupies the visible region will be sufficient for the generation of electron–hole pairs to generate electric current. Therefore, devices made from $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ heterostructure will have less heat losses and higher efficiency as compared to individual $MoS_2$ and $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ based devices.[41] It can also be observed from Fig. 7(c) that the refractive index of the heterostructure is overlapping along the xx and yy tensors, whereas it is different along with the zz tensor. It indicates this anisotropic nature of the heterostructure can also be used for polarizing applications.[32,42,43] Heterostructure made from TMDs such as $MoS_2$ provides outstanding

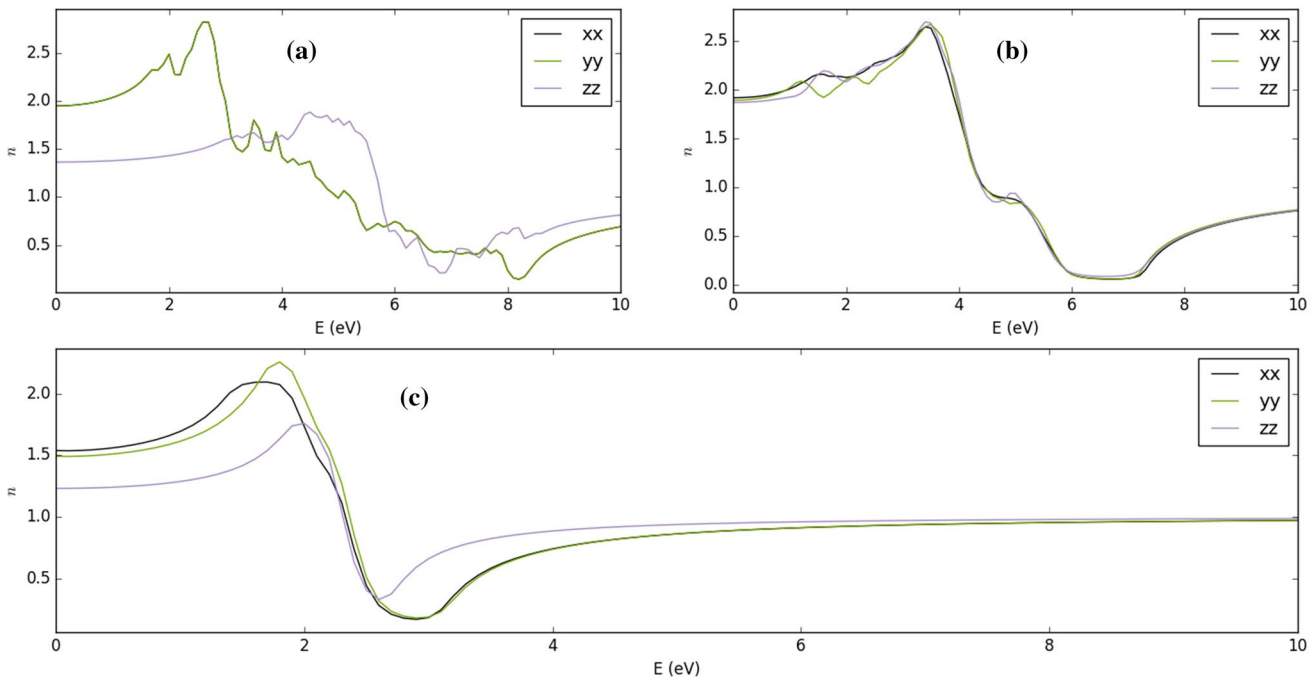



**Fig. 7** Energy-dependent refractive index for (a) $MoS_2$ (b) $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ (c) Heterostructure of $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ in xx, yy and zz tensors.

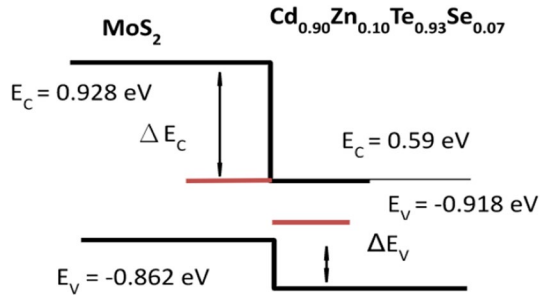



**Fig. 8** The schematic structure of the energy band alignment at $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ heterostructure.

semiconducting properties and mechanical flexibility as compared to traditionally used Ge and Ga-As semiconductors. The strong light-matter coupling in the proposed heterostructure will give higher absorption as compared to indirect band gap semiconductors such as Si and Ge with the same thickness.[32,41]

The refractive index ($\eta$) of the heterostructure and its constituent are calculated in Fig. 7(a-c). As discussed in section II, $\eta$ is related to both the real part ($\varepsilon_1$) and imaginary part ($\varepsilon_2$) of dielectric function; therefore, $\eta$ follows the same trend as that of dielectric constant and is directly proportional to the absorption coefficient. The maximum value of $\eta$ in $MoS_2/Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ heterostructure is reported as ~1.8 eV, starts decreasing after 2 eV and saturates after 4 eV, quite similar to the absorption coefficient and dielectric function in Figs. 5(c) and 4(f), respectively. A higher refractive index ($\varepsilon$) indicates that light will remain in the structure for a longer period of time which corresponds to smaller exciton binding energy.

## Band Alignment

To design the heterojunction-based devices for optoelectronics and solar energy conversion application, it is important to mention the valance and conduction band offset at the junction. For this purpose, Fig. 8 is plotted to show the type of band alignment this heterojunction exhibits.[44] Schematic structure in Fig. 8 clearly shows the type II band alignment at the interface, which makes this heterostructure a promising material for optoelectronics and solar energy conversion.

Band bending calculations have been done at the interface with the expression: $\Delta E_f = W - W_s$ where W is the work functions (potential) of the heterostructure and $W_s$ is the work function (potential) of $MoS_2$ and $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$. Calculated band bending values for $MoS_2$ and $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ are 0.18 eV and -0.862 eV, respectively..

## Conclusion

First-principles DFT calculations have been performed on the vertically stacked monolayer $MoS_2$ on $Cd_{0.90}Zn_{0.10}Te_{0.93}Se_{0.07}$ that constitutes a vdW heterostructure. The heterostructure resulted in shifting of the absorption coefficient towards a higher wavelength or lower energy levels with more confinement and more stability as compared to individual constituent material. A redshift is also observed, and the same trend was also noticed for refractive index ($\eta$) and dielectric constant ($\varepsilon$). These results are conclusive and provide a wide scope of improvement in solar energy conversion materials for optoelectronic devices and applications.

**Authors' Contributions** All authors contributed to the study's conception and design. Material preparation, data collection and analysis were performed by Anurag Chauhan and Prateek Tomar. The first draft of the manuscript was written by Prateek Tomar, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Funding** No funding was received to conduct this study.

**Availability of Data and Material** Not applicable.

**Code Availability** Not applicable.

**Conflicts of interest** The authors declare they have no conflict of interest.

**Ethics Approval** This article does not contain any studies with human participants or animals performed by any of the authors.

**Consent to Participate** Not applicable.

## References

1. K.S. Novoselov, D. Jiang, F. Schedin, T.J. Booth, V.V. Khotkevich, S.V. Morozov, and A.K. Geim, *PNAS* 102, 10451–10453 (2005).
2. S. Das, J.A. Robinson, M. Dubey, H. Terrones, and M. Terrones, *Annu. Rev. Mater. Res.* 45, 1–27 (2015).
3. A.K. Geim, and K.S. Novoselov, *Nat. Mater.* 6, 183–191 (2007).
4. K.F. Mak, C. Lee, J. Hone, J. Shan, and T.F. Heinz, *Phys. Rev. Lett.* 105, 136805 (2010).
5. W. Zhang, J.-K. Huang, C.-H. Chen, Y.-H. Chang, Y.-J. Cheng, and L.-J. Li, *Adv. Mater.* 25, 3456–3461 (2013).
6. K.S. Novoselov, A. Mishchenko, A. Carvalho, and A.H. Castro-Neto, *Science* 353, aac9439 (2016).
7. E. Singh, P. Singh, K.S. Kim, G.Y. Yeom, H.S. Nalwa, and A.C.S. Appl, *Mater. Interface* 11, 11061–11105 (2019).
8. R. Kochar, and S. Choudhary, *IEEE J. Quantum Electron.* 54, 1–6 (2018).
9. S. Choudhary, and A.K. Garg, *IEEE Trans. Nanotechnol.* 18, 989–994 (2019).
10. B. Qiu, X. Zhao, G. Hu, W. Yue, J. Ren, and X. Yuan, *Nanomaterials* 8, 962 (2018).

11. Y. Liu, N.O. Weiss, X. Duan, H.-C. Cheng, Y. Huang, and X. Duan, *Nat. Rev. Mater.* 1, 1–17 (2016).
12. W.S. Yun, S.W. Han, S.C. Hong, and I.G. KimandJ.D. Lee, *Phys. Rev. B* 85, 033305 (2012).
13. Y. Chang, C.H. Grein, C.R. Becker, X.J. Wang, and S. Sivanan, Hard x-ray gamma-ray, and neutron detector physics XII. *Int. Soc. Opt. Photonics* 7805, 78050A (2010).
14. A.J. Syllaios, P.-K.K. Liao, and B.E. Dean, Infrared detectors: State of the art II. *Int. Soc. Opt. Photonics* 2274, 49–54 (1994).
15. T. Takahashi, and S. Watanabe, *IEEE Trans. Nucl. Sci.* 48, 950–959 (2001).
16. S. Del Sordo, L. Abbene, E. Caroli, A.M. Mancini, A. Zappet Tini, and P. Ubertini, *Sensors* 9, 3491–3526 (2009).
17. A.J. Syllaios, P.-K.K. Liao, and B.E. Dean, Infrared detectors: state of the art II. *Int. Soc. Opt. Photonics* 2274, 49–54 (1994).
18. U.N. Roy, G.S. Camarda, Y. Cui, R. Gul, A. Hossain, G. Yang, J. Zazvorka, V. Dedic, J. Franc, and R.B. James, *Sci Rep* 9, 1620 (2019).
19. U.N. Roy, G.S. Camarda, Y. Cui, R. Gul, G. Yang, J. Zazvorka, V. Dedic, J. Franc, and R.B. James, *Sci. Rep.* 9, 7303 (2019).
20. U.N. Roy, A.E. Bolotnikov, G.S. Camarda, Y. Cui, A. Hossain, K. Lee, W. Lee, R. Tappero, G. Yang, R. Gul, and R.B. James, *APL Mater.* 3, 026102 (2015).
21. W. Choi, M.Y. Cho, A. Konar, J.H. Lee, G.-B. Cha, S.C. Hong, S. Kim, J. Kim, D. Jena, J. Joo, and S. Kim, *Adv. Mater.* 24, 5902–5902 (2012).
22. G. Kresse, and J. Furthmüller, *Phys. Rev. B* 54, 11169–11186 (1996).
23. J.P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* 77, 3865–3868 (1996).
24. S. Grimme, *J. Comput. Chem.* 27, 1787–1799 (2006).
25. H.J. Monkhorst, and J.D. Pack, *Phys. Rev. B* 13, 5188–5192 (1976).
26. M.G. Medvedev, I.S. Bushmarinov, J. Sun, J.P. Perdew, and K.A. Lyssenko, *Science* 355, 49–52 (2017).
27. F. Tran, and P. Blaha, *Phys. Rev. Lett.* 102, 226401 (2009).
28. G. Berghäuser, and E. Malic, *Phys. Rev. B* 89, 125309 (2014).
29. D.J. Griffiths, *Am. J. Phys.* 73, 574 (2005).
30. E.S. Kadantsev, and P. Hawrylak, *Solid State Commun.* 152, 909–913 (2012).
31. Dewia HA, Fangben Mengb BS, Guoa C, Norlingc B, Chenb X, Lima S RSC advances RSC Publishing
32. A. Jain, and A.J.H. McGaughey, *Sci. Rep.* 5, 1–5 (2015).
33. M.M. Ugeda, A.J. Bradley, S.-F. Shi, F.H. da Jornada, Y. Zhang, D.Y. Qiu, W. Ruan, S.-K. Mo, Z. Hussain, Z.-X. Shen, F. Wang, S.G. Louie, and M.F. Crommie, *Nat. Mater.* 13, 1091–1095 (2014).
34. G. Berghäuser, and E. Malic, *Phys. Rev. B* 89, 125309 (2014).
35. B.C. Thompson, and J.M.J. Fréchet, *Angew. Chem. Int. Ed.* 47, 58–77 (2008).
36. Park S, Mutz N, Schultz T, Blumstengel S, Han A, Al- Jarb A, Li LJ, List-Kratochvil EJW, Amsalem P, and Koch N, 2D Mater. 5, 025003 (2018).
37. A. Maniyar, and S. Choudhary, *RSC Adv.* 10, 31730–31739 (2020).
38. Z. Sun, A. Martinez, and F. Wang, *Nat. Photonics* 10, 227–238 (2016).
39. M. Stöhr, T.V. Voorhis, and A. Tkatchenko, *Chem. Soc. Rev.* 48, 4118–4154 (2019).
40. M.-Y. Li, C.-H. Chen, Y. Shi, and L.-J. Li, *Mater. Today* 19, 322–335 (2016).
41. M.-L. Tsai, S.-H. Su, J.-K. Chang, D.-S. Tsai, C.-H. Chen, C.-I. Wu, L.-J. Li, L.-J. Chen, and J.-H. He, *ACS Nano* 8, 8317–8322 (2014).
42. M. Zhao, W. Zhang, M. Liu, C. Zou, K. Yang, Y. Yang, Y. Dong, L. Zhang, and S. Huang, *Nano Res.* 9, 3772–3780 (2016).
43. L.-J. Kong, G.-H. Liu, and Y.-J. Zhang, *RSC Adv.* 6, 10919–10929 (2016).
44. Y. Hinuma, A. Grüneis, G. Kresse, and F. Oba, Band alignment of semiconductors from density-functional theory and many-body perturbation theory. *Phys. Rev. B* 90, 155405 (2014).


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

*International Journal of Research in Engineering and Science (IJRES)*
*ISSN (Online): 2320-9364, ISSN (Print): 2320-9356*
*www.ijres.org Volume 10 Issue 8 ǁ August 2022 ǁ PP. 584-587*

# Floating Solar Power Plant Installation and Indian Overview

Vaibhav[1], Vardhan Suri[2], Dr. J.P. Kesari[3]

[1,2]*Student, Dept. of Software Engineering, Delhi Technological University, New Delhi, India*
[3]*Associate Professor, Dept. of Mechanical Engineering, Delhi Technological Engineering, New Delhi, India*

***Abstract -*** *The high energy demand and the constant depletion of the fossil fuels these problems lead us to shift our focus to renewable energy sources which are not only the future unlimited source of energy, it is also eco-friendly and viable for the environment. Hydro and Wind though are renewable sources but majorly depnends on project location and area. Solar energy on the other hand can be extracted from any place. The major issue with the solar energy is the requirement of land which is scarcely available in the world and even costly to get. But floating solar plants can be installed in any water bodies which will not only reduce the cost of the land but will increase the amount of generation with the cooling effect of water. This research paper concentrates upon the global scenario and indian overview on this new technology will also focus upon the effect ofpanel shade on the ecosystem.*

***Key Words:*** *fossil, renewable , Hydro , wind.*

Date of Submission: 18-08-2022 Date of acceptance: 02-09-2022

## I. INTRODUCTION

Floating solar power plants are quite similar to normal solar power plants installations, except for the fact that they are built on a water body, thereby eradicating the need for land. Many times, even in developed countries with large land areas having plenty of lands available for projects, the land in the vicinity of prime energy consumers (i.e. cities, industrial zones, etc.) is notavailable or highly unaffordable. Floating solar power plant,(FSPV) also popularly known as 'floatovoltaic' is an attractive option at such places since it is independent of land availability. Especially at places where land is a roadblock, floating solar installations are a positive step for energygeneration.

At present, among the 60+ countries actively pursuing the deployment of FPV (see Map 1 below), more than 35 countries are home to an estimated 350 operational FPV systems, which up until the end of August 2020 had a cumulative capacity of approximately 2.6 GW. Although still considered a niche, FPV is projected to experience an average growth rate of above 20% in the coming five years.

In this context, Asia is expected to account for roughly two thirds of the global demand, mostly driven by China, India, South Korea, Taiwan, Thailand and Vietnam.

**FSPV Scenario in the UK**

With more than 13 MWP of installed capacity, the UK is the fifth largest country in terms of FSPV installation. Like many other Asian countries, lack of space for land-mounted PV in the country is the main reason for going towards FSPV plants. The majority of the plants in the UK are installed on irrigation and water treatment ponds. In March 2016, the world's first deep water reservoir (maximum depth of 18.4)

.

**FSPV Scenario in Japan**

With more than 210 MWP of installed capacity, Japan is aleading country in terms of total numbers of installation and home to 73 of theworld's 100 largest FSPV plants. The majority of these plants are installed on man-made waterbodies having main purpose to retain rainwater orirrigation.

**FSPV Scenario in China**

To date, China is the world's largest market for FPV. Currently, FPV systems are either deployed as a result of a bidding scheme, thus eligible for a FIT granted over 20 years or as a so-called grid-parity project, i.e., without any form of subsidy support. To illustrate the on-going demand for FPV, in June this year, China's Datang Power released a tender seeking several bids for a total capacity of 820 MW of FPV to be installed across China by December 2021.

| Country | Water body type | Typical range of % ofwater surface area covered by FSPV plants | Typical range of depth of water body on which FSPV plants are installed (in meters) | Typical range of water level variation(in meters) |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| China | Irrigation ponds | 10%–30% | 3.5–14.1 | 3.5– 8.0 |
| | Industrial ponds | No information | No information | No information |
| | Large water bodies | 10%–40% | No information | No information |
| | Mining ponds | 10%–20% | 3–12.5 | 4.8 |
| Japan | Irrigation ponds | 10%–70% | 1.8–15.1 | 1.8–15.1 |
| | Industrial ponds | No information | No information | No information |
| | Water storage reservoirs | 15%–86% | 3.0–5.0 | 3.0–5.0 |
| Taiwan | Irrigation ponds | 11%–28% | 3.6–4.6 | 2.4–4.6 |
| | Industrial ponds | 15% | 14 | 5 |
| | Water storage reservoirs | 7%–10% | No information | No information |
| UK | Irrigation Ponds | 2% –15% | 4.0–18.4 | 4.0–18.4 |
| | Water treatment plant | 48% | 10 | 10 |

**The republic of Korea Scenario in Europe**

Korea was among the early mover to have adopted FSPV. Till date the country has installed more than 79 MWp of FSPV plants and more than 2300 MWp are in pipeline. Korea was the first country to install a tracking-based FSPV plant having a capacity of 465 kW in 2014. Recently, the Ministry of Trade, Industry, and Energy (MOTIE) has announced its plans for developing 2.1 GW of FSPV- based projects at Saemangeum Seawall Dyke located on the southwest coast of the country. The project will be built in two stages . The first stage with a capacity of 1.2 GW is to be completed by 2022 and the remaining 900 MW by 2025 in the second stage. Upon completion, the project will be 14 times larger than the present world's largest 150-MW FSPV-based project of China.

**Indian FSPV Scenario**

The FSPV as a technology is still in the nascent stages of development in India. The journey started with a 10kW FSPV plant on a pond in Rajarhat, Kolkata in 2015. In 2016, NTPC installed country's largest 100kW plant on a reservoir of its combined cycle power plant situated in Kerala's Kayamkulam district. Later in December same year, Kerala State Electricity Board started its operation of 500kW plant at Banasura Sagar reservoir in Wayanad district replacing NTPC's 100kW as a largest FSPV-based plant. The plant is actually a scaled-up version of the 10kW plant commissioned in January 2016 at the same location. The plant was able to bring some confidence to FSPV promoters by successfully surviving the recent flood ravage in the state.

**Investment Requirement in India**

Recent bids results indicate a sharp decline in the investment cost for FSPV. In the latest tender result of the country's first large-scale FSPV plant of 70MW capacity, cost as low as INR 35 per Wp has been quoted by developers. This is the lowest cost achieved in the entire world so far. Further considering last year's trends, one can see the 45 percent reduction in the cost, which is significant. Even though the costs have decreased, it is still very early to come to any generalization, since the FSPV as a technology is still in its early stage of market penetration. The reduction in costs, particularly in India may be linked to many factors like – decrease in the cost of floaters because of improvement in manufacturing process, reduction in the material cost, reduction in thickness of solar panels. While falling cost is a welcome trend, it is also important to check whether this is not impacting the overall quality of the projects, since degradation in the quality of FSPV projects has higher potential to impact local biodiversity compared to ground- mounted solar PV. Hence it is essential to form proper guidelines/standards to keep the things in check from the initial stage of market development in the FSPV technology.

The cost also majorly depends on these factors

1. Depth of Water body
2. Project location
3. Water-level variation
4. Site conditions – speed of wind and direction, of it solar irradiations, ambient temperature, humidity levels, etc.
5. Size of the plant

**2.** Floating Solar and Jobs

A new case study, *Employment Potential of Emerging Renewable Energy Technologies – Insights from the Floating Solar Sector*, examines the direct employment potential across the floating solar project development cycle and captures the sector's skill development needs. Estimates are drawn from project-based surveys and interviews with manufacturers, developers, and engineering, procurement, and construction (EPC) providers. The study provides insight into the operational strategies and team structure, in addition to discussing the typical duration of different phases of project development and the corresponding workforce employed. The case study is developed by the Council on Energy, Environment, and Water (CEEW) and Skill Council for Green Jobs (SCGJ) along with NRDC.

An FPV solar project includes four stages for a successfuldeployment

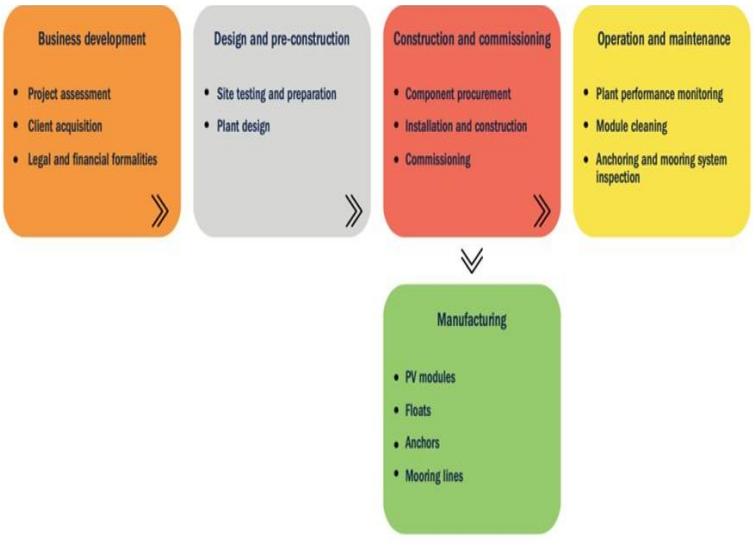
Key Data

- A small-scale FPV plant (capacity <1 MW) directly employs 58 workers while a mid-scale (capacity <10 MW)plant employs 45 personnel, over the course of their deployment.

- FPV generates indirect job opportunities through manufacturers of specialized components like floats, anchors, and mooring system as well as domestic modulemanufacturers.

- FPV offers opportunities for people qualified in hydraulic engineering, marine architecture, and plastic blow- molding techniques in addition to those required in ground- mounted solar operations.

- By setting time-based targets for FPV capacity, the Indian government can widen the employment potential of this sector, bolster efforts to drive the COVID-19 economic recovery and achieve its Paris Agreement climate goals




Key Recommendations

1. The Ministry of New and Renewable Energy (MNRE) should specify standards for PV modules used in theFPV sector.

2. The Ministry of Skill Development and Entrepreneurship should develop training programmes on specialised skills relevant to the FPV industry such as marine architecture and hydraulic engineering.

3. The FPV industry should regularly release data onemployment and skill

| Project phase | Duration (days) | | Number of people engaged | |
|---|---|---|---|---|
| | Small-scale | Mid-scale | Small-scale | Mid-scale |
| Business development | 30 | 110 | 11 | 12 |
| Design and pre-construction | 15 | 14 | 6 | 2 |
| Construction and commissioning | 53 | 195 | 36 | 27 |
| Maintenance | 30* | Not available | 5** | 4** |

## 3. CHALLENGES

1. Unavailability of FSPV specific Standards.
2. Unavailability of water body data.
3. FSPV plant components safety and its long termreliability.
4. Absence of local manufacturing.
5. Unavailability of Bathymetry and other studies ofwater.

Bathymetry – the measurement of depth of waterbody.

6. Also these impacts on various fields due to largeFSPV

1. Impact on local marine aquaculture due to the blockage of sunlight reaching the water surface
2. .Impact on the fishing pattern
3. Also, it can reduce oxygen level of water whichaffects water bodies badly.
4. Impact on drinking water quality

**References**


[1]. http://sunengy.com/
[2]. http://social.csptoday.com/technology/solar-island-groundbreaking-csp-technology
[3]. http://www.ciel-et-terre.net/
[4]. http://www.kseboa.org/kseb/floating-solar-plant-at-banasura-sagar-reservoir-inaugurated- 23014084.html
[5]. http://www.theguardian.com/environment/2014/sep/29/first-floating-solar-farm-built-in-uk
[6]. http://cleantechnica.com/2014/07/02/india-plans- worlds-largest-floating-solar-power-project-50-mw/
[7]. http://www.thehindu.com/news/cities/kolkata/floati ng-solar-power-plant-in-kolkata/article6747770.ece
[8]. http://www.sersc.org/journals/IJSEIA/vol8_no1_2 014/7.pdf [9]http://onlinepresent.org/proceedings/vol25_2013/6 6.pdf
[9]. http://www.ijsgce.com/uploadfile/2013/1231/201 31231022201908.pdf
[10]. http://waset.org/publications/9999567/a-case-study-on-suitable-area-and-resource-for- development-of-floating-photovoltaic-system
[11]. http://techxplore.com/news/2015-04-japan-solar-power-hyogo-prefecture.html
[12]. http://forwardthinkingarchitecture.com/SMART- FLOATING-FARMS-SFF Floating Solar PV Potential in Large Reservoirs in India (IJIRST/ Volume 2 / Issue 11/ 018) All rights reserved by www.ijirst.org
[13]. D.P.Kothari, Rakesh Ranjan and K.C.Singhal, "Renewable Energy Sources and Emerging Technologies", Prentice-Hall of India, New Delhi,2007; 2nd edition, 2011.
[14]. DP Kothari , Majid and Rizwan, "Grid integration of Solar Photovoltaic Systems" Taylor andFrancis.UK.2016
[15]. Analysis of Electricity Consumption under a photovoltaic micro-grid system in India. – Toru Kobayakawa , Tara C Kandpal , Solar Energy 116(2015) 177- 183
[16]. Decentralized v grid electricity for rural India -The economic factors , Chandra Shekhar Sinha and Tara Chandra Kandpal , 0301-4215/91/050441-08 ©1991 Butterworth-Heinemann Ltd
[17]. Providing Electricity access to remote areas in India – an approach towards identifying potential areas for decentralized electric supply , M R Nouni , S C Mullick , T C Kandpal ,Renewable & SustainableEnergy Reviews 12(2008) 1187-1220

*Sādhanā* (2022)47:195
https://doi.org/10.1007/s12046-022-01972-6

© Indian Academy of Sciences

# Heavy metals concentrations and risk assessment in the sediment of Ganga River between Kanpur and Prayagraj, U.P., India

MOHIT AGGARWAL*, S ANBUKUMAR and T VIJAYA KUMAR

Department of Civil Engineering, Delhi Technological University, Delhi 110042, India
e-mail: mohitjvg@gmail.com

MS received 4 April 2022; revised 29 June 2022; accepted 9 August 2022

**Abstract.** This study analyses the heavy metal contamination in freshly deposited sediment of 210 km stretch of River Ganga between Kanpur and Prayagraj, Uttar Pradesh (U.P.), India. As per APHA (2012)30 digested samples were prepared after collecting from 10 sampling sites. Then the samples were analyzed for 8 heavy metals namely Cadmium (Cd), Chromium (Cr), Copper (Cu), Iron (Fe), Manganese (Mn), Nickel (Ni), Lead (Pb), and Zinc (Zn) by Atomic Absorption Spectroscopy (AAS) (Shimadzu AA-6300). Risk assessment was done by employing four pollution indices. The average values of contamination factor (CF) indicate that the level of contamination by Pb is moderate ($1 < CF < 3$) and significant by Cd ($3 < CF < 6$). Whereas, the average value of Geo-accumulation index (Igeo) indicates that the sediments are uncontaminated or in a moderate level of contamination by Pb ($0 < Igeo < 1$) and in a moderate level of contamination by Cd ($1 < Igeo < 2$), which may have harmful effects on Ganga River ecology. The values of Pollution load index (PLI) indicate that the Ganga River sediment has a low level of contamination. Based on various pollution indicators, it has been observed that concentrations of Cd and Pb in the river sediment is a matter of concern and may lead to further deterioration of sediment quality in near future due to increasing anthropogenic activities in the river basin, hence proper management strategies should be taken to control the direct dumping of both domestic and industrial wastewater in the river.

**Keywords.** River health; river sediment; heavy metals; enrichment ratio; contamination factor; geo-accumulation index ($I_{geo}$); pollution load index (PLI).

## 1. Introduction

Heavy metal contamination in rivers is a serious threat to aquatic life as they are highly toxic, long-term persistent in the environment, and get accumulated in water bodies [1, 2]. Causes of heavy metal contamination in water bodies include atmospheric deposition [3] erosion of soil [4], disposal of effluent (treated or untreated) [5], use of fertilizers and pesticides which contain heavy metals [6], surface run-off [7] and various chemicals from agricultural, urban and industrial activities [8, 9]. Some studies have also suggested that the quality of the sediments is an indication of pollution levels in the river and can also help to detect the history of pollution from the past to the present [7]. The presence of elevated levels of toxic heavy metals in sediments can pose a potential risk to the health of humans as they may transfer into aquatic species and then enter the food chain [10, 11].

In India, Ganga is the largest river with a stretch of 2525 km originating from Gangotri Glacier and ending in the Bay of Bengal, catering to around 400 million people of India. River water quality is degraded in the last few decades [12, 13]. Discharge from industries is a primary source of pollution in the Ganga River [14, 15]. Only 3500 MLD of sewage is treated out of 8250 MLD of total sewage which is generated from more than 220 cities in the Ganga River basin. 2460 MLD is dumped directly into the Ganga River, 4510 MLD is discharged into its tributaries, and the remaining is dumped on land or low-lying areas that leach downstream [15]. Since 1985, the government has taken many steps but still, the quantity of untreated sewage being discharged into the Ganga River has increased [12]. In addition to sewage from urban areas, approximately 2500 MLD sewage load of industrial units which include textiles, sugar mills, electroplating, power plants, distilleries, paper and pulp, etc. discharge their waste directly into the Ganga River [16]. The Middle section of the Ganga River i.e. from Kannauj to Varanasi receives most of the industrial discharge to the river majorly from tannery industries [17].

Heavy metal studies have been done on various rivers including the Ganga River, India [18], Indus river, Pakistan [19], Gomti River, India [20], Hindon River, India [21] and Nile River, Egypt [22].

*For correspondence
Published online: 20 September 2022

The main objectives of this study are (i) to find the heavy metal concentrations in the Ganga River sediments of the selected stretch, (ii) to identify their source i.e., naturally enriched or anthropogenic using various pollution indices like enrichment ratio (ER) and geo-accumulation index (Igeo), (iii) to study spatial distribution pattern of heavy metal concentrations in the selected stretch and (iv) to evaluate the risk of heavy metals on the environment in the selected area by comparing the current metal concentrations values with standard quality guidelines. This study is also conducted to access the pollution level in the Ganga River and provides a base for the environmental health risk assessment of the consumers and also to manage the problems related to Ganga River pollution in the studied area.

Though in the past, studies were conducted to assess heavy metal contamination in the Ganga River water [17, 23] and Ganga River bed sediment [24, 25] but very limited data is available for the Jajmau location of Ganga River which is in the downstream of Kanpur city. Jajmau is a hub of tannery industry involving heavy metals in their manufacturing process and other industries like paints, electroplating, leather, etc. which lead to a huge untreated sewage load in the Ganga River which could be a possible source of heavy metal contamination in the Ganga River [26]. Further, this can also lead to contamination to downstream of the Ganga River in the holy city of Prayagraj where lakhs of people take holy baths daily and do many rituals and spiritual activities. This gave us the motivation to select this particular stretch of Ganga River between Kanpur and Prayagraj city for this study. In addition, the entire selected region is used for agricultural purposes using various pesticides and fertilizers which can contaminate River Ganga with heavy metals. Besides, chromium tannery effluent contain some other heavy metal like lead, cadmium and copper which are toxic in excess amount. Cu, Fe, Mn, Ni, and Zn metals are required as nutrients in trace amount for life processes in plants and microorganisms but become toxic at higher concentrations wheras Pb, Cr, and Cd metals have no biological function and are toxic elements. Therefore, in this study, the deposition of these eight heavy metals namely Cd, Cr, Cu, Fe, Mn, Ni, Pb, Zn was analyzed to evaluate its spatial distribution in the Ganga River sediments.

## 2. Methodology

### 2.1 *Study area, sampling sites and collection of sediment samples*

The Ganga River flows through Kanpur and joins the Yamuna river in the Triveni Sangam in Prayagraj city, a sacred confluence in Hinduism. Between Kanpur and Praygraj, the Ganga River flows 230 km through these 10 selected sampling sites: Jajmau (S1), Maharajpur (S2), Sarsaul (S3), Dalmau—Fatehpur Bridge (S4), Unchahar (S5), Nawabganj (S6), Manikpur (S7), Shringverpur (S8), Phaphamau (S9) and Sangam (S10) (table 1; figure 1).

Sediment sampling was done in the summer (March–June) between 7:00 AM and 9:00 AM in the year 2019. From each sampling location, 3 sediment samples were collected from a depth of 10–15 cm from the top surface of the river bank of the running river. In total 30 samples were collected from 10 selected sites using standard procedure. Each sediment sample of approximately 500 g was collected with the help of a grab sampler and then packed in impermeable polythene bags [27]. All sediment samples were then brought to the laboratory and were air dried for 72 h before further analysis.

### 2.2 *Sample preparation and analytical aspect*

According to APHA (2012), collected samples of sediment were digested and analyzed to find the concentrations of heavy metals. Sediment samples were dried in the oven at 100 °C for 1–2 h. to get the constant mass and then crumbled and ground to get $< 63$ µm fractions. Sediment samples each weighing 1 g were digested by adding 10 mL 1:2 solution of perchloric acid and nitric acid initially placed on a hot plate for 60 min at low temperature and then at 120 °C temperature until fumes started arising and the solution became clear and transparent. Then digested samples were filtered using Whatman filter paper 42. Then 0.1 N $HNO_3$ is added to dilute the filtrate till the final volume of 50 mL is reached. Finally, the concentrations of heavy metals namely Cd, Cr, Cu, Fe, Mn, Ni, Pb, and Zn were measured in the digested samples using atomic absorption spectroscopy (AAS) (Shimadzu AA-6300).

### 2.3 *Calculation of enrichment ratio, geo-accumulation index, contamination factor and pollution load index*

Enrichment ratio (ER) helps to know the level of environmental pollution and to know whether the source of pollution is natural or anthropogenic, based on the standards of heterogeneous sediments [7, 28]. Fe is used as reference metal and heavy metal concentrations were found with respect to it. ER of metals in the sediment samples for all sites was calculated using equation (1), [11].

$$\mathrm{ER} = \frac{[(X_X)/(X_{Fe})]_s}{[(X_X)/(X_{Fe})]_b} \tag{1}$$

where $[(X_x)/(X_{Fe})]_s$ denotes the ratio of heavy metal X concentration and Fe concentrations in the sample and $[(X_x)/(X_{Fe})]_b$ denotes the ratio of metal X concentration and

**Table 1.** Sites selected for sampling between Kanpur and Prayagraj with their GIS location.

| Name of site | Code | Location | District and state |
|---|---|---|---|
| Jajmau | S1 | 26°26′17″N 80°27′43″E | Kanpur, U.P. |
| Maharajpur | S2 | 26°22′42″N 80°29′26″E | Kanpur, U.P. |
| Sarsaul | S3 | 26°18′13″N 80°32′35″E | Kanpur, U.P. |
| Dalmau-Fatehpur Bridge | S4 | 26°03′18″N 81°01′56″E | Fatehpur, U.P. |
| Unchahar | S5 | 25°53′22″N 81°12′59″E | Raebareli, U.P. |
| Nawabganj | S6 | 25°49′06″N 81°19′46″E | Pratapgarh, U.P. |
| Manikpur | S7 | 25°46′03″N 81°23′58″E | Pratapgarh, U.P. |
| Shringverpur | S8 | 25°35′20″N 81°38′01″E | Prayagraj, U.P. |
| Phaphamau | S9 | 25°30′24″N 81°51′59″E | Prayagraj, U.P. |
| Sangam | S10 | 25°25′41″N 81°53′20″E | Prayagraj, U.P. |

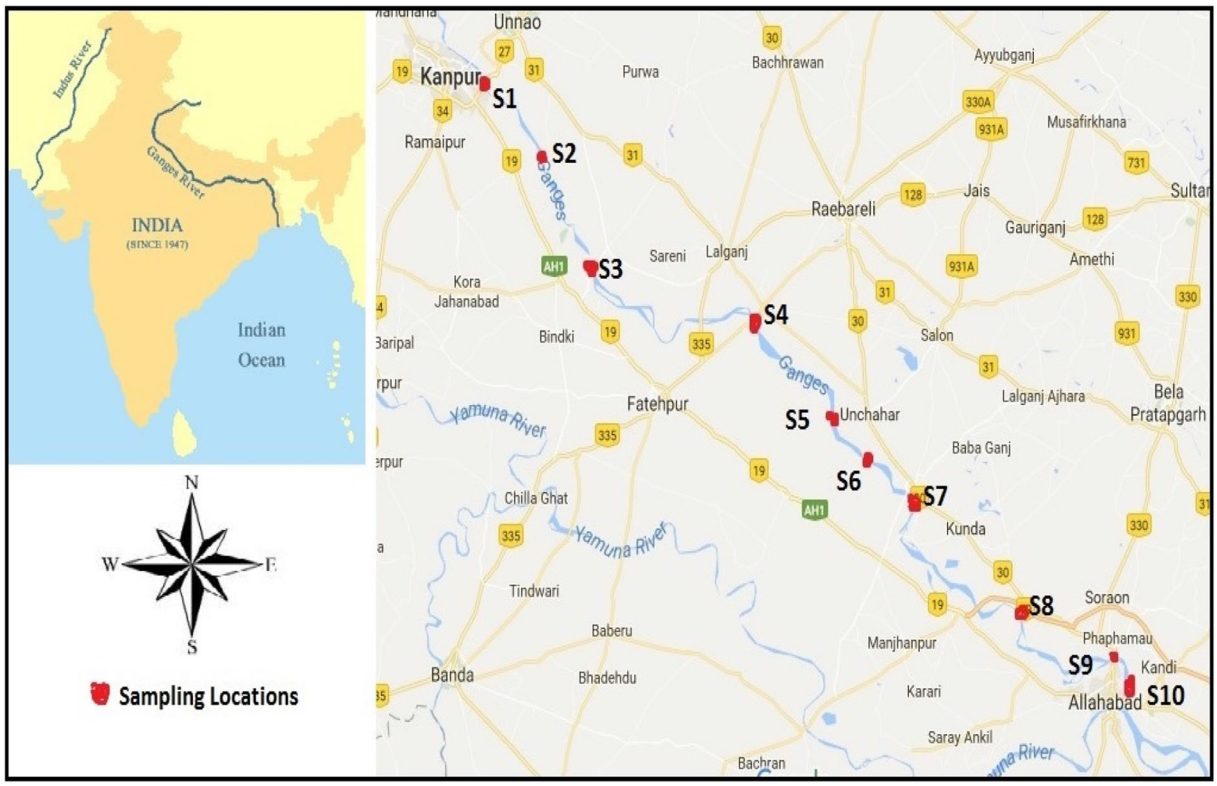



**Figure 1.** Sampling sites between Kanpur and Prayagraj along the Ganga River, Source: Google maps.

Fe concentration taken from the background values. The background values of concentrations of Fe, Ni, Cd, Cu, Cr, Mn, Pb and Zn were taken from geometric mean values given by Singh *et al* [24] mentioned in table 3. The Enrichment levels were determined based on ER values using table 2.

For the determined heavy metal concentrations, the geo-accumulation index (Igeo) was calculated by using equation (2) given below [30] This index is used to assess the anthropogenic impact.

$$\text{Igeo} = \text{Log}_2 \frac{\text{Cn}}{1.5*\text{Bn}} \quad (2)$$

where Cn denotes the concentration of metal in (ppm) determined in the sediment sample, Bn denotes the geo-chemical background concentration value in ppm (table 3) of that metal in the background sample and factor 1.5 is used to reduce the errors in the background values which may be at due to lithogenic effects. Igeo values were used to report the Igeo class and pollution levels as given in table 2.

**Table 2.** Enrichment Levels and Pollution levels corresponding to ER [29] values and Igeo values [30], respectively.

| ER value | Enrichment level | Igeo value | Igeo class | Pollution level |
|---|---|---|---|---|
| Less than 2 | Mineral depletion | Less than 0 | 0 | Unpolluted |
| Between 2 and 5 | Moderate | Between 0 and 1 | 1 | Unpolluted to moderately polluted |
| Between 5 and 20 | Significant | Between 1 and 2 | 2 | Moderately polluted |
| Between 20 and 40 | Very high | Between 2 and 3 | 3 | Moderately to strongly polluted |
| Greater than 40 | Extremely high | Between 3 and 4 | 4 | Strongly polluted |
| | | Between 4 and 5 | 5 | Strongly to very strongly polluted |
| | | Greater than 5 | 6 | Very strongly polluted |

**Table 3.** Concentration (in ppm) of different metals at different sites in Ganga River sediments.

| Name of the site | Code | Fe | Mn | Cr | Zn | Pb | Ni | Cu | Cd |
|---|---|---|---|---|---|---|---|---|---|
| Jajmau | S1 | 19503 | 482 | 31.4 | 34.8 | 35.8 | 33.4 | 5.7 | 1.15 |
| Maharajpur | S2 | 19352 | 541 | 100.2 | 39.1 | 31.5 | 33.4 | 8.55 | 2.3 |
| Sarsaul | S3 | 15110 | 367 | 42.5 | 29.6 | 29.5 | 32.2 | 4.75 | 2.3 |
| Dalmau-Fatehpur Bridge | S4 | 23038 | 654 | 65.8 | 33.8 | 36.6 | 32.2 | 4.75 | 3.45 |
| Unchahar | S5 | 20725 | 568 | 61.7 | 37 | 34.5 | 34.5 | 9.5 | 2.3 |
| Nawabganj | S6 | 30017 | 679 | 85 | 56.9 | 37.8 | 40.8 | 17.1 | 1.15 |
| Manikpur | S7 | 29017 | 462 | 95.1 | 58.1 | 39.6 | 42.6 | 15.2 | 2.3 |
| Shringverpur | S8 | 19700 | 352 | 38.5 | 35.9 | 30.5 | 35.7 | 6.65 | 2.3 |
| Phaphamau | S9 | 17286 | 349 | 34.4 | 30.6 | 29.5 | 33.4 | 5.7 | 0.95 |
| Sangam | S10 | 16974 | 457 | 56.7 | 27.5 | 31.5 | 33.4 | 3.8 | 1.15 |
| Mean concentration | | 21072.2 | 491.1 | 61.13 | 38.33 | 33.68 | 35.16 | 8.17 | 1.94 |
| SD | | ± 200 | ± 10 | ±10 | ±2 | ±3 | ±3 | ±2 | ±8 |
| Background values (Bn)* | | 40346 | 1764 | 147 | 105 | 22 | 46 | 55 | 0.58 |
| Indian River System (IRS) Average** | | – | 607 | 87 | 16 | – | 37 | 28 | – |
| World Surface Rock Average*** | | 35900 | 720 | 71 | 127 | 16 | 49 | 32 | 0.20 |
| SQG**** | TEC | 20000 | 460 | 43.4 | 121 | 35.8 | 22.7 | 31.6 | 0.99 |
| | PEC | 40000 | 1100 | 111 | 459 | 128 | 48.6 | 149 | 4.98 |

*Values taken from Singh *et al* [24]

**Values taken from Subramanian *et al* [34]

***Values taken from Martin and Maybeck [35]

****Values taken from MacDonald *et al* [32]

Pollution load index (PLI) is an indicator used to conclude the overall heavy metal pollution load at each site and is calculated by using equation (3) [31]. PLI = 0 indicates a perfect state of pollution; PLI = 1 points indicate only baseline levels of pollutants present and PLI above 1 would indicate progressive deterioration of sites.

$$\mathrm{PLIn} = (\mathrm{CF1}^{*}\ \mathrm{CF2}{*}\ldots..\ {*}\mathrm{CFn})^{1/n} \tag{3}$$

where CFn is the contamination factor (CF) of the heavy metal 'n' in the sediment sample calculated using equation (4). CF is a quantification of the degree of contamination relative to either average crustal composition of a respective metal or to the measured background values from geologically similar and uncontaminated area.

$$\mathrm{CFn} = \frac{\mathrm{Cn}}{\mathrm{Bn}} \tag{4}$$

CF less than 1 indicates a low degree of contamination; CF between 1 and 3 indicates moderate contamination, CF between 3 and 6 indicates a considerable degree of contamination, and CF above 6 indicates very high contamination.

### 2.4 *Comparison of data using sediment quality guidelines (SQGs)*

The concentrations of heavy metals in the Ganga River sediments are determined and compared with the corresponding SQGs values [32] in table 3. SQGs help in assessing the extent to which concentrations of heavy metals in sediments may adversely affect aquatic plants and animals and are designed to interpret the quality of sediment [33]. Two types of SQGs are developed for the

ecosystem in freshwater [32]: probable effect concentration (PEC) and threshold effect concentration (TEC). PEC represents the concentration above which adverse effects are expected to occur and the TEC represents the concentration below which adverse effects are unlikely to occur.

## 3. Results and discussions

### 3.1 *Concentrations and spatial distribution of heavy metals in the sediments of the selected sites*

Heavy metal concentrations in the sediment samples collected from the selected sites are reported in table 3. On average, the heavy metal mean concentrations followed a decreasing order of i.e. Fe > Mn > Cr > Zn > Ni > Pb > Cu > Cd where Fe is in the highest concentrations and Cd is in the lowest concentration in the collected sediment samples.

3.1.1 *Lead (Pb)* Lead is a toxic and unessential heavy metal. It is generated from both natural and anthropogenic sources. The main sources of lead concentrations are airborne particles, effluents of leather factories, forest fires, emissions from vehicles, volcanoes, waste incineration, paints, and pesticides. Earth's crust contains 15–20 ppm of Pb concentration [36]. Pb concentrations in this study ranged between 29.5 and 39.6 ppm with a mean value of 33.68 ppm. Pb concentration in the sites follows this order S3 < S9 < S8 < S2 < S10 < S5 < S1 < S4 < S6 < S7 with highest value at site S7 and lowest at site S3. Values of Pb concentrations at all sites exceeded the world surface rock average of 16 ppm indicating an increase in anthropogenic activities. Higher levels of Pb is a threat to fisheries resources, whereas, in plants, it initially results in enhanced growth, but from a concentration of 5 ppm onwards, this is counteracted by severe growth retardation, discoloration, and morphological abnormalities (figure 2).

3.1.2 *Cadmium (Cd)* Cadmium is an unessential element and adversely affects the growth of plants. Cd is a by-product of Zn and Pb mining and smelting and is more mobile in aquatic environments than most of the other metals. Cd is released into the environment by metallurgical industries, atmospheric deposits, power plants, fertilizers, natural atmospheric processes, municipal solid waste (MSW) incineration, and discharge of toxic effluents from industries and wastewater treatment units [37]. It is extremely important to study this contaminant because it is highly toxic and water soluble. Earth's crust contains 0.1–0.5 ppm of Cd [38]. The Cd concentration in this study ranged between 0.95 and 3.45 ppm with a mean value of 1.94 ppm. Cd concentration in the sites followed this order S9 < S1 < S6 < S10 < S2 < S3 < S5 < S7 < S8 < S4 with highest at S4 and lowest at S9. Values of Cd concentrations at all sites exceeded the World Surface Rock Average of 0.2 ppm indicating an increase in anthropogenic activities. This can cause toxicity to some aquatic species in the vicinity (figure 3).

3.1.3 *Copper (Cu)* Copper is an essential nutrient in freshwater and river sediments for the growth of aquatic life, but its higher concentration is toxic. It is induced in the environment from natural sources such as decaying vegetation, volcanic eruptions, sea spray, forest fires, and also from anthropogenic activities like wastewater from industries and municipal corporations [39]. After several natural processes, the dissolved Cu finally gets adsorbed in the sediment. In this study, Cu concentration ranged from 3.8 to 17.1ppm with a mean value of 8.17 ppm. Cu concentration in the sites followed this order S10 < S3 < S4 < S1 < S9 < S8 < S2 < S5 < S7 < S6 with highest at S6 and lowest at S10. Values of Cu concentrations at all sites are well below the World Surface Rock Average of 32 ppm and IRS average value of 28 ppm indicating Cu concentrantions in river sediment are safe for aquatic life (figure 4).

3.1.4 *Zinc (Zn)* Sediment may contain up to 100 ppm of zinc in its natural state. Human activities and natural processes are the sources of Zn. The use of fertilizers for agricultural purposes in the catchment of the Ganga River leads to increase in Zn concentrations in river sediments. In

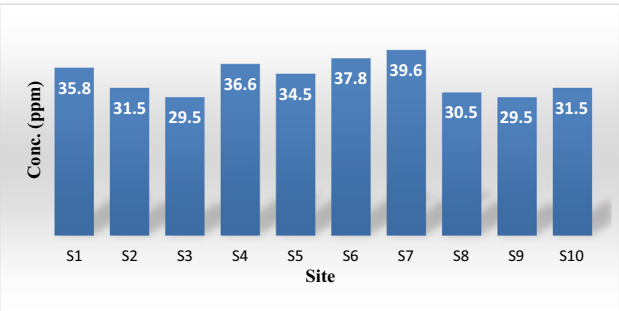



**Figure 2.** Variation in Pb concentrations.

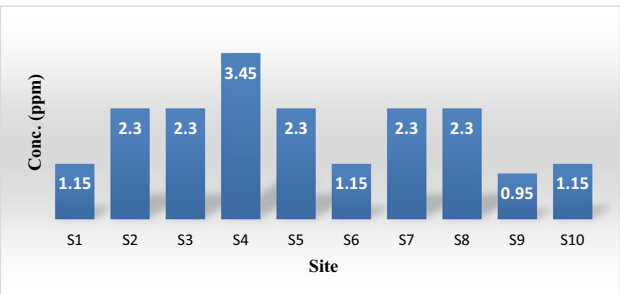



**Figure 3.** Variation in Cd concentrations.

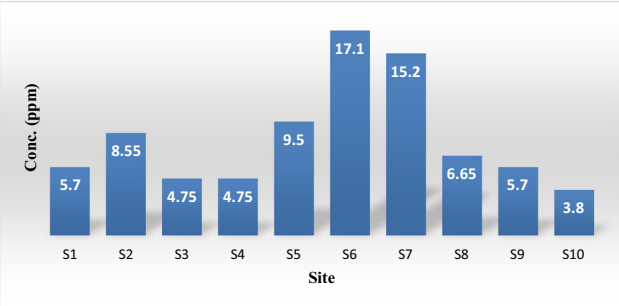



**Figure 4.** Variation in Cu concentrations.

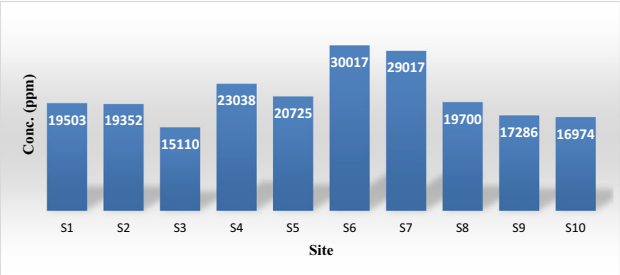



**Figure 6.** Variation in Fe concentrations.

this study, concentrations of Zinc ranged from 27.5 to 58.1ppm with a mean value of 38.33 ppm. Cu concentration in the sites followed this order S10 < S3 < S9 < S4 < S1 < S8 < S5 < S2 < S6 < S7 with highest at S7 and lowest at S10. Values of Zn concentrations at all sites are less than the World Surface Rock Average of 127 ppm; however, it exceeded the IRS average value of 16 ppm [34] which may cause adverse effect on aquatic biota (figure 5).

3.1.5 *Iron (Fe)* Iron (Fe) concentrations were found to be highest among all elements in Ganga River sediment due to many natural reasons like weathering, erosion, etc., and human activities such as urban discharge, industrial effluent, construction and demolition wastes, municipal solid waste, and agricultural activities. In this study, Fe concentration ranged between 15,110 and 30,017 ppm with a mean value of 21072.2 ppm (table 3). Fe concentration in the sites followed this order S3 < 10 < S9 < S2 < S1 < S8 < S5 < S4 < S7 < S6 being highest at S6 and lowest at S3. Values of Fe concentrations at all sites are less than the world surface rock average of 35900 ppm (figure 6).

3.1.6 *Manganese (Mn)* Mn concentration in the earth's crust ranges between 40 and 900 ppm [40]. Sewage sludge, municipal wastewater, mining waste, waste from various metal processing units, alloy production unit emissions, and fossil fuel combustion is the source of manganese in sediments. In this study, Mn concentration lies between 349 and 679 ppm with a mean value of 491.1 ppm. Mn concentration in the sites followed this order S9 < S8 < S3 < S10 < S7 < S1 < S2 < S5 < S4 < S6 being highest at S6 and lowest at S9. Mn concentrations at sites S4 and S6 exceeded the IRS average value of 605 ppm [34] whereas concentrations at all sites are less than the world surface rock average of 750 ppm (figure 7).

3.1.7 *Nickel (Ni)* Ni is used in alloys with chromium, iron, copper, and zinc. It is widely used in the industry such as fuel production, household appliances, food production, electroplating, pigments, ceramics, jewelry manufacturing, heat exchangers, magnets, coins, medical prostheses, batteries, etc. In this study, Ni concentration remained below its baseline (46 ppm) ranging between 32.2 and 42.6 ppm with a mean value of 35.16 ppm, indicating a less contaminated condition by this metal. Ni concentration in the sites followed this order S3 < S4 < S1 < S2 < S9 < S10 < S5 < S8 < S6 < S7 with lowest at S3 and highest at S7. Ni concentrations at sites S6 and S7 exceeded the IRS average value of 37 ppm [34] whereas concentrations at all sites are less than the world surface rock average 49 ppm. This indicates the low level of the contamination of the river sediment (figure 8).

3.1.8 *Chromium (Cr)* Chromium has an average concentration of 100 ppm in the earth's crust [41]. Chromium compounds are used in dyes, paints factories, and leather industries, these compounds get transported in the soil and groundwater of industrial sites. Paints base containing chromium is used for automotive refinishing applications that lead to chromium pollution in the Ganga River. In this study, Cr ranged from 31.4 to 100.2 ppm with

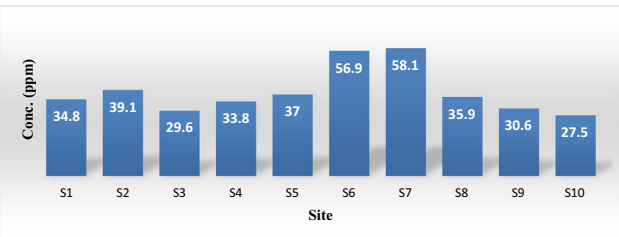



**Figure 5.** Variation in Zn concentrations.

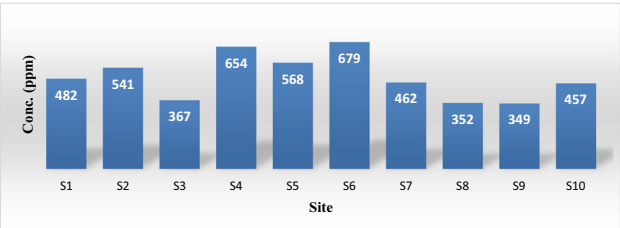



**Figure 7.** Variation in Mn concentrations.

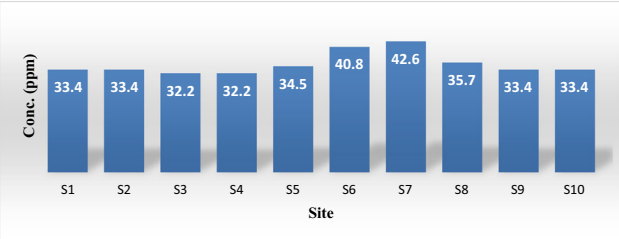



**Figure 8.** Variation in Ni concentrations.

a mean of 61.13 ppm. Cr concentration in the sites followed this order S1 < S9 < S8 < S3 < S10 < S5 < S4 < S6 < S7 < S2 with lowest at S1 and highest at S2. Cr concentrations at sites S2, S6, and S7 exceeded the both IRS average value and world surface rock average of 87 ppm and 71 ppm respectively whereas concentrations at all other sites are less than the IRS average value and the world surface rock average. Increase Cr concentration at some sites is alarming can cause lethality to some aquatic species in the river system (figure 9).

## 3.2 *Distribution of heavy metals in Ganga River sediments*

Following pattern based on the decreasing order of concentration of heavy metals at various sites was observed (see table 3): S7 > S6 > S2 > S5 > S4 > S8 > S1 > S3 > S10 > S9. Among these sites, the Manikpur (S7) site is the most contaminated site receiving the most industrial discharges from the surrounding upstream Kanpur city. All sampling sites receive toxic waste from the upstream flow and the catchment agricultural areas. Among the heavy metals examined, Cd and Pb concentrations exceeded the bottom limit of 0.2 ppm and 16 ppm by up to 15 and 2.4 times, respectively. Industrial cities like Kanpur and Prayagraj lead to industrial and municipal waste in the water of the Ganga River. At some sites, the Cd concentration exceeded the permissible reference point due to low water flow, which also leads to the deposition of different heavy metals in the sediments of the Ganga River at these sites.

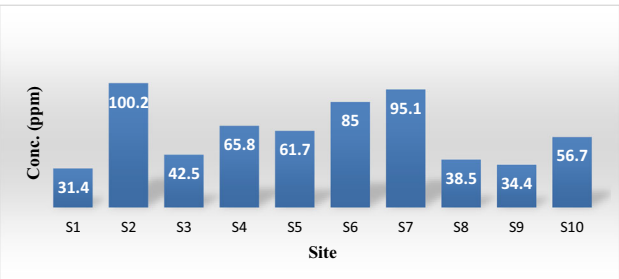



**Figure 9.** Variation in Cr concentrations.

## 3.3 *Estimation of pollutant indices*

### 3.3.1 *Enrichment ratio (ER) and geo-accumulation index (Igeo)*

ER depicts the contamination levels of heavy metals and their source of contamination in a specific environment [42, 43]. ER < 2 indicates the crustal source of the elements in the sediment, whereas ER > 2 reflects the high levels of anthropogenic pollution. The ER values of studied heavy metals in this study are presented in table 4.

The ER values range for Mn (0.36–0.64), Cr (0.44–1.42), Zn (0.56–0.78), Pb (2.50–3.40), Ni (1.17–1.83), Cu (0.15–0.42) and Cd (2.67–10.59). Enrichment levels are identified using table 2 based on calculated values of ER. The ER values for all heavy metals at all sites except for Cd and Pb are less than 2 which indicates the enrichment is at mineral depletion level by the other metals. For most sites, the ER values of Cd show significant contamination (i.e 5 < ER < 20) of which, site S3 and S4 were the most significant sites with ER value of 10.59 and 10.42 respectively indicating the pollution due to anthropogenic activities. Cd has high geochemical activity in the environment and can be transported for a long distance as it is mobile and soluble in water compared to other heavy metals. The values of ER for Pb are between 2 and 5 for all sites which indicates moderate contamination. The difference in ER values for different heavy metals from different locations was probably because of variation in metal input or the variation in the removal rate of each heavy metal from the sediments [44].

The Igeo class was determined with reference to Abrahim and Parker Classification [45], reported in table 2. This index is used to assess the anthropogenic impact. The Igeo values for the studied 8 heavy metals for each sampling location are shown in table 5. The Igeo values for Cr, Cu, Fe, Mn, Ni, and Zn were less than zero for all sites which lie in class 0, indicating that the sediments of Ganga River at these sampling sites are unpolluted by these metals and the Average value of Igeo for Pb is calculated to be as 0.02 which lies in class 1 (i.e. Igeo value lies between 0 and 1) indicate unpolluted stage to moderately pollution stage. Igeo value of Cd at most sites falls in class 2 (i.e. Igeo value lies between 1 and 2) with a mean value of 1.03 which indicates moderate pollution by this metal.

### 3.3.2 *Contamination factor (CF), pollution load index (PLI) and comparison with sediment quality guidelines (SQGs)*

CF values are shown in table 6, CF values for Fe, Mn, Cr, Zn, Ni and Cu in all sediment samples were found to be < 1 with an average of 0.52, 0.28, 0.42, 0.37, 0.75 and 0.15 respectively which shows that the Ganga River sediment is not polluted by these metals. CF values for Pb at all locations lie between 1 and 3 with an

**Table 4.** Enrichment ratio (ER) for heavy metals in sediments of Ganga River.

| Site name | Code | Mn | Cr | Zn | Pb | Ni | Cu | Cd |
|---|---|---|---|---|---|---|---|---|
| Jajmau | S1 | 0.57 | 0.44 | 0.69 | 3.37 | 1.47 | 0.21 | 4.10 |
| Maharajpur | S2 | 0.64 | 1.42 | 0.78 | 2.99 | 1.48 | 0.32 | 8.27 |
| Sarsaul | S3 | 0.56 | 0.77 | 0.75 | 3.58 | 1.83 | 0.23 | 10.59 |
| Dalmau-Fatehpur Bridge | S4 | 0.65 | 0.78 | 0.56 | 2.91 | 1.20 | 0.15 | 10.42 |
| Unchahar | S5 | 0.63 | 0.82 | 0.69 | 3.05 | 1.43 | 0.34 | 7.72 |
| Nawabganj | S6 | 0.52 | 0.78 | 0.73 | 2.31 | 1.17 | 0.42 | 2.67 |
| Manikpur | S7 | 0.36 | 0.90 | 0.77 | 2.50 | 1.26 | 0.38 | 5.51 |
| Shringverpur | S8 | 0.41 | 0.54 | 0.70 | 2.84 | 1.56 | 0.25 | 8.12 |
| Phaphamau | S9 | 0.46 | 0.55 | 0.68 | 3.13 | 1.66 | 0.24 | 3.82 |
| Sangam | S10 | 0.62 | 0.92 | 0.62 | 3.40 | 1.69 | 0.16 | 4.71 |

**Table 5.** Geoaccumulation index (Igeo) of heavy metals for sediments of selected sites of River Ganga.

| Site | Fe | Mn | Cr | Zn | Pb | Ni | Cu | Cd |
|---|---|---|---|---|---|---|---|---|
| Jajmau | −1.63 | −2.46 | −2.81 | −2.18 | 0.12 | −1.08 | −3.86 | 0.40 |
| Maharajpur | −1.64 | −2.29 | −1.14 | −2.01 | −0.07 | −1.08 | −3.27 | 1.40 |
| Sarsaul | −2.00 | −2.85 | −2.38 | −2.41 | −0.16 | −1.13 | −4.12 | 1.40 |
| Dalmau-Fatehpur Bridge | −1.39 | −2.02 | −1.74 | −2.22 | 0.15 | −1.13 | −4.12 | 1.99 |
| Unchahar | −1.55 | −2.22 | −1.84 | −2.09 | 0.06 | −1.03 | −3.12 | 1.40 |
| Nawabganj | −1.01 | −1.96 | −1.38 | −1.47 | 0.20 | −0.79 | −2.27 | 0.40 |
| Manikpur | −1.06 | −2.52 | −1.21 | −1.44 | 0.26 | −0.73 | −2.44 | 1.40 |
| Shringverpur | −1.62 | −2.91 | −2.52 | −2.13 | −0.11 | −0.98 | −3.63 | 1.40 |
| Phaphamau | −1.81 | −2.92 | −2.68 | −2.36 | −0.16 | −1.08 | −3.86 | 0.13 |
| Sangam | −1.83 | −2.53 | −1.96 | −2.52 | −0.07 | −1.08 | −4.44 | 0.40 |
| Mean | −1.56 | −2.47 | −1.97 | −2.08 | 0.02 | −1.01 | −3.51 | 1.03 |

**Table 6.** Contamination factor (CF) of different metals at all sites and their pollution load index (PLI).

| Site | Site | Fe | Mn | Cr | Zn | Pb | Ni | Cu | Cd | PLI |
|---|---|---|---|---|---|---|---|---|---|---|
| Jajmau | 1 | 0.48 | 0.27 | 0.21 | 0.33 | 1.63 | 0.71 | 0.10 | 1.98 | 0.018 |
| Maharajpur | 2 | 0.48 | 0.31 | 0.68 | 0.37 | 1.43 | 0.71 | 0.16 | 3.97 | 0.187 |
| Sarsaul | 3 | 0.37 | 0.21 | 0.29 | 0.28 | 1.34 | 0.69 | 0.09 | 3.97 | 0.016 |
| Dalmau-Fatehpur Bridge | 4 | 0.57 | 0.37 | 0.45 | 0.32 | 1.66 | 0.69 | 0.09 | 5.95 | 0.143 |
| Unchahar | 5 | 0.51 | 0.32 | 0.42 | 0.35 | 1.57 | 0.73 | 0.17 | 3.97 | 0.154 |
| Nawabganj | 6 | 0.74 | 0.38 | 0.58 | 0.54 | 1.72 | 0.87 | 0.31 | 1.98 | 0.660 |
| Manikpur | 7 | 0.72 | 0.26 | 0.65 | 0.55 | 1.80 | 0.91 | 0.28 | 3.97 | 0.964 |
| Shringverpur | 8 | 0.49 | 0.20 | 0.26 | 0.34 | 1.39 | 0.76 | 0.12 | 3.97 | 0.035 |
| Phaphamau | 9 | 0.43 | 0.20 | 0.23 | 0.29 | 1.34 | 0.71 | 0.10 | 1.64 | 0.007 |
| Sangam | 10 | 0.42 | 0.26 | 0.39 | 0.26 | 1.43 | 0.71 | 0.07 | 1.98 | 0.012 |
| Average | | 0.52 | 0.28 | 0.42 | 0.37 | 1.53 | 0.75 | 0.15 | 3.34 | 0.100 |

average of 1.53 which indicates moderate contamination by Pb at all sites. The maximum CF value of for Pb was measured as 1.80 for site S7 (Manikpur) while the minimum value measured was 1.34 for site S3 (Sarsaul) and S9 (Phaphamau). The average value of CF for Cd is calculated as 3.34 which shows river sediments are contaminated by Cd in considerable amount. CF values of Cd for sites S1, S6, S9 and S9 lie between 1 and 3 which indicate moderate contamination by Cd at these sites whereas CF value for Cd at sites S2, S3, S4, S5, S7 and S8 lies between 3 and 6 which indicate a considerable amount of contamination by Cd at these sites. The maximum CF value of 5.95 for Cd was computed for site S4 while the minimum value of 1.64 was recorded at site S9. The

**Figure 10.** Untreated wastewater being dumped in Ganga River near Jajmau, U P, India.

general outcome from calculated CF values demonstrates that the river sediment is not in disturbing condition for metals Fe, Mn, Cr, Zn, Ni and Cu at present. However, it has moderate contamination by Pb and considerable contamination by Cd which may increase in the future and may be alarming to the river ecosystem with the rapidly increasing population in the river basin.

Pollution load index (PLI) indicates the overall heavy metal pollution at any site. PLI values calculated for the selected sites for the investigated heavy metals are shown in table 6. It indicates the highest pollution at Manikpur site (0.964) followed by Nawabganj (0.66), Maharajpur (0.187), Unchahar (0.154), Dalmau-Fatehpur Bridge (0.143), Shringverpur (0.035), Jajmau (0.018), Sarsaul (0.016) Sangam (0.012), whereas the lowest polluted was Phaphamau site (0.007). PLI value for site S7 (Manikpur) is close to 1 indicating close to baseline level of pollutionn and thus require quick attention to control the further deterioration of river sediment.

The reference values TEC and PEC for river sediments are shown in table 3. The concentrations of Zn and Cu were found below the reference TEC values in all sampling locations, indicating rare chances of harmful effects to the aquatic bio-system whereas the concentrations of Fe, Mn, Cr, Pb, Ni, and Cd exceeded the TEC level by 48.6, 44.7, 128, 8.93, 63, and 200% for some samples, indicating harmful effects on the aquatic plants and animals.

## 4. Conclusions

This study focused to assess the risk level, spatial distribution, and geo-accumulation of 8 heavy metals namely Cadmium (Cd), Chromium (Cr), Copper (Cu), Iron (Fe), Manganese (Mn), Nickel (Ni), Lead (Pb), and Zinc (Zn) in sediments of the Ganga River between Kanpur and Prayagraj stretch, U.P (India). The distribution of these metals in the sediments is not uniform throughout the studied stretch of the Ganga River and the change in metal concentration is due to the discharge of these metals from various anthropogenic sources like untreated wastewater from nearby industries of battery manufacturing and electroplating, pesticides, fertilizers, paint, tannery, etc (figure 10).

The concentrations of Cadmium (Cd) and lead (Pb) were found above the recommended SQG values of 0.99 and 35.8, indicating that the Ganga River is contaminated by Cd and Pb which could result in an adverse effect on the ecosystem of Ganga River. The values of Igeo and ER for the analyzed sediment samples revealed that they were moderately contaminated by Cd and Pb and unpolluted by other heavy metals. PLI values reveal that Manikpur is the most polluted site and Phaphamau is the least polluted site.

Heavy metal concentration levels and distribution were found higher at locations that are downstream of industrial and urban areas. This also shows that increase in industries and urban cities near the studied area were the likely causing metal contamination. Industrial units should ensure proper treatment of toxic effluent wastewater before disposing into the Ganga River. Government agencies like Central Pollution Control Borard (CPCB) and State Pollutions Control Board (SPCB) must ensure strict compliance with environmental standards of municipal and industrial effluents to save the Ganga River from further pollution by these metals. Remediation of technology such as In-situ physical/chemical treatments like capping, solidification/stabilization, Ex-situ physical/chemical treatments like soil washing, solidification/stabilization, and chelation, or Ex-situ thermal treatment like vitrification can be used for the majorly of polluted sites.

**Author contributions** All authors contributed to the study conception and design. Sample collection and preparation, testing and analysis were performed by Mohit Aggarwal. The first draft of the manuscript was written by Mohit Aggarwal. All authors read and approved the final manuscript.

**Funding** The authors declare that no funds, grants, or other support were received for this study.

**Declarations**

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

[1] Qadir A and Malik R N 2011 Heavy metals in eight edible fish species from two polluted tributaries (Aik and Palkhu) of the River Chenab, Pakistan. *Biol. Trace Elem. Res.* 143(3): 1524–1540

[2] Fu J, Zhao C and Luo Y 2014 Heavy metals in surface sediments of the Jialu River, China: their relations to environmental factors. *J. Hazard. Mater.* 270: 102–109

[3] Demirak A, Yilmaz F and Tuna A L 2006 Heavy metals in water, sediment and tissues of Leuciscus cephalus from a stream in southwestern Turkey. *Chemosphere.* 63: 1451–1458

[4] Kaushik A, Kansal A and Kumari S 2009 Heavy metal contamination of river Yamuna, Haryana, India: assessment by metal enrichment factor of the sediments. *J. Hazard Mater.* 164: 265–270

[5] Zheng N A, Wang Q and Liang Z 2008 Characterization of heavy metal concentrations in the sediments of three freshwater rivers in Huludao City, Northeast China. *Environ. Pollut.* 154: 135–142

[6] Iqbal J and Shah M H 2014 Occurrence, risk assessment, and source apportionment of heavy metals in surface sediments from Khanpur Lake, Pakistan. *J. Anal. Sci. Technol.* 5: 1–12

[7] Zahra A, Hashmi M Z and Malik R N 2014 Enrichment and geo-accumulation of heavy metals and risk assessment of sediments of the Kurang Nallah—feeding tributary of the Rawal Lake Reservoir, Pakistan. *Sci. Total Environ.* 470: 925–933

[8] Park J and Presley B J 1997 Trace metal contamination of sediments and organisms from the Swan Lake area of Galveston Bay. *Environ. Pollut.* 98: 209–221

[9] Xiao R, Bai J and Huang L 2013 Distribution and pollution, toxicity and risk assessment of heavy metals in sediments from urban and rural rivers of the Pearl River delta in southern China. *Ecotoxicology.* 22: 1564–1575

[10] Varol M and Sen B 2012 Assessment of nutrient and heavy metal contamination in surface water and sediments of the upper Tigris River, Turkey. *Catena* 92: 1–10

[11] Salati S and Moore F 2010 Assessment of heavy metal concentration in the Khoshk River water and sediment, Shiraz, Southwest Iran. *Environ. Monit. Assess.* 164: 677–689

[12] Dwivedi S, Mishra S and Tripathi R D 2018 Ganga water pollution: a potential health threat to inhabitants of Ganga basin. *Environ. Int.* 117: 327–338

[13] Pandey J, Pandey U and Singh A V 2014 Impact of changing atmospheric deposition chemistry on carbon and nutrient loading to Ganga River: integrating land-atmosphere-water components to uncover cross-domain carbon linkages. *Biogeochemistry* 119: 179–198

[14] Yadav A and Pandey J 2017 Contribution of point sources and non-point sources to nutrient and carbon loads and their influence on the trophic status of the Ganga River at Varanasi, India. *Environ Monit. Assess.* 189: 1–19

[15] CPCB 2013 Pollution assessment: River Ganga

[16] Trivedi R C 2010 Water quality of the Ganga River-an overview. *Aquat. Ecosyst. Heal & Manag.* 13: 347–351

[17] Pandey J, Shubhashish K and Pandey R 2010 Heavy metal contamination of Ganga river at Varanasi in relation to atmospheric deposition. *Trop. Ecol.* 51: 365–373

[18] Pandey J and Singh R 2017 Heavy metals in sediments of Ganga River: up-and downstream urban influences. *Appl. Water Sci.* 7: 1669–1678

[19] Tariq J, Ashraf M and Jaffar M 1996 Pollution status of the Indus River, Pakistan, through heavy metal and macronutrient contents of fish, sediment and water. *Water Res.* 30: 1337–1344

[20] Singh V K, Singh K and Mohan D 2005 Status of heavy metals in water and bed sediments of river Gomti—a tributary of the Ganga river, India. *Environ. Monit. Assess.* 105: 43–67

[21] Suthar S, Nema A K and Chabukdhara M 2009 Assessment of metals in water and sediments of Hindon River, India: impact of industrial and urban discharges. *J. Hazard. Mater.* 171: 1088–1095

[22] Rifaat A E 2005 Major controls of some metals distribution in sediments off the Nile Delta, Egypt. *Egyptian Journal of Aquatic Research* 31(2): 16–28

[23] Aktar W, Paramasivam M, Ganguly M, Purkait S and Sengupta D 2010 Assessment and occurrence of various heavy metals in surface water of Ganga river around Kolkata: a study for toxicity and ecological impact. *Environ. Monit. Assess.* 160: 207–213

[24] Singh M, Müller G and Singh I B 2003 Geogenic distribution and baseline concentration of heavy metals in sediments of the Ganges River, India. *J. Geochemical Explor.* 80: 1–17

[25] Singh H, Yadav S, Singh B K, Dubey B, Tripathi K, Srivastava V and Shukla D N 2013 Assessment of geochemical environment from study of river sediments in the middle stretch of River Ganga at Ghazipur, Buxar and Ballia area. *Proceedings of the National Academy of Sciences, India Section B: Biological Sciences.* 83(3): 371–384

[26] Ansari A A, Singh I B and Tobschall H J 2000 Importance of geomorphology and sedimentation processes for metal dispersion in sediments and soils of the Ganga Plain: identification of geochemical domains. *Chem. Geol.* 162: 245–266

[27] USEPA 2001 Methods for collection, storage and manipulation of sediments for chemical and toxicological analyses: technical manual. EPA 823-B-01-002 US

[28] Zhang L, Ye X, Feng H, Jing Y, Ouyang T, Yu X, Liang R, Gao C and Chen W 2007 Heavy metal contamination in western Xiamen Bay sediments and its vicinity, China. *Mar. Pollut. Bull.* 54: 974–982

[29] Hakanson L 1980 An ecological risk index for aquatic pollution control. A sedimentological approach. *Water Res.* 14: 975–1001

[30] Muller G 1969 Index of geoaccumulation in sediments of the Rhine River. *GeoJournal.* 2: 108–118

[31] Usero J, Gonzalez-Regalado E and Gracia I 1997 Trace metals in the bivalve molluscs Ruditapes decussatus and Ruditapes philippinarum from the Atlantic Coast of Southern Spain. *Environ. Int.* 23: 291–298

[32] MacDonald D D, Ingersoll C G and Berger T A 2000 Development and evaluation of consensus-based sediment quality guidelines for freshwater ecosystems. *Arch. Environ. Contam. Toxicol.* 39: 20–31

[33] Wenning R J 2005 Use of sediment quality guidelines and related tools for the assessment of contaminated sediments. *SETAC*

[34] Subramanian V, Van't Dack L and Van Grieken R 1985 Chemical composition of river sediments from the Indian sub-continent. *Chem. Geol.* 48: 271–279

[35] Martin J M and Meybeck M 1979 Elemental mass-balance of material carried by major world rivers. *Mar. Chem.* 7: 173–206

[36] Abadin H, Ashizawa A, Llados F and Stevens Y W 2007 Toxicological profile for lead. *US Dep. Heal Hum. Serv*

[37] WHO 2010 Exposure to cadmium: a major public health concern. *Prev. Dis. through Heal Environ.* 3–6

[38] Faroon O, Ashizawa A, Wright S, Tucker P, Jenkins K, Ingerman L and Rudisill C 2013 Toxicological profile for cadmium. *US Dep. Heal Hum. Serv*. US

[39] Dorsey A and Ingerman L 2004 Toxicological profile for copper. *US Dep Heal Hum. Serv*. US

[40] Williams M, Todd G D, Roney N, Crawford J, Coles C, McClure P R, Garey J D, Zaccaria K and Citra M 2013 Toxicological profile for manganese. *US Dep. Heal Hum. Serv*. US

[41] Abadin H, Fay M, Ingerman L, Tencza B, Yu D and Wilbur S B 2012 Toxicological profile for chromium. Agency for toxic substances and disease registry. Public Heal Serv *US Dep. Heal Hum. Serv*. US

[42] Feng H, Han X, Zhang W and Yu L 2004 A preliminary study of heavy metal contamination in Yangtze River intertidal zone due to urbanization. *Mar. Pollut. Bull.* 49: 910–915

[43] Chen C W, Kao C M, Chen C F and Dong C D 2007 Distribution and accumulation of heavy metals in the sediments of Kaohsiung Harbor, Taiwan. *Chemosphere* 66: 1431–1440

[44] Ghrefat H A, Abu-Rukah Y and Rosen M A 2011 Application of geoaccumulation index and enrichment factor for assessing metal contamination in the sediments of Kafrain Dam, Jordan. *Environ. Monit. Assess.* 178: 95–109

[45] Abrahim G M S and Parker R J 2008 Assessment of heavy metal enrichment factors and the degree of contamination in marine sediments from Tamaki Estuary, Auckland, New Zealand. *Envion. Monit. Assess.* 136: 227–238

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-13900-1

# HISNet: a Human Image Segmentation Network aiding bokeh effect generation

Shaurya Gupta[1] · Dinesh Kumar Vishwakarma[1]

Received: 28 November 2021 / Revised: 7 June 2022 / Accepted: 12 September 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

**Abstract**
The bokeh effect in photography has gained unquestionable popularity since improvements in smartphone cameras, for this effect brings out the attention of the image onto the subject and enhances the overall quality of the photo. Generally, these effects are applicable via dual-lens cameras for auto-focusing onto the subject. However, smartphones with a single lens rely on software to generate such an effect. This paper proposes a deep learning pipeline to generate depth-aware segmentation maps in human images via segmentation and depth estimation networks. The given paper provides a concatenations-based decoder for segmentation applying and experimenting with features learned through state-of-the-art encoder architectures, further we form an encoding concatenation between two prominent encoders to provide an ensemble model for learning segments. Adding to the effect we use a prominent depth estimation architecture and combine it with our segmentation results to generate dept-aware segmentation maps for achieving photos with more focus on human subjects, where the out-of-focus regions appear to be blurred out. The methodology produces compelling bokeh effects, comparable with shots taken via a dual-lens mobile camera or DSLR. During the experimentations of human segmentation, some benchmark results are reported with our best-considered model. Training on Supervisely Persons dataset achieved an IOU score of 95.88%, whereas training the same network on the EG1800 dataset achieved a state-of-the-art IOU of 96.89%. The final segmentation model thus provided some very closely accurate segmentation maps suitable for our task.

**Keywords** Bokeh/depth effect · Depth estimation · Dice coefficient · Human segmentation · IOU score · Transfer learning

✉ Dinesh Kumar Vishwakarma
dinesh@dtu.ac.in; dvishwakarma@gmail.com

Shaurya Gupta
shaurya62000@gmail.com

[1] Biometric Research Laboratory, Department of Information Technology, Delhi Technological University, Delhi, India

Published online: 19 September 2022

## 1 Introduction

Depth Effect, often referred to as Bokeh Effect (Japanese for blur), allows the subject in images to appear to be treated in higher focus than the relative background, as evident from Fig. 1. Generally, these effects are captured via a DSLR (Digital Single-Lens Reflex camera) which utilizes substantial focal lengths to help attain the desired depth effect in images [6]. This effect can also be produced with a dual-lens camera system, allowing many smartphones to replicate these results, where two images are captured via the dual lenses for computation of depth maps to create the bokeh effect, implying the need for extra hardware for generating such pictures thus adding to the camera costs.

Why we need bokeh effect generation? according to studies, about 85% of total photos taken around the world are taken via a smartphone, with the number still increasing with each passing year [33]. Thus, there is a need to make more cost-effective devices that can rely on software to generate such appealing photos, for computational image enhancement is on the rise due to the progress in deep learning research. Bokeh is the prominent effect in such enhancement because it appears just like a DSLR image in many cases. Also, deep networks are widely used for classification and prediction purposes [8, 38].

Using machine learning based techniques, a captured picture with human subjects can be post-processed to generate a bokeh effect in the background using various filters and stylization as portrayed by Shen et al. [31], where the subject area is under the higher focus than the relative environment. Though results are seen through matting technique, they are restricted for single human subject and only in a portrait setting. Many techniques even lack the fine boundaries for human subjects and don't include the near vicinity region to show the feel of depth in image. This paper demonstrates a method for depth effect generation by introducing a semantic human image segmentation network (covering all close-by subjects in a frame) and leveraging depth estimation (mainly for rear camera shots) to generate a depth-aware segmentation map using which we can blur the out-of-focus regions and replicate the bokeh effect. The given paper, using some state-of-the-art encoders as backbone and a shallow decoder form a U-net-like structure, introducing several concatenation layers in between, for our segmentation task. The aim is to use pre-learned filters through the encoder and improve them by concatenating them to a shallow decoder, easing the learning process and form some very fine segments. We refine the segmentation maps using depth estimation networks to include the background regions close to the human subject proximity, allowing us to gain a DSLR-like bokeh effect in our images. Generally, selfie shots do not require us to include a background that much, and thus in such scenarios, our segmentation approach is alone capable of generating the desired effects. In this paper, we only consider human subjects because most

**Fig. 1** Depth focused image in right and the normal image on left, for the two samples. The Depth effect can be seen to make subject appear a little more in focus compared to the original images on left (These effects have been rendered by our proposed methodology)

of the pictures taken by a smartphone capture human images via selfies or rear camera shots, thus making human subjects our priority. The main contribution of our work presented is as under:

- We propose a new U-net-like segmentation architecture called HISNet, where we apply concatenation connections between encoder and decoder layers for accurate region boundaries predictions and quick convergence of the model on the considered loss function.
- The employed encoder network is selected by comparing three pre-trained state-of-the-art architectures. We also introduce an ensemble encoder architecture that concatenates the encoded input features to observe the applicability of such encodings in segmentation networks.
- A thorough pipeline is provided, describing the techniques followed to generate the appealing bokeh/ depth effect in human subjected images.
- The proposed segmentation network achieves the state-of-the-art result on one of the prominent human segmentation datasets, with very-fine segmentation maps.

Through this paper, we follow a structure as follows: in Section 2, we give a brief overview of the current related methods for image segmentation in human images and depth estimation in general; Section 3 provides the methodology of image segmentation technique used in our work with a brief description of each applied encoder; Section 4 provides details about the pre-trained depth estimation network used and reason for its selection; Section 5 provides the proposed technique followed for depth-effect rendering using the depth and segmentation maps; then Sections 6 and 7 provide information on the dataset, training, experiments, and results; Finally, we conclude our the paper with future work and discussions in Section 8.

## 2 Related works

Since the work in this paper applies image segmentation and depth estimation, a brief overview of the dominant methodologies in the respective titles is provided, with some prominent work related to human image segmentation mentioned in the segmentation section. Though we do not introduce a novel depth estimation network, we review its literature to consider the best architecture for usage in our bokeh effect generation pipeline.

### 2.1 Image segmentation

Image segmentation is considered a classification problem at a pixel level. Over the years' segmentation has evolved from a general computer vision problem to a deep learning one, where initially techniques like global and local thresholding were being used to generate the segmentation maps, which later advanced to using machine learning techniques like k-means clustering to get cluster regions of those segments based on the color and other properties of an image. However, the most closely accurate segmentations are currently being built using deep learning, first done using FCNs (Fully Convolutional Networks) introduced by Shelhamer et al. [30] and improved structurally since then.

Deep Learning has provided many promising results in semantic segmentation [22, 41] via Convolutional Neural Networks. Some general architectures prominent in the field include the likes of SegNet [2], U-net [36], E-net [26], DeepLab [3], etc. Recently, with the advent of

transformers and developments in attention mechanism for computer vision[10] there has been rapid development to incorporate these techniques in segmentation as well. Attention U-net [24], for instance, considers the use of an attention gate to focus the model target on different shapes and structure better. Similarly, DA-Net [9] applies self-attention over its channel and spatial dimensions to improve its segmentation results. Some other works based on these methods includes the likes of Trans4Trans [45], SegFormer [39], TransUnet [5], etc. Further many methods are developing towards weakly supervised segmentation through aggregation of cues [37] for their effectiveness. However, because our technique in this paper is more focused on deep learning methods for human image segmentation, we view a few significant developments inclined towards this specific problem statement.

As mentioned in the introduction, the work contributed by Shen et al. [31] for image stylization is prominent in human segmentation. It provides a Portrait FCN network to detect the human masks learned by training on a dataset, also contributed by them called EG1800, containing portrait images. Even though this technique achieves very smooth boundaries on human images, their predictions were limited to closely bound single human images, i.e., selfie-like image shots. The contributions by Xu et al. [40] on the other hand, have a wide range of human poses in their dataset and use a portrait segmentation network to generate a coarse output and improve on that via a Spatial Propagation Network (SPN) that produces some accurate segmentation maps. Their method then applies depth estimation for rendering the bokeh and applies depth estimation network to produce a coarse output and again use SPN to improve the prediction and finally applies those outputs to generate the final depth effect in the image. Even though the results seem suitable for bokeh their technique had a few depth overlays over the human subject which are refined in our technique by focusing mainly over the human subject. Many of these techniques even require much processing before generating the final result. Zhang et al. [44] also proposed models based on Portrait FCNs and contributed a Human Video Dataset with a total of 1000 frames, but they created the dataset with only four distinct single-person videos that make the model prone to the prediction of two or more human subjects.

Though many of the networks mentioned above had benchmark segmentation metric scores in their respective data, they lacked segmentation for multiple subjects and thus were limited to segment only a single human subject. Our considered dataset has a broader range of human segment maps; we even add multiple concatenations of convolutional filters from the encoder to decoder layers to help get more accurate boundaries forming a U-net-like architecture but experimenting on different encoder architectures. Also, encoder networks are pre-trained; thus, fine-tuning on the considered dataset helps our model converge in merely 26 epochs of training compared to training in 100s of epochs in the works as mentioned earlier. We also study the architectural differences contributing to the outputs obtained and compare visual results to infer the effect of these different convolutional methods.

### 2.2 Depth estimation

Compared to image segmentation, Depth Estimation is a task of pixel-wise depth calculation learned by training from the depth maps created using Hardware devices like Microsoft Kinect. Depth Maps provides a 3D perspective to the image in the study. Again, like depth effect generation in photos taken via single-lens smartphones, depth estimation reduces the hardware reliance to estimate depth on a scenery. Thus, depth estimation is a very significant problem in deep learning.

Some of the early and prominent works currently in this problem include works like MonoDepth [11], an unsupervised learning technique for depth prediction contributed by Godard et al., correspondingly for supervised learning; there are many architectures like a ResNet [12] based model by Laina et al. [18] utilizing the Residual blocks for learning depth maps. Lee et al. [19] estimated depths using a technique of relative depth maps while using a custom architecture. Another network called DenseDepth by Alhashim et al. [1] utilized DenseNets [15] for depth prediction is one of the state-of-the-art ones itself. Furthermore, as mentioned in human image segmentation, Xu et al. [40] also shows the use of CNN (Convolutional Neural Network) for coarse depth generation and improving the prediction with SPN, as applied in its segmentation task. For our depth estimation purpose, we use the pre-trained MiDaS [28] model to get the depth maps since the visually analyzed results were suitable for our task to generate depth-aware segmentation maps discussed in the following sections.

## 3 Image segmentation technique

This section provides the overview of our employed technique, introducing HISNet (Human Image Segmentation Network) used to segment human subjects in a focusable range of camera images. We thoroughly explain the use of deep learning-based encoders and decoder structure applied to generate the final segmentation map. The different encoder structures are compared with each other for the final selection of the segmentation model.

Prior to deep learning, generally machine learning algorithms were used for segmenting regions based on different feature selection. As mentioned in the Section 2, k-means algorithm was used to segment objects based on the color attributes present in the image with the drawback of selection of number of clusters to segment regions. Another technique called watershed algorithm used thresholding parameter with contour regions to identify the region for segmentation, though with complex backgrounds the technique tended to fail to segregate regions and foreground. Chan-Vese Algorithm tries to determine segment maps with a more different approach by forming it as an energy minimization problem for the regions present in the image to segment them, which despite being much more effective than k-means and watershed still lacks the ability to define the segment boundaries to fine-precision. Thus, we target the task through deep learning-based methods to learn the underlying mapping through a vast dataset.

For the generation of accurate segmentation maps, we utilize fully convolutional encoder-decoder architecture, wherein we use image-net [7] pre-trained weights for the encoder networks to exploit the benefits of transfer learning [4] and ease the understanding by our model for quick convergence. Each architecture discussed applies the same decoder network to compare performance with each of the considered encoders. These architectures appear like SegNet and FCN, Fig. 2, but with explicit concatenation connections helping quick learning and accurate predictions. The networks are trained on Supervisely Persons Dataset [35]; we provide more details on the dataset in the Section 6.

In this paper, the considered three pre-trained encoder architectures are namely, VGG16 [32], MobileNetv2 [29], and Densenet121 [15]. We compare these three networks' performance with the same decoder network by testing on the supervisely test-set discussed in the Section 6. We name our network as HISNet, i.e., Human Image Segmentation Network. We chose image-net pre-trained weights for our encoders because the wideness of classes in the

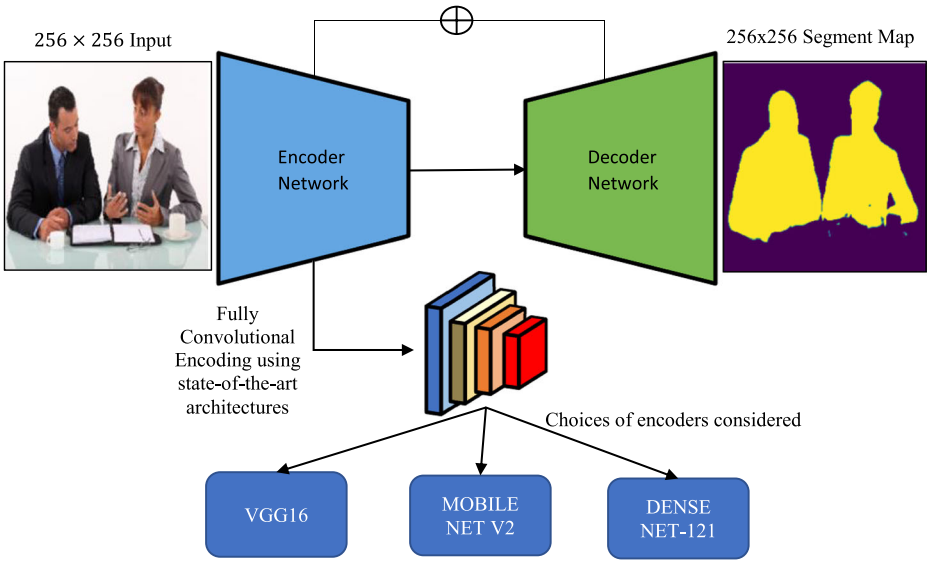



**Fig. 2** Basic working architecture for Fully Convolutional Neural Network used for our Human Segmentation task. The ⊕ symbol here denotes the concatenation between the encoder and decoder network. The segmentation map covers the part where the human body is present

dataset makes fine-tuning on any other dataset relatively easier. Apart from these three encoders, we also use an ensemble of two encoders to generate an encoded feature map produced by concatenating those considered encoders' output, The considered ensemble of encoders is also pre-trained on image-net.

For each of the considered encoders, the feature maps are processed up to a dimensional equivalence of $16 \times 16$ resolution, which acts as input to the decoder for extracting the desired feature maps to generate pixel-wise predictions. The architecture applies 'same' padding throughout their convolutions, thus preserving the dimensions while processing. Therefore, the dimensional change occurs only at the max-pooling layers with a stride of 2.

The features processed by the encoder part are concatenated to the decoder to process the feature maps more easily since these concatenated connections help in quick feature transfer from the encoder to the decoder network. There are two concatenations for each dimension, ranging from $16 \times 16$ to $64 \times 64$, followed by only one concatenation for the $128 \times 128$ dimension.

### 3.1 VGG-16 encoder

Since the success of AlexNet [17] in 2012, with more layered depth, compared to LeNet, it was evident that more depth contributed to better feature learning. Thus, VGG Networks were introduced by Simonyan et al. [32] in 2014, which is more than twice the size of AlexNet and applies $3 \times 3$ receptive fields with stride one as compared to $11 \times 11$ with stride 4 in AlexNet, reducing the number of parameters as well as increasing the performance. The two significant Networks introduced were VGG-16 and VGG-19, with the respective number of layers. In this paper, we use transfer learning on VGG-16 architecture for the encoder. Since we want to use as few parameters as possible, we use only the first 12 convolutional layers to reach the dimensional size of $16 \times 16$. Among all the convolutional layers, ReLU activation is used. For

each dimension of the image processed by the convolution layer, a few are saved for concatenation in the decoder part of the network to help improve the segmentation maps generated.

### 3.2 Mobile Net v2 encoder

The mobile net architecture proposed by Howard et al. [13], was introduced for good inferencing on mobile devices. The architecture used by MobileNetV2 [29] is based on an inverted residual structure where the input and output of the residual blocks are thin bottleneck layers. This architecture utilizes lightweight depthwise convolutions to filter features in the intermediate layers, as shown in Fig. 3a. Depthwise Separable Convolutions, introduced initially in MobileNetV1 [13], dramatically reduced the complexity and size of the network, which is suitable to low computational power devices. With MobileNetV2 as the encoder architecture for feature extraction, segmentation becomes feasible since the number of parameters used is significantly less than VGG-16. Even with fewer parameters, MobileNetV2 has a depth of 88 layers making network learning sufficient. Again, utilizing Transfer learning provides faster training and better convergence. Though the results of this pre-trained network are on par with VGG-16 on the image-net dataset, the results on the segmentation task here are observed to be far more superior to the VGG-HISNet.

### 3.3 DenseNet 121 encoder

As the number of layers in neural networks increased, it was found that there was a decline in model learning instead of getting better results. It was due to the vanishing gradients problem, i.e., As we progressed, the computed values in each subsequent layer decreased significantly, making the model difficult to learn further in deeper layers. To tackle this problem, a common practice was to use residual blocks, introduced by He et al. in ResNets [12]; these helped the model retain the previous input values and thus tackle the vanishing gradients problem by not letting the network die out. DenseNets [15], by Huang et al., takes this concept to another level by considering dense blocks wherein instead of doing element-wise addition, we perform

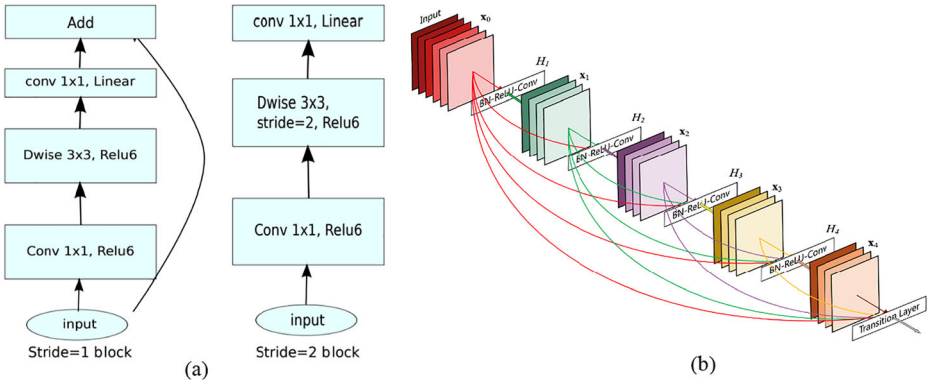



**Fig. 3** **a** Depthwise Convolutional Blocks utilized by the MobileNetv2 encoder [44] and **b** Example of Dense Block used in DenseNet121 [45], We can see how each following layer get concatenated inputs from previous layers, helping tackle the vanishing gradient problem

channel-wise concatenation for each layer from the preceding layers, as visible in Fig. 3b. Thus, each layer receives a 'collective knowledge'; this allows the model network to be thinner and achieve higher accuracy with lower parameters. Each layer in a DenseNet comprises Batch Normalization, ReLU, and convolution layers. With multiple Dense Blocks, a pre-trained model of depth 121 layers was established and found to have higher accuracy than the other two models used as our encoders. The results are discussed, shown in the following sections. Thus, we use this method and compare it with the other two models and find inference on the best model for HISNet. Again, each model only runs up to processed 16-dimensional size and a few convolutions for further processing learning.

### 3.4 Ensemble encoder

An Ensemble model aims to combine the results of different models, helping improve the overall performance. Generally, in an ensemble architecture, the average of several models is considered to find the most probable result among the number of results presented by each model. For our ensemble, we concatenate the encoded features maps of each encoder to form a final encoded vector, to be decoded by our decoder (discussed in the next section). To construct our ensemble model, we use DenseNet121 and Mobilenetv2 encoders previously reviewed in Sections 3.2 and 3.3. We did not use the VGG16 encoder, for the number of parameters was very high in the combined architecture, with more than 50% of parameters contributed by VGG.

Our methodology, in particular, utilizes concatenation connection from an encoder to a decoder providing quick convergence over the loss by transfer of previously learned features. In a single bottleneck encoder, we select some layers to be concatenated later to the same dimensional decoder layers. However, for the considered ensemble bottleneck in Fig. 4, we concatenate both encoders' selected layers to the corresponding decoder layer, as followed for a single bottleneck. Thus, our decoder would get two concatenation connections instead of one, as shown in Fig. 4. The benefits of this method can be observed quantitatively in Section 6 and by visible results in Fig. 6.

### 3.5 Decoder network

After encoding the processed image, the remaining part of the image segmentation network is to perform the decoding of the encoder's output tensor. The proposed decoder architecture part is commonly used by all of the four encoder networks, with a different number of filters being concatenated from each of the considered encoder architectures.

Each of the up-sampling layers in the decoder uses bilinear interpolation [16], and each up-sampling layer is followed by two convolutions, with the use of batch normalization in between to ease the learning process. All the layers utilize leaky ReLU activation except for the final layer being Sigmoid activated, for it is a binary segmentation problem. The total depth of the decoder network sums up to just 14 layers (excluding batch normalization) which is two more layers than the used VGG encoder and far smaller than the mobile and dense nets. The decoder Network is shown in Fig. 5. Again, we call our model HIS (Human Image Segmentation) considering the task being performed. We can observe some visual results of segmentation by each of the encoder network, as shown in Fig. 6. Visually, DenseNet and the Ensemble techniques shows us the most accurate and stable results.

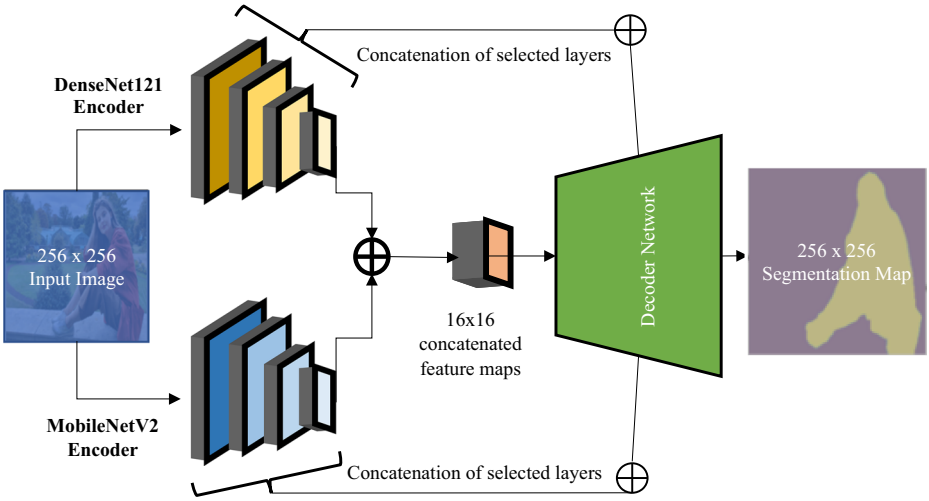



**Fig. 4** Ensemble architecture using concatenation of encoded features from two bottleneck encoders namely, Densenet121 and MobileNetv2. The ⊕ symbol represents the concatenation connections

### 3.6 Loss function

In our segmentation problem, there is only a single object to be segmented since we are interested only in human subjects for segmentation. Therefore, we apply sigmoid activation in the final layer of the decoder network shown in Fig. 5. Thus, to account for the binary problem,

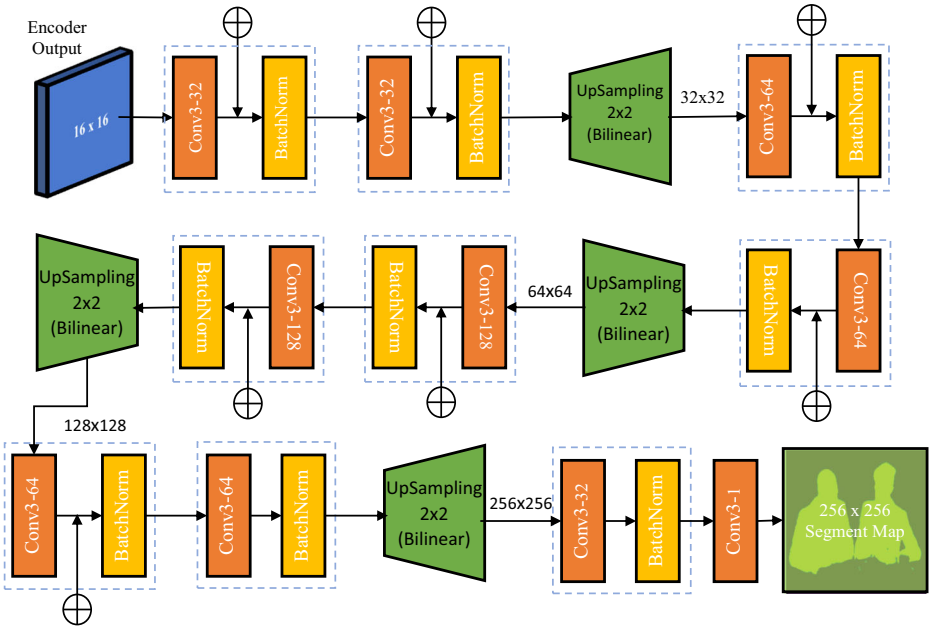



**Fig. 5** Decoder Network used with each by mentioned encoding method. The ⊕ symbol denotes the concatenation connection from encoder layers with appropriate dimensions. Here, Conv3-32 means kernel size of 3 with 32 filters

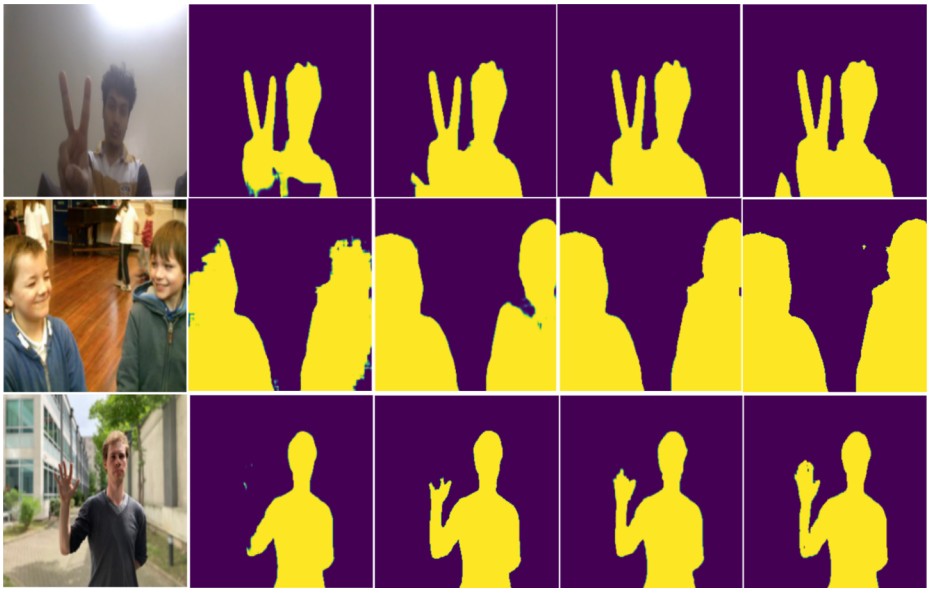



**Fig. 6** Results of Segmentation between VGG16, MobileNetV2 DenseNet121 and Ensemble encoding methods. Clearly, Ensemble and DenseNet are the networks architectures to be considered best working on unseen images. First image is a shot taken from webcam and other two images are referred from google images

we use Binary Cross-Entropy (BCE) loss for image segmentation in this case, as signified by Eq. (1).

$$L_{BCE}(y,\widehat{y}) = -(ylog(\widehat{y}) + (1-y)log(1-\widehat{y})) \tag{1}$$

Where, $\widehat{y}$ is used for each pixel value predicted in the spatial segmentation map.

Though BCE is a widely used loss function for both segmentation and classification, it is based on data distribution. In tasks like segmentation, our greater target is to get accurate regional boundaries. Therefore, to improve the results, we fine-tune our models on Dice Loss, a region-based loss that improves the near boundary segmentation and leads to more exemplary accurate segments.

Dice coefficient or F-1 score, which is similar to the IOU (Intersection over Union) score, is a commonly used metric for evaluating image segmentation maps. It can be seen that dice coefficient like IOU lies between 0 and 1, with more value indicating better overlap with the ground truth. If we subtract the dice coefficient from one, we form the loss function known as Dice Loss [34], referred to as the region-based loss as shown in Eq. (2).

$$DL(y,\widehat{y}) = 1 - \frac{2|y \cap \widehat{y}|}{|y| + |\widehat{y}|} \tag{2}$$

Again, $\widehat{y}$ here is the predicted segment map, and y is the ground truth map. If the ground truth and predicted masks do not have any region to be segmented, then the expression leads to an undefined 0/0 case; thus, a small value $\epsilon$, of the order 1e-6, is added to both numerator and denominator to avoid such scenarios.

## 4 Depth estimation network

MiDaS v2 large (Ranftl et al., [28]) is the considered model for our Depth estimation task in this work, for it provided visually and by estimated values some great benchmark results. MiDaS made use of multiple datasets for training and trained using zero-shot cross-dataset transfer methodology, i.e., the model trained on a considered dataset and its performance tested on an unseen dataset that is not used during training. MiDaS v1 utilized five datasets which were increased to ten by the authors recently in their updated version. Due to the greater size of data used by MiDaS, we considered it for depth estimation since training on more data helps a model to generalize better. The depth estimation results on the latest version provide excellent results on unseen data and estimate on human images correctly to a great extent. The MiDaS method, like ours, also applied encoder-decoder architecture for depth estimation with the encoder being pre-trained on the image-net dataset. The proposed work also experimented with state-of-the-art loss functions and studied the relative performance to provide the best model.

The depth maps are used to include near vicinity background with the human subjects (discussed briefly in Section 5). For selfie image shots, i.e., images taken via the front camera, our HISnet is strong enough to predict the human segmentation map, and since the human subject is very close in the camera shot, there is generally no need for the inclusion of near background regions. However, we have many instances where the close background, when included, gives a better bokeh effect for images taken through the primary rear camera. One question arises, why not just use the depth estimation model for getting the depth-aware segment map? This is because depth estimation models provide a higher value for objects close to the camera; thus, for close objects, it becomes easy to segment them, but for an object far away, it becomes a difficult task to find the appropriate threshold and generate a binarized segment map containing all the human subjects. We can find all the major problems associated in Fig. 7, where we take the minimum of mean and half-range value as the threshold to binarize and generate the depth-aware segment maps. The first image excludes the baby head considered by the segmentation approach using HISnet, shown in the Pipeline of our method in Fig. 8. For the second image, we do not require a depth-aware map, but as visible from the depth map, we cannot segregate the human subject as opposed to the segmentation done in Fig. 6. Finally, in the last image, the binarized depth map, even with the minimum of the two threshold values, cannot segregate the human subjects in the background, thus showing the need for our segmentation architecture to improve the final results and include the human subjects. We also observe how the close vicinity background gets added in the first image, thus making results appear as good as a dual-lens camera, shown in Fig. 8. We also observe that the hat which HISNet did not consider is covered by the depth map, thus helping preserve essential features.

## 5 Depth effect rendering

For an image captured through a smartphone, the foreground is almost always farther from the subject person. For Depth Effect Realization, we blur out the out-of-focus regions to make the subject appear in focus. These results are obtained by following the pipeline shown in Fig. 8. However, for a selfie image, where the human is very close to the camera, this pipeline can be avoided, and we may use HISNet to get the non-segmented region blurred out.

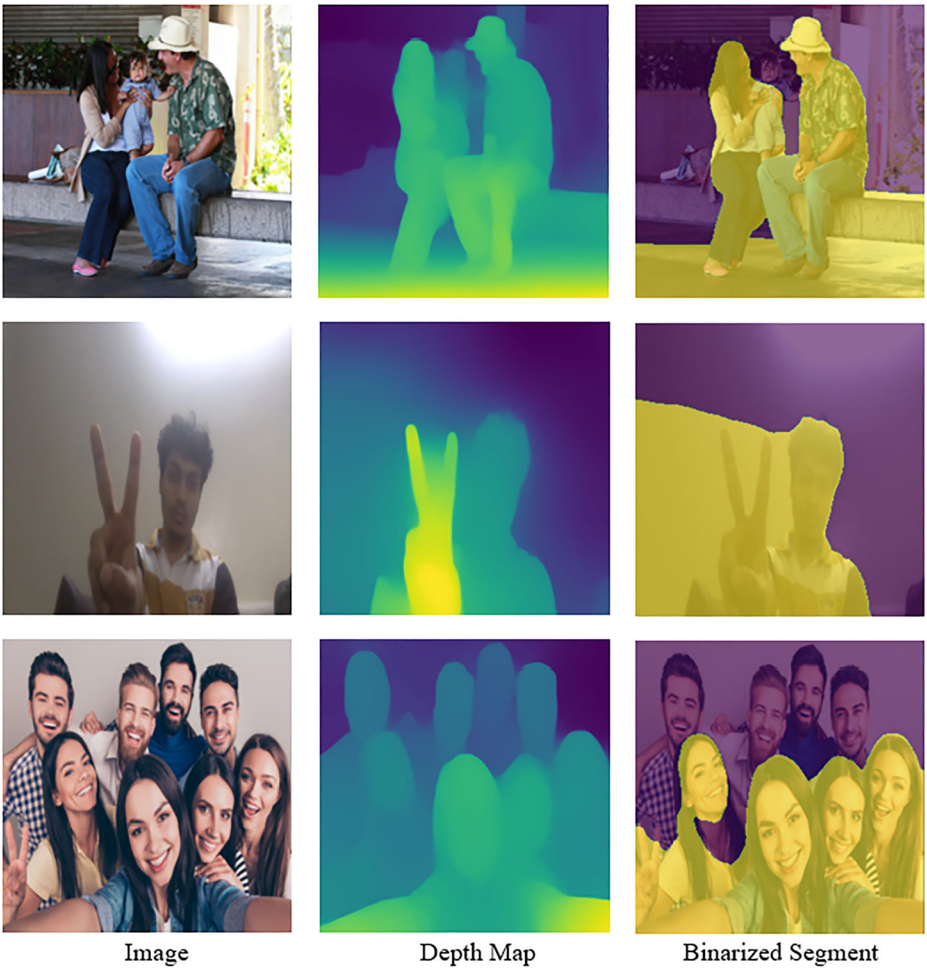



**Fig. 7** Visualization of Depth Maps produced by MiDaS. The binary segment maps are produced by taking the minimum of mean and half-range of the depth map values. The yellow color in Binarized segment corresponds to the segmented region

For generation of blur, we use the Gaussian Blur technique. Where a two-dimensional Gaussian function is utilized, as shown in Eq. (3), to form an $n \times n$ dimensional Gaussian Matrix; where n is odd.

$$f(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3}$$

For the considered $n \times n$ Gaussian Matrix, the central coordinate is assumed as the origin (0,0). Now, for our task, we apply a filter window of $9 \times 9$ dimension; the larger the dimension more the blur. The Gaussian matrix is then calculated using the formula given in Eq. (3) with the considered filter size. Further, the matrix values are averaged out to produce the sum of elements in the matrix as one. The central pixel would have the maximum value since a Gaussian is maximum at the center as observed by the exponential term associated in Eq. (3)

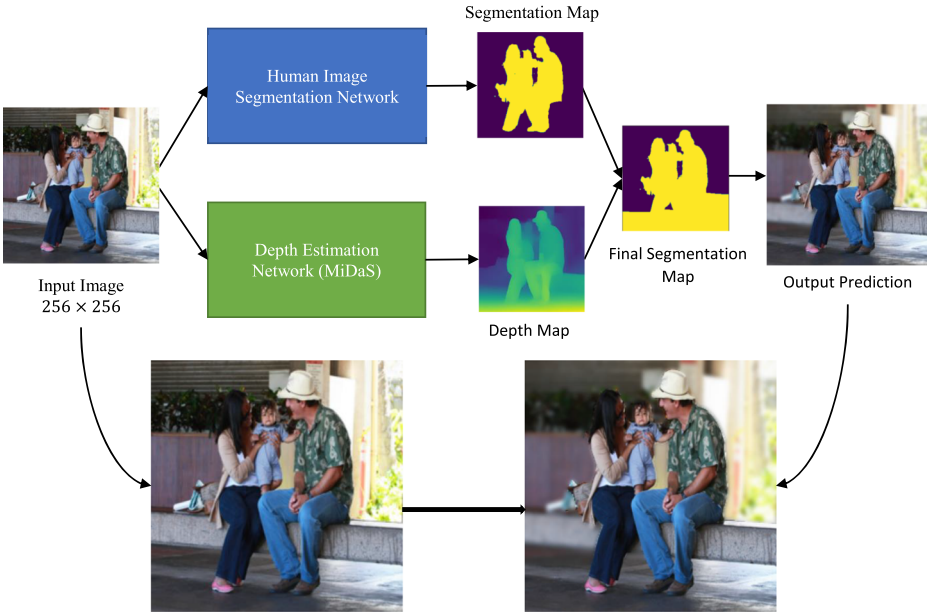



**Fig. 8** Pipeline for Depth Effect generation. We can observe that the baby head is almost not observable in depth map, whereas the considered HISNet gives full human segmentation. Combination of these two results provide us with the desired depth-aware segment map, to blur the out of focus regions

which has a maximum value of 1 at (0,0). Thus, as we move away from the central pixel, the values start to fade away, i.e., decrease in value but never reach zero.

This $9 \times 9$ Gaussian filter then convolves the image area, and the respective convolution takes place over the image to generate a blurred-out picture. This technique is considered over the regular averaging operation of pixels because this process preserves the visible boundaries better in comparison. We can then map the blurred image over the depth-aware segmentation map to generate the bokeh effect in our images.

To generate the final segmentation map containing the foreground regions, we consider half the range i.e., half the difference of maximum and minimum value of the depth map as a threshold parameter for binarizing it, since it maintained most of the near boundary region compared to average depth map value. The threshold parameter calculated for each depth map is shown in Eq. (4).

$$\theta = \frac{\max\left(\text{depth}_{\text{map}}\right) - \min\left(\text{depth}_{\text{map}}\right)}{2} \tag{4}$$

After binarizing the depth map using the threshold in Eq. (4), we simply perform the OR operation between the binarized depth map and the segmentation map from HISNet to generate the region of interest where we need to avoid the Gaussian Blur.

Some rendered Results obtained on images are given in Fig. 9. The first and third images benefitted from the depth map, for it helped include the near vicinity background, thus improving the resulting image and appearing more like a DSLR one. The second image, in comparison, does not benefit that much from the depth network since the significant background region lies very far in comparison. Images with the human subject at a farther distance benefit much more in comparison. Thus, the segmentation network and depth estimation

**Fig. 9** Depth Effect Rendering, by using the combination of Depth and Segment maps to produce depth-aware segmentation and blur the out-of-focus regions to produce the desired bokeh effect. The first two images are taken via smartphone cameras. The third image is randomly selected from internet

network go hand in hand to help each other, to make the human subject and any other object in the near vicinity appear in focus.

## 6 Experiments

In order to test the performance of the built approach, we discuss the Experimentations applied for our HISNet and the dataset utilized in training. The programming and implementation were done using Python with Tensorflow and Pytorch as the deep learning frameworks with segmentation architectures built using Tensorflow and depth estimation used available in Pytorch. All the models were trained on a system configuration comprising 2 CPUs @ 2.2Ghz, 14GB RAM, and a NVIDIA Tesla K80 GPU card with about 12 GB of Memory containing 2500 CUDA cores.

### 6.1 Dataset

As mentioned in the Section 1, we make use of the Supervisely Persons Dataset [35]; the selection of this dataset is very crucial, for the binary mask images included in this considered dataset are far more fine-grained compared to most of the benchmark segmentation data sets

like MS-COCO [20]. Additionally, the regions covered by humans in this dataset have an extensive range of coverage compared to the selfie image dataset, EG1800, given by Shen et al. [31]. A small sample of segmentation maps in the dataset can be observed in Fig. 10.

The range of human coverage in the considered dataset did pose issues based on our specific task. In some images, the presence of some very small-sized human subjects meant small area coverage, which made the architecture prone to predict those subjects that were irrelevant with our bokeh effect task since depth effect appeals when the subject is in near vicinity to the camera. For Example, in Fig. 9, the first image had irrelevant small-sized human subjects in the background, which must be excluded from being segmented. Since the considered dataset had a higher percentage of data with the human subjects in close range to the camera, we removed the human area coverage of less than 1.5%; Since below this, the images infer almost no subject for feasible focus and appealing effects. Therefore, this task reduces our training data by almost 170 images. The coverage of the human subject in the dataset can be inferred from Table 1.

The original data, the Supervisely Persons dataset [35], had approximately 6 K images with 7 K smooth and accurate annotations for human subjects, but after removing all the redundant far away subject images from the dataset, we were left with almost 5.6 K images. The dataset is then split using hold-out validation into an $85 - 15\%$ ratio for the training and testing purpose (with 4802 and 851 images, respectively). The images in the dataset are of variable sizes therefore, the images were processed and resized to be fixed at a size of $256 \times 256$ pixels for our model inference.

### 6.2 Training

All the four encoder architectures are joined with the decoder in Fig. 5, with respective concatenations. Each model is then trained on the input image of resolution $256 \times 256$ with a batch size of 32 samples. The training is done in two phases. In phase-I, the models are trained for 17 epochs on Binary Cross-Entropy Loss as shown in Fig. 11, left column. In phase-II, the models are fine-tuned on Dice Loss for only 9 epochs shown in Fig. 11, right column. Application of transfer learning can be observed to have helped the models quickly converge within such a small number of epochs.

RMS-prop optimizer was used in phase-I of training with a learning rate of 1e-3; the significant convergence over the loss values occurs in this phase. Finally, Adam optimizer is used for fine-tuning in phase-II with a decreased learning rate of 1e-4. It was evident from the graphs that pre-trained encoders help the models converge quickly. MobileNetV2 and DenseNet121 are on par with each other. The Ensemble architecture is slightly better than the other two architectures. VGG-16 under similar training conditions does not provide excellent results, even though it has more than double the number of parameters compared

**Fig. 10** A few samples of segmentation from the Supervisely Persons Dataset [37], where yellow region represents the segmentations and purple tint shows the foreground area

**Table 1** Variation of human subjects covered in supervisely persons dataset

| Area coverage of humans in the dataset | |
|---|---|
| Range | Percentage area covered |
| Minimum area | 0.1% |
| Maximum area | 99.7% |
| Average area | 26.1% |

to the models in comparison. The architectural differences thus help in the segmentation task here. We also observe that MobileNetv2 overfits in phase-I but overall improves in phase-II. DenseNet and Ensemble are the best models, when fine-tuned, with minimal differences in training and validation data.

We did not consider training further in phase-II, for later epochs lead to overfitting on the data, thus producing unwanted results. Also, because Dice Loss is just dice coefficient subtracted from one, we can consider the inversed graph of phase-II in Fig. 11 as the Dice-Coefficient metric during training in the second phase.

### 6.3 Evaluation metrics

The segmentation results are measured by the standard metric used in image segmentation, i.e., Intersection-over-Union (IOU). This metric is computed as the intersection of predicted and ground truth segments over their union. The same is defined in Eq. 5.

$$IOU(\%) = \frac{area(output \cap groundtruth)}{area(output \cup groundtruth)} \times 100 \tag{5}$$

We could have also used Dice Coefficient or (F1 score) as mentioned in our loss functions, for it is also used as an evaluation metric since both the Dice Coefficient and IOU are very similar.

### 6.4 Ablation study

To further understand and show the effectiveness of the proposed techniques presented, we provide an ablation study through this section wherein experiments are performed over three parameters to observe their influence on the produced metrics for each of the encoding techniques used. Those parameters include loss function fine-tuning, the influence of data augmentation, and effect of concatenation connections. The following study show results over test data divided via hold-out validation using the mIOU metric discussed in the Section 7.

#### 6.4.1 Influence of loss function fine tuning

As discussed in the Section 3.6, our task is that of a binary segmentation one thus binary cross-entropy is the first choice of loss to optimize upon but despite its wide usage this type of loss is not able to account for class imbalance i.e., regions in the image where segments are less or more compared to other images in the dataset. Thus, to account for the imbalance dice coefficient loss is used to further fine-tune the BCE loss trained architectures. Since dice coefficient loss is a regional loss aimed to improve the region boundary instead of pixel-wise classification in BCE, it produces much better results in comparison. Though starting the training directly through dice coefficient, in

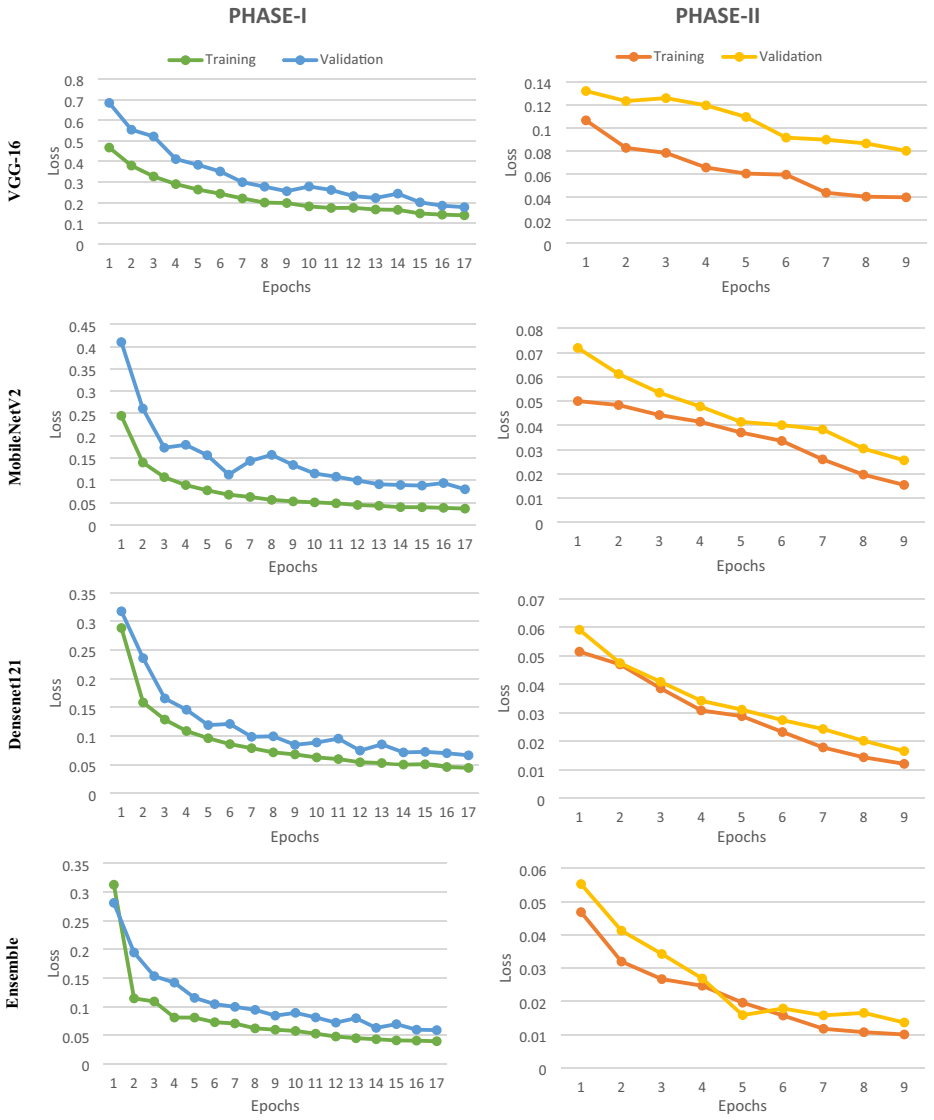



**Fig. 11** Variation of Loss in both Phases for all the considered encoder networks. Clearly, Densenet and Ensemble provide the best convergence and minimal overfitting

many tasks, turns out to be difficult to optimize thus tuning provides better results. The results for the fine-tuning can be observed in Table 2, where each technique is trained for 26 epochs following the same phases for the learning rate discussed before. The bold text in Tables 2, 3, 4, 5 and 6 shows the highest performance parameters achieved by the different models.

### 6.4.2 Effect of concatenation connections

Concatenation connections play a major role in helping the HISNet converge over the loss function quickly and achieve respectable accuracies. Though, the selection of concatenation

**Table 2** Accuracy of the 4 models using BCE, dice coefficient, and tuning method for 26 epochs each

| HIS Encoder Architecture | Loss technique | mIOU (%) |
|---|---|---|
| HIS-VGG-16 | BCE | 86.54 |
| | Dice Coeff. | 88.65 |
| | BCE-Dice-Tune | **89.62** |
| HIS-MobileNetv2 | BCE | 89.74 |
| | Dice Coeff. | **93.36** |
| | BCE-Dice-Tune | 92.54 |
| HIS-DenseNet121 | BCE | 88.43 |
| | Dice Coeff. | 91.01 |
| | BCE-Dice-Tune | **92.17** |
| HIS-Ensemble | BCE | 89.19 |
| | Dice Coeff. | 92.15 |
| | BCE-Dice-Tune | **93.89** |

connection plays a big role in determining the accuracies too. For the decoder shown in Fig. 5, we are concatenating features from each dimensional block in the encoder to make sure we are easily able to transfer the knowledge learned in the encoder to each respectable decoder block. For this purpose, we are selecting intermediate encoder convolutional blocks for each dimension which provided us with the best overall results as observed in Table 3, compared to the use of initial and end convolutional layers in each dimensional block for concatenation. The selection of these concatenation layers through each of the backbone structure is done regressively near each location to get the best accuracy in each case.

### 6.4.3 Influence of data augmentation

Augmentation of data involves the use of existing data to generate new samples. For image data, there are many instances including the likes of horizontal/vertical flipping, random cropping, rotation, scaling, etc. The experimented dataset contains many samples from large to small with many samples including multiple subjects thus we are not using cropping or scaling techniques. To improve the accuracy of the architecture we apply horizontal flipping and rotation [-50°, 50°] augmentations, since training on data instances covering different regions can help generalize the architectures better, the same can be observed in Table 4.

**Table 3** Accuracy of the 4 models using different concatenation connections trained on same fine-tune loss method discussed before

| HIS Encoder Architecture | Concatenation location | mIOU (%) |
|---|---|---|
| HIS-VGG-16 | Initial | 88.14 |
| | Mid | 89.62 |
| | End | **90.49** |
| HIS-MobileNetv2 | Initial | 91.46 |
| | Mid | **92.54** |
| | End | 90.79 |
| HIS-DenseNet121 | Initial | 90.83 |
| | Mid | **92.17** |
| | End | 91.95 |
| HIS-Ensemble | Initial | 92.72 |
| | Mid | **93.89** |
| | End | 92.31 |

**Table 4** Accuracy of the 4 models with image augmentation to improve the accuracy

| HIS Encoder Architecture | Regularization | mIOU (%) |
|---|---|---|
| HIS-VGG-16 | None | 90.49 |
| | Augmentation | **91.21** |
| HIS-MobileNetv2 | None | 92.54 |
| | Augmentation | **94.65** |
| HIS-DenseNet121 | None | 93.17 |
| | Augmentation | **95.23** |
| HIS-Ensemble | None | 93.89 |
| | Augmentation | **95.88** |

## 7 Results

We analyze the segmentation results on the four HISNets, applying different encoders. The models performed great on unseen images, evident from visual results in Fig. 6. The depth estimation network then aids in-depth effect generation, and the resultant image gives the desired results, evident from the pipeline shown in Fig. 8 and rendered results in Fig. 9.

### 7.1 Quantitative and visual analysis

Based on the data-splitting, we test our models on a test-set with about 851 images. We compute the mean IOU (mIOU) score for each of the four models on test data. Further, to test with other state-of-the-arts, we train a few techniques over given dataset using the same training strategy applied by HISNet. From the visual results in Fig. 6 and qualitative results in Table 5, it is evident that the choice of the encoder for our human image segmentation is with the ensemble architecture and the proposed techniques provide a much higher mIOU over other state-of-the-arts under similar conditions while using pre-trained backbone like with the proposed techniques. Thus, we utilize our ensemble model for segmentation in HISNet, in all the results rendered in Figs. 1, 8 and 9.

Next, to test our results, for comparison of our HIS-Ensemble network, we test our model on the Portrait FCN dataset, also known as EG1800. Since in selfie images, the subject is almost always very close to the camera, it becomes an easy task for our model to predict for such examples without training on that particular dataset, but for a fair comparison, we compare a trained and un-trained (our Supervisely trained) version of the network. Also, the Portrait FCN model by Shen et al. [31] required detection of the cropped regions of humans for its model to generate the segmentation map; thus, the dataset comprised of only near sight

**Table 5** Accuracy of the 4 models on supervisely test set mentioned in Section 6.1

| Model Architecture | Parameters (M) | mIOU (%) |
|---|---|---|
| HIS-VGG-16 | 14.6 | 91.21 |
| HIS-MobileNetv2 | **1.3** | 94.65 |
| HIS-DenseNet121 | 6.2 | 95.23 |
| HIS-Ensemble | 7.7 | **95.88** |
| PSPNet (ResNet-50) [46] | 3.8 | 79.20 |
| PSPNet (ResNet-152) [46] | 4.9 | 85.38 |
| FPN (ResNet-152) [21] | 61.6 | 85.93 |
| Attention Unet [24] | 10.1 | 68.14 |

portrait images, since for multiple humans, it applied object detection. Our model overcomes those steps, for our network is trained on more variations of human sizes in the Supervisely Persons Dataset. The variation in images present in the dataset helps the model learn the human subject with very close and accurate segments.

The Portrait dataset or EG1800 has many recent networks for comparison. We refer to the data of some of the current state-of-the-art models, given by Park et al. [25], to compare our HISNet's performance on F1-score and mean IOU as shown in Table 6. We have shown our results for training on both the datasets, i.e., we test HISNet trained on Supervisely Persons Dataset with testing on EG1800 test set (provided in the dataset), and another comparison being HISNet trained on EG1800 training-data to get the comparison with the other networks trained on the same data distribution. We consider the Ensemble (Section 3.4) as the encoder, for it appeared as the best one compared to the other three encoders, as shown in Table 5. The training on EG1800 data is similar to the one discussed in Section 6, done in 2 phases with the same number of epochs and optimizing techniques. Also, the ensembled encoders are again pre-trained on the 'image-net' dataset as we did for Supervisely.

All the networks mentioned in Table 6 were inferred by Park et al. [25] at $224 \times 224$, except for PortraitNet and BiSeNet, whose mIOU scores are referenced from their original work. Our network, HIS-Ensemble, inference at input dimension of $256 \times 256$, for two different datasets, the network trained on the Supervisely Persons dataset produced some significant comparable results with the highest F1-score metric produced compared to the older networks. Even though the mean IOU score was just the second last among the considered networks, the results were significant considering the training data was different from other models. The results for HISNet trained on EG1800 surpassed the F1-score of Supervisely trained network, with mIOU being the highest among other networks. Thus, our HISNet produced significant results for human image segmentation with greater benchmarks on the EG1800 dataset. Figure 12 provides some results on different images present in the EG1800 dataset to visualize our segmentation results. The high F-1 scores can be attributed to the second phase of training since the dice coefficient is the same as the F-1 score.

From Fig. 12, most ground truth has very accurate segmentation maps refined up to a single hair strand. In contrast, Supervisely annotation covers a whole hair region as single-segment, for even a single strand. Thus, these shortcomings may account for some reductions in mIOU scores for a model trained on Supervisely Dataset. However, we can even see that the results are comparable for the two different data-trained models.

Overall, it can also be observed that in portrait subject images, the depth network can provide a similar but rough segmentation map which can be corrected to finer and more accurate segments by our proposed HIS network. However, as the subject appears to be distant

**Table 6** Comparison of performance with state-of-the-art models on EG1800 test-data

| Model | Training Dataset | F1-Score | mIoU (%) |
|---|---|---|---|
| BiSeNet [42] | EG1800 | 0.908 | 95.25 |
| ContextNet [27] | EG1800 | 0.896 | 95.71 |
| PortraitNet [43] | EG1800 | 0.919 | 96.62 |
| ESPNetV2(2.0) [23] | EG1800 | 0.872 | 94.71 |
| MobileNetV3 [14] | EG1800 | 0.854 | 94.19 |
| SINet+ [25] | EG1800 | 0.892 | 95.29 |
| **HIS-Ensemble(Ours)** | **EG1800** | **0.9788** | **96.89** |
| **HIS-Ensemble(Ours)** | **Supervisely Persons Dataset** | **0.9676** | **94.41** |

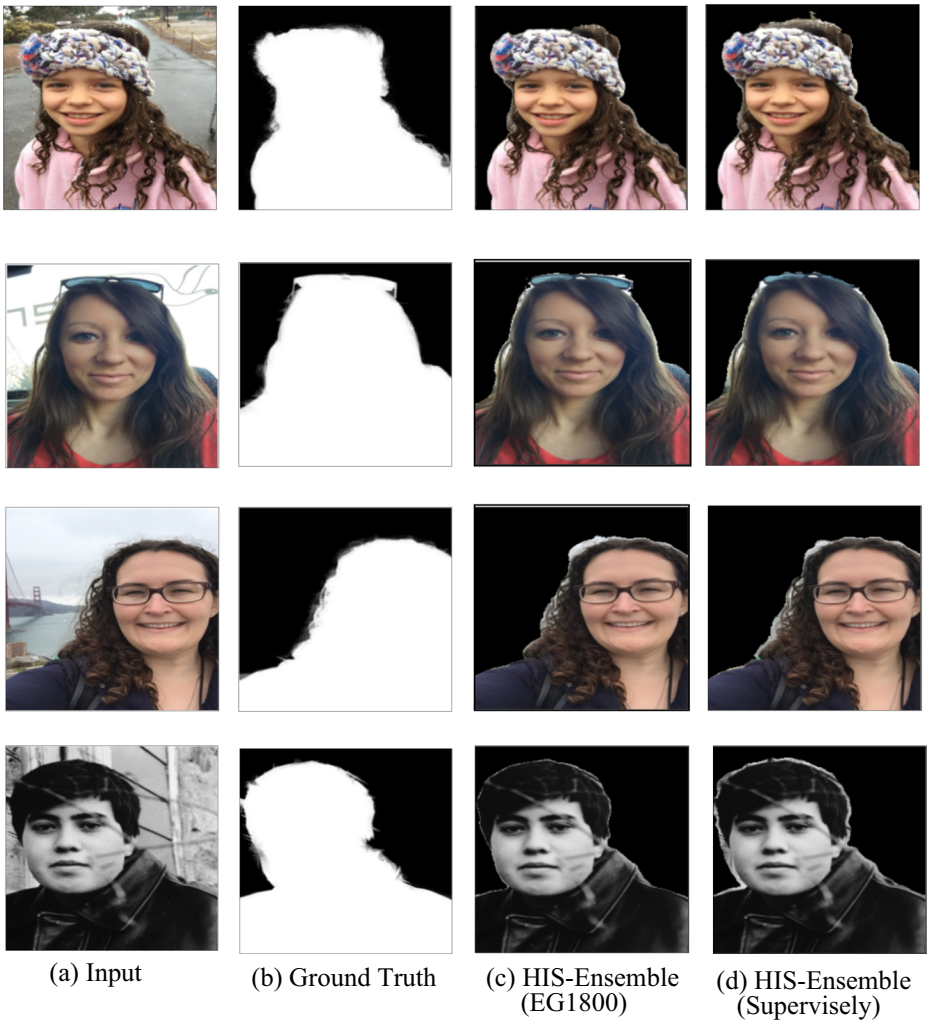



**Fig. 12** Some visual results on EG1800 dataset

or when the image has too much variation in the background, the depth network gets confused in such cases; here, the segmentation network helps include human subjects neglected by the depth network. Thus, both the networks go hand-in-hand to correct the depth effect generated via the segmentation maps. As previously discussed in Section 4.

## 8 Conclusion and future work

In this paper, we demonstrated an approach to use depth estimation and image segmentation methods to apply the bokeh effect on images captured using single-lens mobile cameras. First, we introduced our novel HISNet, which appeared similar to an FCN but with a very shallow decoder with explicit concatenation connections and application of some state-of-the-art encoders. We applied four different encoder architectures and applied their pre-trained

networks for quick convergence on the loss function and compared the performance between each of them to get the best HISNet for human segmentation. We applied the method of encoder concatenation to get Ensemble HISNet as our best network. We also benchmarked our model on the EG1800 dataset with some significant results for comparison.

The developed solution achieves significant performance on unseen images, produces accurate segmentation maps, and thus help generate beautiful depth effect images. However, the presented approach is capable of generating effects in human subject images only. In the future, we plan to improve our decoder network for higher accuracy and extend our dataset to more classes for the classification of more photo-specific subjects like flowers, animals, food, etc. Also, the experiments here are carried on an image size of $256 \times 256$; given the advent in the resolution of cameras, this size is not enough for high-definition pictures, so we plan to increase the image resolution used during training for better results.

**Data availability** It is declared that the dataset used in this manuscript is obtained from following sources and there are freely available.

The Supervisely Persons Dataset that supports the findings of this study are available in/from supervisely-ecosystem/persons GitHub Repository, [https://github.com/supervisely-ecosystem/persons].

The EG1800 Dataset that supports the findings of this study were extracted from the link provided by the original authors: http://xiaoyongshen.me/webpage_portrait/index.html, in their paper available at DOI: https://doi.org/10.1111/cgf.12814.

The citation of these datasets are given and used ethically.

## Declarations

**Conflict of interest** It is declared that there is no conflict of interest.

## References


1. Alhashim I, Wonka P (2018) High quality monocular depth estimation via transfer learning
2. Badrinarayanan V, Kendall A, Cipolla R (2017) SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615
3. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018) DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848. https://doi.org/10.1109/TPAMI.2017.2699184
4. Chen J et al (2021) A transfer learning based super-resolution microscopy for biopsy slice images: the joint methods perspective. IEEE/ACM Trans Comput Biol Bioinforma 18(1):103–113. https://doi.org/10.1109/TCBB.2020.2991173
5. Chen J et al (2021) TransUNet: transformers make strong encoders for medical image segmentation, pp 1–13
6. Digital single-lens reflex camera - Wikipedia
7. Fei-Fei L, Deng J, Li K (2010) ImageNet: constructing a large-scale image database. J Vis 9(8):1037. https://doi.org/10.1167/9.8.1037
8. Feng R et al (2021) ChroNet: A multi-task learning based approach for prediction of multiple chronic diseases. Multimed Tools Appl. https://doi.org/10.1007/s11042-020-10482-8
9. Fu J et al (2019) Dual attention network for scene segmentation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019:3141–3149. https://doi.org/10.1109/CVPR.2019.00326
10. Gao H, Xu K, Cao M, Xiao J, Xu Q, Yin Y (2022) The deep features and attention mechanism-based method to dish healthcare under social IoT systems: an empirical study with a hand-deep local-global net. IEEE Trans Comput Soc Syst 9(1):336–347. https://doi.org/10.1109/TCSS.2021.3102591
11. Godard C, Aodha OM, Brostow GJ (2017) Unsupervised monocular depth estimation with left-right consistency. Proc. – 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol 2017, pp 6602–6611. https://doi.org/10.1109/CVPR.2017.699

12. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition
13. Howard AG et al (2017) MobileNets: efficient convolutional neural networks for mobile vision applications
14. Howard A et al (2019) Searching for mobileNetV3. Proc IEEE Int Conf Comput Vis 2019:1314–1324. https://doi.org/10.1109/ICCV.2019.00140
15. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. https://doi.org/10.1109/CVPR.2017.243
16. Kirkland EJ (2010) Advanced computing in electron microscopy: second edition. Adv Comput Electron Microsc Second Ed 1–289. https://doi.org/10.1007/978-1-4419-6533-2
17. Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet classification with deep convolutional neural networks. ACM Int. Conf. Proceeding Ser., pp 145–151. https://doi.org/10.1145/3383972.3383975
18. Laina I, Rupprecht C, Belagiannis V, Tombari F, Navab N (2016) Deeper depth prediction with fully convolutional residual networks. Proc. – 2016 4th Int. Conf. 3D Vision, 3DV 2016, pp 239–248. https://doi.org/10.1109/3DV.2016.32
19. Lee JH, Kim CS (2019) Monocular depth estimation using relative depth maps. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019:9721–9730. https://doi.org/10.1109/CVPR.2019.00996
20. Lin TY et al (2014) Microsoft COCO: common objects in context. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol 8693 LNCS, no PART 5, pp 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
21. Lin T-Y, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp 2117–2125
22. Martinez M, Yang K, Constantinescu A, Stiefelhagen R (2020) Helping the blind to get through covid-19: Social distancing assistant using real-time semantic segmentation on rgb-d video. Sens (Switzerland) 20(18):1–17. https://doi.org/10.3390/s20185202
23. Mehta S, Rastegari M, Shapiro L, Hajishirzi H (2019) ESPNetv2: a light-weight, power efficient, and general purpose convolutional neural network. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2019:9182–9192. https://doi.org/10.1109/CVPR.2019.00941
24. Oktay O et al (2018) Attention U-Net: learning where to look for the pancreas, no. Midl
25. Park H, Sjosund LL, Yoo Y, Monet N, Bang J, Kwak N (2020) SINet: Extreme lightweight portrait segmentation networks with spatial squeeze modules and information blocking decoder. Proc. – 2020 IEEE Winter Conf. Appl. Comput. Vision, WACV 2020, vol 2, no 1, pp 2055–2063. https://doi.org/10.1109/WACV45572.2020.9093588
26. Paszke A, Chaurasia A, Kim S, Culurciello E(2016) ENet: a deep neural network architecture for real-time semantic segmentation, pp 1–10
27. Poudel RPK, Bonde U, Liwicki S, Zach C (2019) ContextNet: exploring context and detail for semantic segmentation in real-time. Br. Mach. Vis. Conf. 2018, BMVC 2018
28. Ranftl R, Lasinger K, Hafner D, Schindler K, Koltun V (2020) Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer. IEEE Trans Pattern Anal Mach Intell XX(Xx):1. https://doi.org/10.1109/tpami.2020.3019967
29. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) MobileNetV2: inverted residuals and linear bottlenecks. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp 4510–4520. https://doi.org/10.1109/CVPR.2018.00474
30. Shelhamer E, Long J, Darrell T (2015) Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 39(4):640–651. https://doi.org/10.1109/TPAMI.2016.2572683
31. Shen X et al (2016) Automatic portrait segmentation for image stylization. Comput Graph Forum 35(2):93–102. https://doi.org/10.1111/cgf.12814
32. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp 1–14
33. Smartphones Cause Photography Boom by Felix Richter (Aug 31, 2017), Statista
34. Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M (2017) Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10553 LNCS, pp 240–248. https://doi.org/10.1007/978-3-319-67558-9_28
35. Supervisely Person Dataset - Datasets - Supervisely
36. Weng W, Zhu X (2015) UNet: convolutional networks for biomedical image segmentation. IEEE Access 9:16591–16603. https://doi.org/10.1109/ACCESS.2021.3053408
37. Xiao J, Xu H, Gao H, Bian M, Li Y (2021) A weakly supervised semantic segmentation network by aggregating seed cues: the multi-object proposal generation perspective. ACM Trans Multimed Comput Commun Appl 17(1s). https://doi.org/10.1145/3419842

38. Xiao J, Xu H, Fang DK, Cheng C, Gao HH (2021) Boosting and rectifying few-shot learning prototype network for skin lesion classification based on the internet of medical things. Wirel Netw 0123456789:1–15. https://doi.org/10.1007/s11276-021-02713-z
39. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P (2021) SegFormer: simple and efficient design for semantic segmentation with transformers, no. NeurIPS, pp 1–14
40. Xu X et al (2018) Rendering portraitures from monocular camera and beyond Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol 11213 LNCS, pp 36–51. https://doi.org/10.1007/978-3-030-01240-3_3
41. Yang K et al (2018) Unifying terrain awareness for the visually impaired through real-time semantic segmentation. Sens (Switzerland) 18(5):1–32. https://doi.org/10.3390/s18051506
42. Yu C, Wang J, Peng C, Gao C, Yu G, Sang N (2018) BiSeNet: Bilateral segmentation network for real-time semantic segmentation. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 11217 LNCS:334–349. doi: https://doi.org/10.1007/978-3-030-01261-8_20
43. Zhang SH, Dong X, Li H, Li R, Yang YL (2019) PortraitNet: real-time portrait segmentation network for mobile device. Comput Graph 80:104–113. https://doi.org/10.1016/j.cag.2019.03.007
44. Zhang T, Lang C, Xing J (2019) Realtime human segmentation in video, vol 11296 LNCS. Springer International Publishing
45. Zhang J, Yang K, Constantinescu A, Peng K, Muller K, Stiefelhagen R (2021) Trans4Trans: efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world. Proc IEEE Int Conf Comput Vis 2021:1760–1770. https://doi.org/10.1109/ICCVW54120.2021.00202
46. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. Proc. – 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol 2017, pp 6230–6239. https://doi.org/10.1109/CVPR.2017.660

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

Springer Nature or its licensor 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.

FOREX Publication
Open Access | Rapid and quality publishing

International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 10, Issue 3 | Pages 447-453 | e-ISSN: 2347-470X

# Identifying and Mitigating the Barriers for Vehicle-to-Grid Adoption in India

**Aditya[1], Rachit Bansal[2], Akshit Goel[3], N.K. Jain[4] and Uma Nangia[5]**

*[1,2,3]Department of Electrical Engineering, Delhi Technological University, Delhi, India, aditya_2k18ee008@dtu.ac.in[1], rachitbansal_2k18ee152@dtu.ac.in[2], akshitgoel_2k18ee017@dtu.ac.in[3]*
*[4]MIE, LMISTE, FMIETE, Department of Electrical Engineering, Delhi Technological University, Delhi, India, vnarender84@yahoo.com*
*[5]MIE, LMISTE, Member IEEE, Department of Electrical Engineering, Delhi Technological University, Delhi, India, umanangia@dce.ac.in*

***Correspondence:** Aditya; Email: aditya_2k18ee008@dtu.ac.in

**ABSTRACT-** This study gauges the current scenario of EVs in India as a precursor to the adoption of V2G on a large scale. It outlines the barriers to complete EV adoption under three challenge categories. It discusses the motivation for the use of vehicle-to-grid by describing the technology in detail and discussing an overview of how it works. Lastly, the study outlines how popular optimization techniques have been employed to solve individual optimization and scheduling tasks to optimize power, cost, and emissions for V2G.

**General Terms:** Particle Swarm Optimization, Double-Layer Optimal Charging, Cost and Emission Optimization.

**Keywords:** Vehicle-to-Grid Technology, Electric Vehicles, Smart Grid and Power Optimization.

ARTICLE INFORMATION

**Author(s)**: Aditya, Rachit Bansal, Akshit Goel, N.K. Jain and Uma Nangia.
**Received**:05/04/2022; **Accepted**: 18/07/2022; **Published**: 10/08/2022;
**e-ISSN**: 2347-470X;
**Paper Id**: IJEER100307;
**Citation:** 10.37391/IJEER.100307
**Webpage-link**:
https://ijeer.forexjournal.co.in/archive/volume-10/ijeer-100307.html

This article belongs to the Special Issue on **Recent Advancements in the Electrical & Electronics Engineering**

**Publisher's Note:** FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

## 1. INTRODUCTION

Recent years have seen significant interest in two booming directions- renewable energy and electric vehicles (EVs). The two naturally reinforce each other. As larger portions of our energy start to come from renewable sources---the cost, efficiency, and environmental control with the adoption of EVs is expected to be more prominent. To this end, the advent of Vehicle-to-Grid (V2G) technology promises to further mitigate the individual foreseen challenges. V2G considers EVs as small portable power plants (S3Ps) or energy storage systems, thus having the ability to provide much-needed stability to renewable energy sources. Not only does this allow us to make best use of our resources, but also removes the reliance on conventional energy sources and the associated pitfalls.

With the rapidly expanding Indian Automobile sector, there is great momentum for Passenger Electric Vehicles (PEVs). The Automobile sector is projected to reach INR 16.16-18.18 lakh crores (US $251.4-282.8 billion) by 2026 [5], and along with it, electric vehicles are expected to have a significant share in the future. The industry data for the first half of FY2021-22 shows that a total of 6,261 PEVs were sold during the period, i.e., April to September 2021, exhibiting a 234% Year on Year growth although it has a low base of the previous year still, PEVs contribute less than even 1% of total passenger vehicles that are sold in India.

The first half of the financial year 2021-22 experienced a total sale of 13A, 87,714 Passenger vehicles where electric vehicles accounted for only 0.45 percent of the total passenger vehicles market. But there is an expected growth as the 6,251 PEVs sold during the first half of FY 2021-22 have already surpassed the total PEVs sold in FY 2020-21 of 5,905 units [16].

In the past 10-15 years, Wind and Solar energy have proved to be a trusted sources of sustainable energy production, but there is an associated problem with these energy sources, i.e., these sources are not regular and reliable, which means that they cannot provide energy to the electrical grid at the need of peak hours and we have to build reserve power plants, which demands price hikes during peak hours but these plants prove to be a pricey measure. To mitigate a significant percentage of variable renewable energy generation and fluctuations, an electric car can be employed as a flexible load to help standardize the grid, which is the main motive behind this study and the **Vehicle to Grid** (V2G) technology.

With the support of investments from various Original Equipment Manufacturers (OEMs), there is an exception for the rise of more and lower-cost EVs in the next years due to hazardous emissions coming out from the transportation sector. There are some challenges and barriers in the road ahead for the growth of PEVs in India; some of them include Battery Optimization, battery state estimation, and equalization, large-scale charging infrastructure. Technology breakthroughs, government support, incentives for purchase, parking benefits,

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 10, Issue 3 | Pages 447-453 | e-ISSN: 2347-470X

lower car costs, and enough public charging infrastructure could help India in adoption of Electric Vehicles.

# 2. RELATED WORK

## 2.1 Vehicle-to-Grid

Vehicle-to-Grid technology is a technology that draws unused power from the cars back into the power grid. With this technology, a car's battery can be charged as well as discharged on the basis of the need or excess of energy nearby [13].

This technology can help the power grid during peak hours on National Level and can help to power nearby buildings or fans and lighting at the local level [11].

For example, In a region where a large percentage of power is being generated via means of renewable energy sources like Solar and Wind Power but since, they can't supply power during the night and the power output largely depends on the external environmental conditions that are hard to predict and can sometimes harm the power grid as well [13].On the other hand, the batteries in electric cars are just lying as potential energy stored in batteries that are portable and can be employed as per the need. V2G technology can act as a bridge to solve the issue of fluctuations in power production by acting as a secondary source of power when needed [17].

Our study is particularly inspired by similar assessments and analysis done for other regions of the world, considering economic and behavioral factors [7].

## 2.2 Electric Vehicle Adoption in India

The electric car market in India is at a nascent stage. For the previous two years, the sales of electric cars have been hovering at around 2000 units per year. However, by 2030, it is anticipated to have sales of majority electric vehicles, with a compound annual growth rate of 28.12% since 2020.The Reva (Mahindra) which was India's first electric vehicle, was debuted in 2001 but the sales have remained muted since then. In 2010, Toyota debuted the Prius hybrid, succeeded by the Camry hybrid in 2013. As part of a pilot project to test the feasibility, electric buses and hybrid cars have been tested and tried in a few locations. The government is offering a scheme that allows consumers to purchase an electric vehicle at a cheap cost. The Indian government has created the National Electric Mobility Mission 2020, which aims to lower the cost of electric vehicles and increase their sales in India. The Indian government's Niti Ayog plans the policies that will be implemented for electric mobility. Niti Ayog is also working hard to improve the conditions for the sale of electric automobiles in India. In India, low-cost financing is a serious hindrance to the growth of electric vehicles. In comparison to other countries, India is a late adopter of electric mobility technology. And Indians are having a hard time adjusting to this new technology. Even if they acquire an electric car, they will not be able to utilize it on a daily basis or for long-distance travel. Poor battery technology is to blame for this, as it delivers less range and takes longer to fully charge. In India, there is also a scarcity of charging stations where consumers may power their vehicles.

Several recent studies look at the resistive adoption from an economic perspective [8, 6] while others perform intensive analysis of other potential factors that influence this change [14, 15, 2, 1].

## 2.3 Optimization

Currently, Vehicles that are involved in V2G technologies are more generally referred to as gridable vehicles, and they essentially act as small portable power plants (S3P). In order to bring into effect, the true potential of V2G and have a tangible impact with respect to power generation and storage, it is essential to employ intelligent and effective optimization techniques [17].

Several parameters including cost [22], emissions [24], operations [12], and scheduling [23] are necessary to ensure an effective system without underlying inconvenience to users.

# 3. ELECTRIC VEHICLES

The power electronic traction inverter, which places an electric motor, in place of internal combustion engine, which is powered by energy stored in batteries, is the ultimate goal of electric vehicles. The electric motor is extremely efficient, using 90–95 percent of the input energy to successfully power the vehicle. An electric vehicle's battery, power electronics controller, charging port, charger, DC/DC converter, regenerative braking, and drive system are all required. An electric car's objective is to use the energy stored in batteries to power the vehicle. EVs allow the environment it is recharged with low emissions. Cells are charged in the electrical grid. The charging port is an area that allows the car to connect to an external charging system to charge the battery.

The charger function uses the charging connector to connect to a power source and transforms to DC power to charge the battery. It also keeps track of voltage, current, temperature, and battery charging status. The DC/DC converter transforms high DC voltage extracted from the battery into low DC power that can be used to power automotive accessories. The towing truck's speed and torque are controlled by the electronics power controller, which regulates the electrical energy flow from the towing battery. Regenerative braking is crucial for keeping the car's strength and improving its strength. This braking method recharges the battery by converting kinetic energy into electrical energy using mechanical energy from the engine. The drive system's job is to move the vehicle by distributing the traction wheels' power. An electric automobile can have a variety of interior configurations based on its components and does not require a standard transmission.

## 3.1 Types of Electric Vehicles

### 3.1.1 Battery Electric Vehicles

They are entirely electric vehicles that are equipped with rechargeable batteries and without any gasoline engine, sometimes known as BEVs and more commonly referred to as EVs. The energy that is required to operate the car is provided by the battery pack, which is recharged from the grid. Battery Electric Vehicles are considered emission less cars since they

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 10, Issue 3 | Pages 447-453 | e-ISSN: 2347-470X

produce no harmful emissions coming out from exhaust or contaminate the air in the same manner as other vehicles with gasoline-powered engine do.

#### 3.1.2 Plug-in Hybrid Electric Vehicles

PHEVs (or plug-in hybrid electric vehicles) have an engine and an electric motor to power the vehicle. They, similar to conventional hybrid cars, can recharge their batteries by the help of regenerative braking technology. But they do differ from other hybrid automobiles in a manner that they possess a much larger battery that can be recharged by connecting it to the grid. PHEVs have a range typically from 10 to 40 miles before switching to gasoline (considering the drive at moderate speeds), but regular vehicles can only sustain 1-2 miles and then we need to switch to gasoline engine.

#### 3.1.3 Hybrid Electric Vehicles

A gasoline engine along with an electric motor power HEVs. All of the energy consumed for the battery is provided by regenerative braking, which recovers power that is supposed to be lost otherwise and helps to assist the gasoline engine during acceleration in the process of braking. Heat is dissipated as breaking energy in the heat pads and rotors in a standard IC engine car.

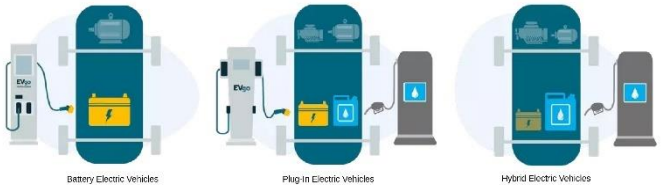



**Figure 1:** Types of Electric Vehicles

## 4. ELECTRIC VEHICLES IN INDIA

A technology like vehicle-to-grid is largely dependent on the mass population switching to electric vehicles and adopting bi-directional charging. When looked at from an individual perspective, the benefits of V2G seem insignificant--only when millions of individuals jointly adopt such a framework does this technology reap its true potential. Thus, in order to assess the feasibility and practicality of this technology, it is essential to gauge where we stand. In this section, we would take a look at the current scenario of electric vehicles in India, the most prominent barriers in the way of their mass adoption, and a perspective on the future.

### 4.1 Barriers

Several past studies have reviewed the barriers to complete EV adoption in India [3, 4]. In this report, we classify the challenges of EV adoption in India into three main categories: (i) Operation and Maintenance Challenges, (ii) Technical Challenges, and (iii) Infrastructural Challenges.

#### 4.1.1 Operation and Maintenance

Switching from traditional gas-powered vehicles to electric vehicles not only involves a large change with respect to how an individual operates and maintains the vehicle but also how the entire market around transportation and repair functions.

Firstly, a significant amount of expertise and human skill is already present around gas-powered vehicles in India. Considering the structural differences between such vehicles and EVs, this manual skill-set and market would need to change immensely.

Moreover, frequent, and well-operated charging stations would need to be set up throughout the city. Governments and organizations are already giving thought in this direction. For instance, the New Delhi government announced its vision of setting more than a hundred public charging stations in the state this year. This is encapsulated in the larger goal of converting at least one-fourth of the current vehicles to EVs and developing a charging point at a radius of every 3 KM.

Further, electric vehicles involve a high capital cost in comparison with gas-powered vehicles. The battery packs of EVs are costly and need to be changed after every several years in order to ensure the proper functioning of the vehicle. This makes the position of EVs less certain among the economic divisions of India.

#### 4.1.2 Operation and Maintenance

One of the most prominent concerns around the adoption of EVs revolves around the driving range. In comparison to similar standard gas-powered vehicles, EVs have a much smaller range of distance that can be travelled with full charging. The current state-of-the-vehicles offer up to 450 KMs per charge, however, the average across their lifetime remains a low of 250 KMs. Due to this reason, consumers must plan their journey with precision and longer trips remain difficult to perform [10, 20].

This concern of low driving range is strongly coupled with the large charging time. Gas-powered vehicles take minutes to “recharge” because they just need their fuel tanks to be filled up. EVs, on the other hand, require several hours to recharge their electric batteries.

Several types of technologies exist around charging EVs. Particularly, the most popular three types are Level 1, Level 2, and DC fast. Level 1 charging makes use of a standard 120 V outlet by performing AC-DC conversion using an onboard converter. Generally, it has been shown to take up to 8-9 hours for charging an EV for a range of 120-130 KMs, depending upon the vehicle. This kind of charging is usually designed for a home or workplace setup. On the other hand, Level 2 chargers are usually set up at public places, employing a 240 V outlet. It takes up to 4 hours to charge the battery for a similar driving range. DC fast charging usually occurs in a charging station, allowing for the fastest charging arrangements with special AC to DC conversion. Stations are modified that can charge the vehicles quickly by supplying more power. It can also easily charge the battery in as low as 30 minutes for a range of 145 KMs. Currently, this method of charging is the most efficient in terms of time and convenience, however, it is also the most expensive to set up and maintain.

#### 4.1.3 Infrastructural

The batteries used in EVs generally last for a limited amount of time and are likely to wear out before the vehicle. Currently, the

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 10, Issue 3 | Pages 447-453 | e-ISSN: 2347-470X

costs of battery replacement is not being disclosed transparently by manufacturers and whenever the instance arises for battery replacement outside a given period, it adds to the expenses for the consumer. Not only that, the chemical elements like Lithium, Nickel, Manganese, Cobalt, and Titanium that are used in these batteries pose a concern to the environment. All these elements are scarce and their availability might rapidly decline in the future. Lithium-ion batteries alone consume more than five million tons of nickel per year, and this is estimated to reach 10–20 times in the future from current levels, with the advent of EVs [16].

# 5. VEHICLE-TO-GRID

It is a technology that draws unused power from the cars back into the power grid. With this technology, a car's battery can be charged as well as discharged on the basis of the need or excess of energy nearby.

This technology can help the power grid during peak hours on National Level and can help to power nearby buildings or fans and lighting at the local level.

For example, in a region where a large percentage of power is generated via the help of renewable energy sources like Wind and Power but since, they can't supply power during the night and the power output largely depends on the external environmental conditions that are hard to predict and can sometimes harm the power grid as well. On the other hand, the batteries in electric cars are just lying as potential energy stored in batteries that are portable and can be employed as per the need. V2G technology can act as a bridge to solve the issue of fluctuations in power production by acting as a secondary source of power when needed.

## 5.1 How it Works

The basic principle behind the working of Vehicle to Grid technology involves the extension of the electrical grid and considering the electric vehicles' batteries as a part of the electrical grid where we can store the power as well as draw the power from it when needed. In this way, the charging of cars evolves to a bidirectional process, where the power grid not only supplies electricity to the vehicle's battery but also utilize it as a power source that can meet various needs of energy consumption.

It is also an important component in the development of an intelligent distribution network, which is commonly known as a smart grid, where energy flows are constantly optimized by the help monitoring at every step along the way.

The most important thing to remember is that EV drivers have enough energy in their car batteries when they need it. When they go to work in the morning, the automobile battery must be properly charged to drive them to and from work at any time. A major requirement of V2G and any other charging system is that the EV driver be able to specify when they want to unplug the car and the battery percentage at the time.

In this technology, the smart system works in a way to optimize the power transfer and frequency to meet the needs of the electrical grid such that whenever there is a need for power in the Grid, the system draws the power from electric vehicles' cars and supplies to the nearest location where power is needed because the Power Grid also works in the same way of transferring the power via the shortest route possible. In the same way, whenever there is an excess supply of power in the grid, the system stores the excess power in the car batteries and uses them as energy storage devices.

This system works because the batteries are not 100% charges, rather fluctuate between the minimum power capacity required for sustained performance of the battery and the maximum power that is optimum for the battery life while keeping in mind to maintain the battery percentage at the level indicated by the user at the time when the user needs to use the vehicle.

## 5.2 Optimization Techniques

Vehicles that are involved in V2G technologies are more generally referred to as grid-able vehicles, and they essentially act as small portable power plants (S3P). In order to bring into effect, the true potential of V2G and have a tangible impact with respect to power generation and storage, it is essential to employ intelligent and effective optimization techniques.

### 5.2.1 Optimal Vehicle Scheduling

The use of Vehicle to Grid technology poses a problem of unexpected triggering events, that is, as the technology involves the user to indicate the time of departure and required battery percentage beforehand while leaving the car at charging station but in an event when the user unplugs the car before the scheduled departure time or if a new vehicle gets connected to the system, these are known as unexpected triggering events which might put a load on the power grid, destabilize the grid and thereby, negating the whole purpose of using V2G technology.

To deal with this problem, we can employ an Optimal scheduling scheme where there will be scheduling centers that communicate with the charging facilities [18] During the normal course of events, the charging facilities will send the inputs like time of departure, state of charge (SoC) of the car at the time of initiating the charging process, the required battery percentage as indicated by the user.

Whenever a car disconnects from the network unexpectedly before the scheduled departure time, the mathematical model is updated based on the changes in the inputs, and re-scheduling is done in the scheduling center using the revised model. Finally, to coordinate energy exchange between PEVs and the power grid, revised charging facility scheduling instructions will be provided. Until the next triggering event occurs, the procedures will continue.

However, if a PEV is disconnected at or after the expected time indicated by its owner, the rescheduling operation will not be triggered.

- The problem is solved by upgrading the optimization model and rescheduling the triggering events, which include the PEV when it is connected to the grid and unexpectedly when it is disconnected.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 10, Issue 3 | Pages 447-453 | e-ISSN: 2347-470X

- This method can deal with the uncertainty that arises from electric car connections that are stochastic.

Plug-in electric vehicles look forward to functioning as a revolutionary distributed energy storage system in V2G operation, assisting them in achieving the balance between supply and demand of the power grid while smoothing out power load profile fluctuations.

### 5.2.2 Optimizing Cost and Emissions

The most salient attributes of the V2G technology can be inculcated together in a unifying objective function, that would be minimized using the most appropriate optimization technique. Here, the objective function is composed of three such attributes, given as a function of the output power $P_i(t)$ at a time $t$: fuel cost $(FC_i(P_i(t)))$, emission $(EC_i(P_i(t)))$, and start-up cost $(SC_i(t))$. Each of them is given as a polynomial equation:

$$FC_i(P_i(t)) = a_i + b_i + c_i P_i^2(t) \quad \dots (1)$$

$$EC_i(P_i(t)) = \alpha_i + \beta_i P_i(t) + \gamma_i P_i^2(t) \quad \dots (2)$$

$$SC_i(t) = \begin{cases} h - cost_i & MD_i \le X_i^{off}(t) \le H_i^{off} \\ c - cost_i & X_i^{off}(t) > H_i^{off} \end{cases}$$

Hence, the net objective function that unifies cost and emissions is given by:

$$\min \prod = W_c \times (Fuel + Start - up) + W_e \times Emission$$

$$= \sum_{i=1}^{N} \sum_{t=1}^{H} [W_c(FC_i(P_i(t)) + SC_i(1 - I_i(t-1))) + W_e(\psi_i EC_i(P_i(t)))] I_i(t) \quad \dots (3)$$

where, $\psi_i$ represents the unit $i$ emission penalty factor. The weighing factors $W_c$ and $W_e$ can be used to consider ($W = 1$), not consider ($W = 0$), or consider with varying precedence ($0 < W < 1$) the cost/emissions in the fitness function, thus adding to the system flexibility.

Further, this optimization occurs under particular well-defined constraints in the V2G setup. Particularly, 12 constraints are considered for this optimization problem, as outlined in *Table 1*

The Particle Swarm Optimization (PSO) method can be effectively used to solve the problem of optimization that has been outlined above [21]. PSO is a biologically-inspired evolutionary algorithm. Each of the potential solution (termed as a particle) travels in a multi-dimensional space ans that too with a velocity that varies dynamically according to the particle's own experience and that of other particles in the neighborhood. PSO is an iterative method where the position as well as velocity of each particle is updated as follows:

$$v_{ijt} = w \times v_{ijt} + c_1 \times rand_1 \times (pbest_{ijt} - x_{ijt}) + c_2 \times rand_2 \times (gbest_{jt} - x_{ijt}) \quad \dots (4)$$

where, $v_{ijt}$ represents the current velocity of the considered particle (inertia). The cognitive element of the particle is given by ($pbest_{ijt}$ - $x_{ijt}$) where the agent changes the velocity in correspondence to its own memory. While the social element of PSO is depicted in the term ($gbest_{jt}$ - $x_{ijt}$), where the particle is changing its velocity at every step which is based on the social-psychological knowledge adaption from the entire swarm.

An N × H binary matrix which is used here for generating units can be easily optimized using PSO, owing to the 0/1 nature of possible states. Whereas, standard PSO has less balance b/w local and global entities to search for optimization of a given H × 1 column vector with integer states (as for gridable vehicles). This is because the number of vehicles connected with each other can range from 0 to $N_{V2G}^{max}(t)$ at an hour t. Thus, binary PSO has been used for the better optimization of vehicles of V2G here. The complete algorithms is given in *Figure 2*.

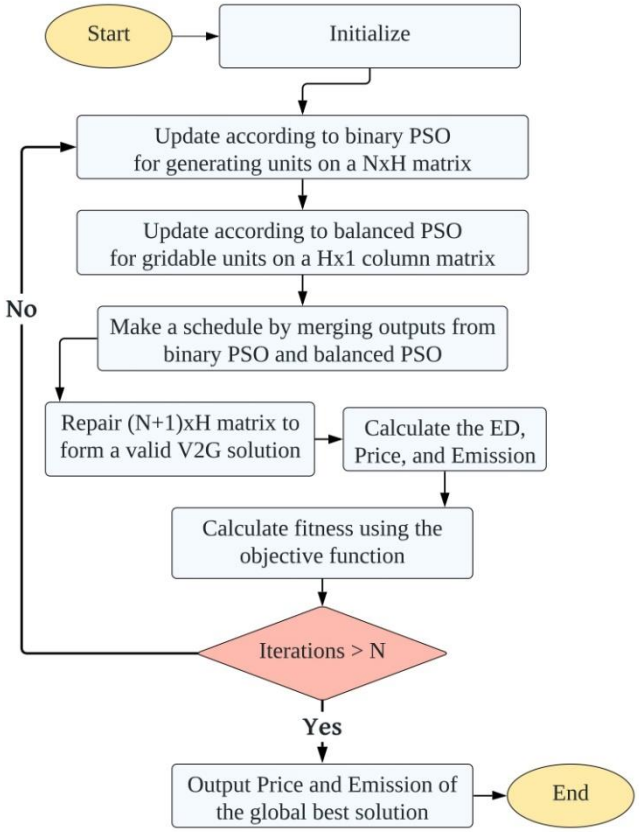



**Figure 2:** Particle Swarm Optimization for the given Vehicle-to-Grid cost and emission optimization problem. Joint use of balanced and binary PSO has been made to solve the problem efficiently

### 5.2.3 Optimizing Charging Schedule

When providing charging services for PEVs, the best charging strategy is found to be by calculating the charging power over each individual time slot as well as for separate charging port. The charging capacity of each charging point in the same time slot is however kept constant here. The goal of this technique is to reduce the regional grid's total load variation throughout the next one-day cycle.

A double-layer optimal charging (DLOC) technique can be employed to address the daunting problem of excessively high computational complexity caused by the increase large-scaled PEVs than ever and charging posts required by them [19].

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 10, Issue 3 | Pages 447-453 | e-ISSN: 2347-470X

Central concept used here is to group all of the regional grids' charging ports under the control of a few charging stations.

The CCC plans ideal operating power schedule for each charging station as a whole in the first layer optimization to decrease total load variation.

The station control server can then plan the required charging power for each charging post hosted under its control in optimization second layer, trying to satisfy the instructions provided by CCC in the first layer optimization. Charging centers that are located in the same streets, parking spaces, or the same residential complex might be considered the same charging station if they are in the exact location or linked to the same node transformer.

**Table 1: Constraints considered for the optimization of cost and emissions in V2G**

| Constraint | Formulation |
|---|---|
| Vehicle Balance | $\sum_{t=1}^{H} N_{V2G}(t) = N_{V2G}^{max}$ |
| Discharging Frequency | F=1/day |
| Power Balance | $\sum_{i=1}^{N} I_i(t)P_i(t) + P_v N_{V2G}(t) = D(t) + Losses$ |
| Spinning Reserve | $\sum_{i=1}^{N} I_i(t)P_i^{max}(t) + P_v^{max} N_{V2G}(t) \geq D(t) + R(t)$ |
| Generation Limits | $p_t^{min} \leq P_i(t) \leq p_t^{max}$ |
| State of Charge | - |
| Number of Discharging Vehicles | $p_t^{min} \leq P_i(t) \leq p_t^{max}$ |
| Efficiency | - |
| Minimum Up/Down Time | - |
| Ramp Rate | $p_t^{min} \leq P_i(t) \leq p_t^{max}$ |
| Prohibited Operations | - |
| Initial Status | - |

## 6. CONCLUSION

Compared to conventional automobiles, all variants of electric vehicles like Hybrid, Plug-in Hybrid can help in dealing with threats like Global Warming and Climate change as well as other issues like Energy Security, better deployment of renewable sources of energy. enhance fuel efficiency although increasing the cost of ownership. In certain cases where non-conventional energy sources might be unavailable to consume, we can still use a better Vehicle-to-Grid concept to either deliver surplus electricity whenever there is a need to the grid or simply charge the battery. In this paper, we've explored numerous technologies that can optimize various parts of the electric vehicle space, such as cost, rate of emission, and load on charging stations, with the goal of optimizing these parameters for the Indian electric vehicle market.

## 7. FUTURE WORK

There are a number of research possibilities in this area as we can better compare the most effective optimization techniques for the Vehicle to Grid technology once we have better measuring equipment and enough datasets to satisfy various variables required for different optimization methods. There is a need of fast and real-time data transfer with respect to the load variance by conventional and renewable sources of energy which will enable researchers to devise more accurate forecasting models for the optimal load sharing and scheduling of charging stations.

## REFERENCES

[1] Marc Barbar et al. "Scenarios of future Indian electricity demand accounting for space cooling and electric vehicle adoption". In: Scientific Data 8.1 (2021), pp. 1–11.

[2] Rubal Dua et al. "Enablers and disablers to plug-in electric vehicle adoption in India: Insights from a survey of experts". In: Energy Reports 7 (2021), pp. 3171–3188.

[3] Sonali Goel, Renu Sharma, and Akshay Kumar Rathore. "A review on barrier and challenges of electric vehicle in India and vehicle to grid optimisation". In: Transportation Engineering 4 (2021), p. 100057. ISSN: 2666-691X. DOI: https://doi.org/10.1016/j.treng.2021.100057. URL: https://www.sciencedirect.com/science/article/pii/S2666691X21000130)

[4] Vishal Singh Patyal, Ravi Kumar, and Shiksha Kushwah. "Modeling barriers to the adoption of electric vehicles: An Indian perspective". In: Energy 237 (2021), p. 121554. ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2021.121554. URL:https://. sciencedirect.com /science /article /pii /S0360544221018028.

[5] IBEF. Automobile Industry in India. https://www. ibef.org/industry/india-automobiles.aspx. [Online; accessed Nov-2021]. 2021.

[6] Som Sekhar Bhattacharyya and Shreyash Thakre. "Exploring the factors influencing electric vehicle adoption: an empirical investigation in the emerging economy context of India". In: foresight (2020).

[7] Chien-fei Chen et al. "Assessing the socio- demographic, technical, economic and behavioral factors of Nordic electric vehicle adoption and the influence of vehicle-to-grid preferences". In: Renewable and Sustainable Energy Reviews 121 (2020), p. 109692.

[8] Rupesh Kumar et al. "Addressing the challenges to electric vehicle adoption via sharing economy: an Indian perspective". In: Management of Environmental Quality: An International Journal (2020)

[9] Nele Rietmann, Beatrice Hügler, and Theo Lieven. "Forecasting the trajectory of electric vehicle sales and the consequences for worldwide CO2 emissions". In: Journal of Cleaner Production 261 (2020), p. 121038.

[10] Lo Zhenbo et al. "Optimal Driving Range for Battery Electric Vehicles Based on Modeling Users' Driving and Charging Behavior". In: Innovative Methods for Data Informed Multimodal Transport Flow and System Analysis (2020).

[11] Sid-Ali Amamra and James Marco. "Vehicle-to-grid aggregator to support power grid and reduce electric vehicle charging cost". In: IEEE Access 7 (2019), pp. 178528–178538.

[12] Riccardo Iacobucci, Benjamin McLellan, and Tetsuo Tezuka. "Optimization of shared autonomous electric vehicles operations with charge scheduling and vehicle-to-grid". In: Transportation Research Part C: Emerging Technologies 100 (2019), pp. 34–52. ISSN: 0968-090X. DOI:https://doi.org/10.1016/j.trc.2019.01.011. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18309197

[13] Hasan Mehrjerdi and Elyas Rakhshani. "Vehicle-to- grid technology for cost reduction and uncertainty management integrated with solar power". In: Journal of Cleaner Production 229 (2019), pp. 463–469.

International Journal of
Electrical and Electronics Research (IJEER)
Review Article | Volume 10, Issue 3 | Pages 447-453 | e-ISSN: 2347-470X

[14] Amit Shankar and Pooja Kumari. "Exploring the enablers and inhibitors of electric vehicle adoption intention from sellers' perspective in India: A view of the dual-factor model". In: International Journal of Nonprofit and Voluntary Sector Marketing 24.4 (2019), e1662

[15] Rachana Vidhi and Prasanna Shrivastava. "A review of electric vehicle lifecycle emissions and policy recommendations to increase EV penetration in India". In: Energies 11.3 (2018), p. 483

[16] T Donna Chen, Kara M Kockelman, and Josiah P Hanna. "Operations of a shared, autonomous, electric vehicle fleet: Implications of vehicle & charging infrastructure decisions". In: Transportation Research Part A: Policy and Practice 94 (2016), pp. 243–254.

[17] Kang Miao Tan, Vigna K Ramachandaramurthy, and Jia Ying Yong. "Integration of electric vehicles in smart grid: A review on vehicle to grid technologies and optimization techniques". In: Renewable and Sustainable Energy Reviews 53 (2016), pp. 720–732.

[18] Linni Jian et al. "Optimal scheduling for vehicle-to- grid operation with stochastic connection of plug-in electric vehicles to smart grid". In: Applied Energy 146 (2015), pp. 150–161. ISSN: 0306-2619. DOI: https: //doi.org/10.1016/j.apenergy.2015.02.030. URL: https://www.sciencedirect.com/science/article/pii/S030626191500207X.

[19] Linni Jian et al. "A scenario of vehicle-to-grid implementation and its double-layer optimal charging strategy for minimizing load variance within regional smart grids". In: Energy Conversion and Management 78 (2014), pp. 508–517. ISSN: 0196-8904. DOI: https://doi.org/10.1016/j. enconman.2013.11.007. URL: https://www.sciencedirect.com/science/article/pii/ S0196890413007280T

[20] Zhenhong Lin. "Optimizing and diversifying electric vehicle driving range for US drivers". In: Transportation Science 48.4 (2014), pp. 635–650.

[21] Ahmed Yousuf Saber and Ganesh Kumar Venayagamoorthy. "Intelligent unit commitment with vehicle-to-grid —A cost-emission optimization". In: Journal of Power Sources 195.3 (2010), pp. 898–911. ISSN: 0378-7753. DOI: https://doi.org/10.1016/j.jpowsour.2009. 08.035. URL: https://www.sciencedirect.com/science/article/pii/S037877530901341X.

[22] Ahmed Yousuf Saber and Ganesh Kumar Venayag-amoorthy. "Intelligent unit commitment with vehicle-to-grid—A cost-emission optimization". In: Journal of Power Sources 195.3 (2010), pp. 898–911.

[23] Ahmed Yousuf Saber and Ganesh Kumar Venayagamoorthy. "Optimization of vehicle-to-grid scheduling in constrained parking lots". In: 2009 IEEE Power Energy Society General Meeting. 2009, pp. 1–8. DOI: 10.1109/PES.2009.5275205.

[24] Ahmed Yousuf Saber and Ganesh Kumar Venayagamoorthy. "Unit commitment with vehicle-to-grid using particle swarm optimization". In: 2009 IEEE Bucharest PowerTech. IEEE. 2009, pp. 1–8.

© 2022 by Aditya, Rachit Bansal, Akshit Goel, N.K. Jain and Uma Nangia. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

This article has been accepted for publication in IEEE Transactions on Industrial Informatics. 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/TII.2022.3203059



# Identifying Influential Nodes for Smart Enterprises using Community structure with Integrated Feature Ranking

Sanjay Kumar, Akshi Kumar *Senior Member, IEEE* and B. S. Panda

**Abstract**—Finding influential nodes reshuffles the very notion of linear paths in business processes and replaces it with networks of business value within a smart enterprise system. There are many existing algorithms for identifying influential nodes with certain limitations for applying in large-scale networks. In this paper, we propose a community structure with an Integrated Features Ranking (CIFR) algorithm to find influential nodes in the network. Firstly, we use the community detection algorithm to find communities in the system and then we rank the nodes of network based on three factors, namely- local ranking, gateway ranking, and community ranking, collectively termed as integrated features. Our algorithm intends to select influential nodes, which are both globally and locally optimal, leading to overall high information propagation. We perform the experimental results on total eight networks using various evaluation parameters. The obtained results validate superior performance against contemporary algorithms adding value to smart enterprises.

**Index Terms**—Smart enterprises, Community structure, Business networks, Influential nodes, Node centrality

## 1 Introduction

A smart enterprise system is a built on the pillars of mobile workforce, flexible on-demand delivery, collaborative communities with trustworthy and secure services. Many enterprises have outspread their business activities to network platforms [1]. A smart enterprise must engage by involving and investing in people who are instrumental in reaching personal, group and organizational goals. Social footprints of key influencers can accelerate returns, assure high level of trust, and a wider reach within a smart enterprise design journey. The influencer pool in collaborative communities can reinvent the customer engagement models and act as the key driver for reshaping smart enterprises.The advantages of various social networks for the promotion of business and viral marketing are universally accepted primarily because of its low cost and fast information diffusion characteristics [2] [3]. Undoubtedly, social networks (SNs) play a crucial role in information propagation and user interactions building smart networks and finding influential nodes in these networks has gained much popularity among researchers because of the business value. The maximum spread of the information in the network can be achieved through the influential nodes. There are many node centrality-based and greedy-based algorithms for identifying influential nodes with certain limitations for applying in large-scale networks.

• *Sanjay Kumar is with the Department of Computer Science and Engineering, Delhi Technological University, New Delhi-110042,India. E-mail: sanjay.kumar@dtu.ac.in, Akshi Kumar is with Department of Computing and Mathematics, Faculty of Science and Engineering, Manchester Metropolitan University, Manchester, United Kingdom. E-mail: akshi.kumar@mmu.ac.uk, B. S. Panda is with Computer Science and Application Group, Department of Mathematics, Indian Institute of Technology Delhi, Hauz Khas, New Delhi-110016, India, E-mail: bspanda@maths.iitd.ac.in*

*Manuscript received March 30, 2022; revised Jul 27, 2022*

A social network is a complex and dynamic network where users produce and consume a massive amount of content creating economic outcomes. These quickly (re-)configuring links can be modeled as a graph $G = (V, E)$, where $V$ represents a set of people or entities present and $E$ denote edges. An edge connects two nodes if both the nodes have a social connection like friendship, follow-followee, coauthorship, etc. Based on the notion of word-of-mouth strategy in information diffusion and the trust between users, finding influential nodes is one of the most studied problems in the domain of network science in recent years because of its potential business value [4]. Identifying influential nodes is the task of choosing a constant number ($k$) of seed nodes with a high spreading capability such that if a piece of information originates from them, the information can reach the optimal number of vertices in the system through the diffusion cascade [5], [6]. It is a kind of optimization problem, and Kempe et al. [6] proved that getting an optimal solution is NP-hard under conventional information diffusion models. Mathematically, the problem of finding influential nodes is expressed as follows: "Given a social network $G(V, E)$, information diffusion model $M$, and a small and constant positive integer $k$, the aim is to select a subset $S \subset V$ of $k$ users as the seed set to maximize the information spread such that for any other seed subset $S^* \subset V$ of $k$ nodes, the following condition is satisfied:

$$\sigma_{G,M}(S^*) \leq \sigma_{G,M}(S) \tag{1}$$

where $\sigma_{G,M}(S)$ is the expected influence spread or numbers of nodes influenced by the nodes in set $S$."

Centrality-based methods are the most common way to rank the spreading ability of nodes to uncover influential nodes. These methods assign a rank or provide a score to nodes based on the importance of their topological position under various considerations in the network [7]. Numerous greedy-based algorithms [6], [8] are proposed to

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:48:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Industrial Informatics. 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/TII.2022.3203059



find influential nodes using a greedy strategy and discrete optimization technique. Besides, many community-based, and deep learning based influence maximization methods are proposed with considerably improved performance [9]–[11].

The presence of community structure in business networks where several nodes in a community or module are relatively densely connected and have some similar features allow picking the best capabilities from many business network actors [12]. Community structure can give rise to insight into the system and understand its function. The appearance of such community structures in a network can accelerate the information propagation in the system [13]. In this paper, we propose Community-structure with Integrated Feature Ranking (CIFR) algorithm for finding influential nodes for smart enterprises. We divide the network into various communities using the Leiden algorithm [14], then focus on the influence of nodes by considering their contribution to their local community and the nodes of other communities. After considering communities as building blocks of the network, we also rank all those communities based on their relative importance. The proposed method considers the spreading ability of a node based on the following three parameters: (a) how good a node influences the nodes within the same community, i.e., the local ranking of the node?, (b) how well a node within a community infects the nodes of other communities, i.e., gateway ranking of the node?, and (c) how well is a community based on its number of connections to other communities, i.e., community ranking? Thus, we consider a combined effect of the above-said parameters to compute the final ranking of all the nodes in the network. We term these features of the nodes as integrated features. The proposed work for finding influential nodes for smart enterprises has many augmented intelligence applications as well. One prominent application is viral marketing, where influential nodes can change the decision, opinions, and preferences of many other nodes due to mutual trust. The influential nodes can work as business enterprise leaders, making a product or idea viral and leading many adopters. This can generate high profit for the business and its teams to keep up with the rapid pace of business disruption.

Tab 1 lists the various mathematical terms used in this work.

TABLE 1: Mathematical terms used

| Notation | Description |
|---|---|
| $\sigma$ | Influence spread |
| $DC$ | Degree centrailty |
| $KS$ | K-shell centrality |
| $EV$ | Eigenvector centrality |
| $HI$ | $h$-index centrality |
| $GLR$ | Gateway local rank |
| $LR$ | Local Ranking |
| $GR$ | Global ranking |
| $CR$ | Community ranking |
| $F(t)$ | Infection scale |
| $\tau$ | Kendall tau correlation |
| $\beta$ | Infection rate |

The rest of the paper is organized as follows: Section 2 presents the related works and information diffusion models. Section 3 describe the proposed method named CIFR in detail, along with time complexity analysis. Section 4 reports the results and analysis of the proposed method, along with various popular existing methods based on multiple performance parameters. Finally, Section 5 presents the conclusion of the paper with its application in large scale networks.

## 2 Related Work

In recent years, identifying influential nodes have been studied extensively. Centrality-based algorithms are popular methods for ranking the spreading abilities of nodes. Such methods assign a score to each node by utilizing the topological structure of networks, and nodes having high score values are chosen as the seed nodes. Degree centrality, Betweenness Centrality, and Closeness Centrality, K-shell (KS) centrality [15], [16] are quite common centrality based methods for finding influential nodes. Chen et al. proposed local centrality (LC) [17] as a semi-local measure, which is an extension of degree centrality and considers the 1-hop and 2-hop neighbors into account. The notion of Eigenvector (EV) centrality [18] relies on the fact that a node is vital if it has connectivity with nodes that are considered significant in the system. PageRank [19] is a popular method to rank the web pages used by Google, which is an improved variant of Eigenvector centrality. The $h$-index (HI) centrality is a popular measure to assess the importance of a node [20]. The value of $h$-index of a node can be defined as maximum value $h$ such that it should have a minimum of $h$ neighbors with each having degree $h$ or more.

In literature, several algorithms of finding influential nodes are proposed based on the idea of combining local and semi-local features of nodes and produced good results. Berahmand et al. [21] proposed the DCL algorithm for finding influential spreaders by utilizing a combination of local measures such as the number of common neighbors, degree, and inverse clustering coefficient. The reversed node ranking (RNR) method [22] employed the ranking information and the reversed rank as the importance for calculating the influence spread of node. The reversed rank relies upon the idea that the node with a higher rank has more information spreading ability. Wen and Deng [23] introduced a technique named LID for influence maximization using local information dimensionality by considering the local structural properties around the central node. They assumed the quasilocal information and lowered the computational complexity, and further utilized the notion of Shannon entropy.

By utilizing the presence of community structures in networks, the information propagation and influence spread in the system can be enhanced [24], [25]. In recent years, many community-based influence maximization algorithms are proposed. Such algorithms assume independence between communities and perform parallel execution. Cai et al. [10] recommended a heuristic influence maximization technique, which combines community detection and topic awareness into influence diffusion modeling. Salavati et al. [26] proposed the gateway local rank (GLR) method, which improves closeness centrality by community detection and utilizing the local structure of nodes. GLR method first identifies the community, and then in each community, it obtains one best local node and one gateway node using

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:48:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Industrial Informatics. 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/TII.2022.3203059



centrality measures. Generally, the influence maximization method using community structure gives equal importance to each community irrespective of the number of nodes present in them and the number of outgoing edges to other communities.

## 3 Proposed Work

This section presents our Community structure with Integrated Feature Ranking (CIFR) algorithm for finding influential nodes in smart enterprises. A community in a business network corresponds to a group of nodes that are closely connected among themselves and sparsely connected to the rest of the network. Uncovering community structure in such networks can be utilized to assess the people's behavior in the group and anticipate their future activities. Due to the closed connection, information can rapidly spread in the community. Node with a prime position and sharing many edges with the other members in its community or module may play a vital role in shaping opinions within the group. The vertices present at the boundaries and connecting communities play a crucial role in information exchanges between different modules. Hence, a node with its connections to the nodes of other communities can act as a bridge or gateway for information propagation, and the role of such nodes becomes crucial for influence maximization. We assume each community has different importance for information propagation. Hence, we assign a ranking to each community based on the number of members and outgoing edges to other communities. We define the local importance of a node in its community as local ranking, bridging role between communities of nodes as gateway ranking, and the ranking of community in which a node is present as community ranking. The adopted definitions are as follows.

***Defintion 1:*** **Local ranking**: We define local ranking or local betweenness centrality of a node $i$ in its community ($c$) is defined as the extent to which the shortest path between a pair of nodes $j$ and $k$ in the same community ($c$) through node $i$, i.e.,

$$LR^c(i) = \sum_{i,j,k \in c \cap j \neq k} \frac{\sigma^i_{j,k}}{\sigma_{j,k}} \tag{2}$$

where $\sigma_{j,k}$ is the total number of shortest path between nodes $j$ and $k$ and $\sigma^i_{j,k}$ is the total number of such shortest path that passes through node $i$ in community $c$. Local ranking can be used to identify crucial nodes for information propagation, as usually, information is flowing through the shortest paths in the network [27]. The high value of the local ranking of a node corresponds to control over a major fraction of the information flow through that node.

***Defintion 2:*** **Gateway ranking**: We define the gateway ranking of a node $i$ lying in community $c$ is defined as the total number of neighbors, which belongs to other communities. Assume input graph is $G(V, E)$ then gateway ranking can be computed using the following steps:

$$G_2 = (V, E \setminus E'), \tag{3}$$

$$GR(i) = DC(i), i \in G_2(V) \tag{4}$$

where $G_2$ is a graph obtained from input graph $G$ after removing all edges ($E'$) in the same community. Here, the new graph $G_2$ contains only those edges which connect nodes of different communities or inter-community edges. The node of this graph is a kind of gateway node which is connected to the nodes of other communities as well. Hence, $GR(i)$ is the gateway ranking of node $i$ that is equal to the degree centrality of node $i$, i.e., $DC(i)$, in the new arrangement. The high value of gateway ranking of a node signifies its importance in influence maximization as it can share the information among nodes of different communities. Hence, local ranking is used to rank the nodes based on the information propagation capability within their community. However, gateway ranking is used to rank the nodes based on the information propagation capability outside the community. Therefore, local ranking is a kind of local measure, and gateway ranking is a type of global measure to assess the importance of a node for information spraed.

***Defintion 3:*** **Community ranking**: we define community ranking that assigns a score to each node based on the significance of the community to which it belongs. To understand the essence of community ranking, we create a multi-graph ($G_3$) from the input graph $G(V, E)$ as follows:

$$G_3 = (V_c, E_c) \tag{5}$$

where $G_3 = (V_c, E_c)$ is a multi-graph with $V_c$ is the set of communities, and $E_c$ is the edge between these communities. Basically, each node of the graph, $G_3$, is a community, and each edge connects nodes of two different communities. Hence, there can be a situation where multiple nodes of the same community ($c_i$) connect to one or more nodes of other communities ($c_j$), and $G_3$ is a multi-graph. Now, community ranking of a node $i$ in the community $c$ is defined as the degree of the community $c$ where $c$ is a vertex in the multi-graph $G_3$ created as per Eq. (7), i.e.,

$$CR(i) = DC(i), i \in c \text{ and } c \in G_3(V) \tag{6}$$

Hence, the community ranking of a node $i$ lying in community $c$ is equal to the number of outgoing connections to other communities from the vertices of community $c$. Here, all the nodes belonging to a particular community $c$ gets the same value of community ranking. The high value of community ranking of a node implies that the node belongs to a community with more connections to the members of other communities, and the node can play a vital role for influencing other nodes. Since the size of each community may vary in real-life situations, and a core node chosen from a community having greater size can contribute more influence rather than a core node selected from a small community.

We term local ranking, gateway ranking, and community ranking features of nodes as an integrated features. Therefore by considering three factors, namely, the influence and importance of every node in its community, the significance of the node as per inter-community connections or gateway capabilities, and the relative importance and influence of each community, we propose Community structure with Integrated Features Ranking (CIFR) method for finding influential nodes in smart enterprises. The outline of the proposed algorithm is presented in Algorithm 1.

The detailed steps of the proposed algorithms are as follows:

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:48:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Industrial Informatics. 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/TII.2022.3203059



**Algorithm 1** : CIFR

**Input:** $G = (V, E)$ with $\|V\| = n, \|E\| = m$
**Output:** $\mathcal{R}$ : Ranked nodes
1: $C \leftarrow$ Detecting Communities;
2: **for** each $c \in C$ **do**
3: $LR_i^c \leftarrow$ compute local ranking for each node $i$ in $c$ using Eq. (2)
4: $LR_i^c \leftarrow LR_i^c \times A^c$, where $A^c$ is a normalization parameter computed using Eq. (9)
5: **end for**
6: $l_1 \leftarrow$ $LR$ value of each node as obtained in step $4$.
7: $l_1 \leftarrow$ sort the $l_1$ in descending order of $LR$ values and assign rank 1, 2, and so on.
8: **for** each node $v \in V$ **do**
9: compute $GR(v)$, i.e., gateway ranking score using Eq. (4)
10: **end for**
11: $l_2 \leftarrow$ gateway ranking ($GR$) of all nodes
12: $l_2 \leftarrow$ sort the $l_2$ in descending order of $GR$ values and assign rank 1, 2, and so on.
13: **for** each $c \in C$ **do**
14: compute $CR(c)$, and assign community ranking score to each node in $c$ using Eq. (6)
15: **end for**
16: $l_3 \leftarrow$ community ranking ($CR$) of all nodes.
17: $l_3 \leftarrow$ sort the $l_3$ in descending order ofcommunity ranking ($CR$) values and assign rank 1, 2, and so on.
18: **for** each node $v \in V$ **do**
19: compute $CIFR(v)$ score using Eq. (8)
20: **end for**
21: $\mathcal{R} = sort(CIFR \text{ score})$
22: **return** $\mathcal{R}$

(i) Step 1: We use an efficient Lieden Algorithm to find the community in large-scale networks, which improves the classical Louvain algorithm and guarantees that communities are well connected. There exist numerous community detection methods. However, the recently developed Lieden method only visits those nodes whose neighborhood has changed [14]. By relying on a fast local move approach, the Leiden algorithm runs in linear time and is a faster algorithm. We can execute community detection process in parallel using Lieden method enabling overall processing faster. the Leiden algorithm can be used to find communities in the network with overlapping communities.

(ii) Step 2: After finding the communities, we consider the communities are separated from each other. Now, within each community, we calculate the local ranking ($LR$) of each node using Eq. (2), as per line number 3 of the Algorithm 1. This step can be done in parallel in all communities obtained.

(iii) Step 3: Since the size of each community may vary in real-life situations, and a core node chosen from a community having greater size can contribute more influence rather than a core node selected from a small community. Hence, we need a proper normalization parameter to address the difference in the size of communities. We introduce a normalization parameter, $A^c$, for each community $c \in C$ as follows:

$$A^c = \frac{n_c}{n} \tag{7}$$

where $n_c$ is the number of nodes in community $c$ and $n$ is the total number of nodes in the network. Within each community, the value of local betweenness centrality of each node is multiplied by the parameter $A^c$, as per line number $4$ of the algorithm. This may normalize the effect of the size of the community with the influence of a node in that community.

(iv) Step $4$: This step explains line number $6$ and $7$ of the proposed algorithm. We make a dictionary $L_1$ that contains the final local ranking score of each node of the network after doing the normalization, as obtained in the previous step $3$. Sort the dictionary $L_1$ based on values in descending order, and assign the value: $1$, $2$, $3$, and so on, where $1$ is the new rank value given to the best node, $2$ corresponds to a second-best node and so on according to the measure. Till now, we prioritize the nodes of the network based on its local importance for the influence maximization by taking care of the normalization needed to handle the difference in the size of various communities.

(v) Step $5$: In this step, which covers line number $8$ to $12$ of the proposed Algorithm $1$, we want to find the global importance of nodes for information propagation based on their bridging capability. A node with its attachments to the vertices of other communities can serve a bridge for information propagation, and the role of such nodes becomes vital for the information spread in the network. We compute the gateway ranking ($GR$) of each node of the input graph $G(V, E)$ using Eq. (4). Now, we use the dictionary $L_2$ to contain the nodes and their gateway ranking score. Here, all non-gateway node has a gateway rank value as $0$. Similar to step $4$, we sort the dictionary $L_2$ based on values in descending

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:48:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Industrial Informatics. 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/TII.2022.3203059



order and assign the values: 1, 2, 3, and so on, where 1 is the new rank value given to the best node, 2 corresponds to a second-best node and so on. In this arrangement, the nodes which are not gateway nodes, they each get a score as zero, and they get last or maximum rank. Hence, in this step, we prioritize the nodes of the network based on its bridging role or global importance for the information exchange.

(vi) Step 6: In this step, we rank the nodes based on the number of inter-community connections of the community in which they belong. This step discusses the line number 13 to 17 of the proposed Algorithm 1. We compute the community ranking ($CR$) of each community $c$ using Eq. (6), and assign the same score to all the nodes in that particular community. We utilize the dictionary $L_3$ to contain the nodes and their community score.
Similar to steps 4 and 5, we sort the dictionary $L_3$ based on values in descending order, and assign the values: 1, 2, 3, and so on, where 1 is the new rank value given to the best node according to this measure, 2 corresponds to a second-best node, and the last rank is given to the least important nodes. The nodes present in the same community get the same rank in this arrangement. Here, the notion of $L_3$ ranking signifies, if a node is present in a higher-ranked community, it should get higher priority by this measure.

(vii) Step 7: In the above three measures, namely $L_1$, $L_2$, and $L_3$, we follow the same strategy to handle the case when there are ties, i.e., ranking of nodes having the same score in a particular measure. For example, if there are four nodes, if three of them have the same score, which is higher than the fourth node, e.g., 5,5,4,2, then we rank them as 1,1,3,4. Here, all the nodes having the same score, they each get the same rank, and the next rank is skipped.
Finally, based on three different measures for prioritizing the nodes namely, $L_1$, $L_2$ and $L_3$, we introduce $CIFR$ score of each node ($v$) as:

$$CIFR(v) = \frac{1}{L_1(v) + L_2(v) + L_3(v)}, v \in V \quad (8)$$

Here, we assign equal weightage to all three measures, namely, local ranking, global ranking, and community ranking of a node, to compute its CIFR score as per Eq.(8). The reason for the same weightage assignment is that all three measures are equally crucial to identifying influential nodes, and there should be no preference for one measure over others. Eventually, we select top-$k$ nodes in this raking as the source influential nodes.

### 3.1 Time complexity analysis:

We discuss the time complexity of the proposed algorithm based on the pseudo-code of the Algorithm 1. Assume the number of nodes in the network is $n$, and the number of edges is $m$. Community detection using the Lieden method in line number 1 can be performed in parallel and takes $O(m + n)$ time. Line number 2 to 5 for computing the local ranking for each node from each community requires $O(m + n)$ time. Line number 6 takes $O(n)$ for maintaining a dictionary containing the $LR$ value of each node, and line 7 uses $O(n \times logn)$ time for sorting activity and prioritizing the nodes. The line numbers from 8 to 10 for computing gateway ranking of all the nodes demands $O(m + n)$ time. The line numbers 11 and 12 for maintaining a dictionary containing the gateway ranking ($GR$), and then sorting and assigning ranks to all nodes requires $O(n \times logn)$ time. Similarly, line numbers 13 to 15 for computing community ranking requires $O(n)$ times followed by maintaining a dictionary containing the $CR$, and then sorting and assigning ranks to all nodes requires $O(nlogn)$ time for line numbers 16 and 17. To compute the $CIFR$ score for each node needs $O(n)$ time in the loop of the lines 18 to 20. Finally, the ranking of nodes based on the $CIFR$ score can be done in $O(n \times logn)$ time.

Hence, the overall complexity of the proposed Algorithm is: $O(m + n + n \times logn)$. Since most of the real-life graph is sparse, we can also consider $O(m)$= $O(n)$. Hence, the time complexity of the proposed algorithm is $O(n \times logn)$.

## 4 Experimental Results and Analysis

To exhibit the effectiveness of the proposed algorithm, we conducted investigations on seven real-life networks, namely PGP, Hep-Th, Cond-Mat, Gr-Qc, Gnutella08, Gnutella09, Facebook social, and one synthetic network, LFR dataset. These datasets are of diverse size, and application, and are listed in Tab. 2 where $< k >$, $< c >$, and $d$ denotes the average degree of the nodes, average clustering coefficients, and diameter, i.e., longest shortest path of the network respectively. We compared the performance of the proposed Community-structure with Integrated Feature Ranking (CIFR) algorithm with the well-known approaches like local centrality (LC) [17], K-shell (KS) centrality [15], Eigenvector (EV) centrality [18], gateway local rank (GLR) [26] and local information dimensionality (LID) [23]. The Lieden method for performing community detection in the proposed algorithm has been exectuted in parallel.

The performance of all the methods are compared using three evaluation criteria, namely the infection scale ($F(t)$) originating from source nodes, the number of influenced nodes vs. the number of source influential nodes, and Kendall tau ($\tau$) correlation. To compute the CIFR score of each node Eq. No. 8 is used, and top $k$-nodes are chosen as the seed influential nodes. We employed two popular the information diffusion model, namely, Susceptible-infected-recovered (SIR) [33] and independent cascade (IC) model to calculate scale of influence achieved by selected influential nodes using different considered algorithms. SIR model is a well-studied stochastic epidemic-based information propagation model to investigate the performance of the influence maximization algorithm. SIR model splits all nodes into three categories, namely, Susceptible ($S$), Infected ($I$), and Recovered ($R$). This model takes three inputs: initial spreaders, infection rate ($\beta$), and recovery rate ($\gamma$). The independent cascade (IC) model is a classical information diffusion paradigm where information spreads in the network through a cascade originating from some seed nodes. Every node can be in one of the two states-

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:48:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Industrial Informatics. 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/TII.2022.3203059



TABLE 2: The basic statistical features of used network datasets

| Dataset | Type of Network | #nodes | #edges | $< k >$ | $< c >$ | d |
|---|---|---|---|---|---|---|
| PGP [28] | Trust network | 10638 | 24301 | 4.57 | 0.023 | 10 |
| Hep-Th [29] | Collaboration network | 9877 | 25998 | 5.2 | 0.47 | 17 |
| Cond-Mat [29] | Collaboration network | 23133 | 93439 | 8.08 | 0.63 | 14 |
| Gr-Qc [29] | Collaboration network | 5242 | 14496 | 5.52 | 0.529 | 17 |
| Gnutella08 [30] | File-sharing network | 6301 | 20777 | 6.59 | 0.01 | 9 |
| Gnutella09 [30] | File-sharing network | 8114 | 26013 | 6.40 | 0.009 | 10 |
| Facebook Social [31] | Social network | 4039 | 88234 | 43.7 | 0.60 | 8 |
| LFR-0.2 [32] | Synthetic network | 10000 | 64936 | 12.987 | 0.130 | 16 |

active or inactive. The active nodes refer to those nodes that got the information, whereas inactive nodes have not yet received the information. The spread of infection scale and the final number of influenced nodes at the end of the spreading process depend on the number of seed nodes and the infection rate ($\beta$) in the SIR model and probability $P_{uv}$ in IC model. As the selection of susceptible neighbors for the infection in the SIR and IC model is random, we run the information diffusion model 200 times, and results are averaged over. The experiments have been performed on a personal computer with primary memory 8 GB, and 1.6 GHz Intel Core i5 processor. We performed the simulations of our experiments using Python programming language using various packages like Networkx, Scikit-Learn, matplotlib, panda, etc. The details of all four evaluations are as follows.

### 4.1 Infection Scale:

The infection scale ($F(t)$) at any time $t$ is the sum of the number of infected nodes ($n_{I(t)}$) at time $t$, and the number of recovered nodes ($n_{R(t)}$) till time $t$. The following equation computes the Infection scale, $F(t)$:

$$F(t) = n_{I(t)} + n_{R(t)} \tag{9}$$

To estimate the infection scale ($F(t)$) in terms of the increasing timestamp of the information diffusion process, we run the SIR model using the initial number of influential spreaders, and infection rate ($\beta$). We considered different numbers of influential spreaders as the seed nodes for different datasets based on their size. Here, influential spreaders are such nodes that are identified by the various considered algorithms. We considered the top $2\%$, and $5\%$ nodes as the source spreader nodes depending upon the size of the network. For a relatively large dataset having $10,000$ nodes or more like Cond-Mat, PGP, and LFR top $2\%$ nodes are used as spreader nodes. However, for the remaining datasets, the top $5\%$ nodes are taken as source nodes. The value of the epidemic threshold, i.e., $\beta_{th}$ are different for different datasets, as mentioned in Tab. 2. To maintain consistency in our evaluation process, we use the same value of $\beta$ as $0.08$ for all the datasets, i.e., when a node is in the infected state, then it can influence $8\%$ of its neighbors randomly. Fig. 1 depicts the infection scale ($F(t)$) with respect to the time for different datasets considered. It is noticeable from the result that in most of the datasets, the proposed algorithm (CIFR), outperforms all approaches, namely, local centrality, K-shell, $h$-index, Eigenvector, GLR, and LID. LID method almost remains in the second position after CIFR. In the case of Facebook and LFR datasets, which

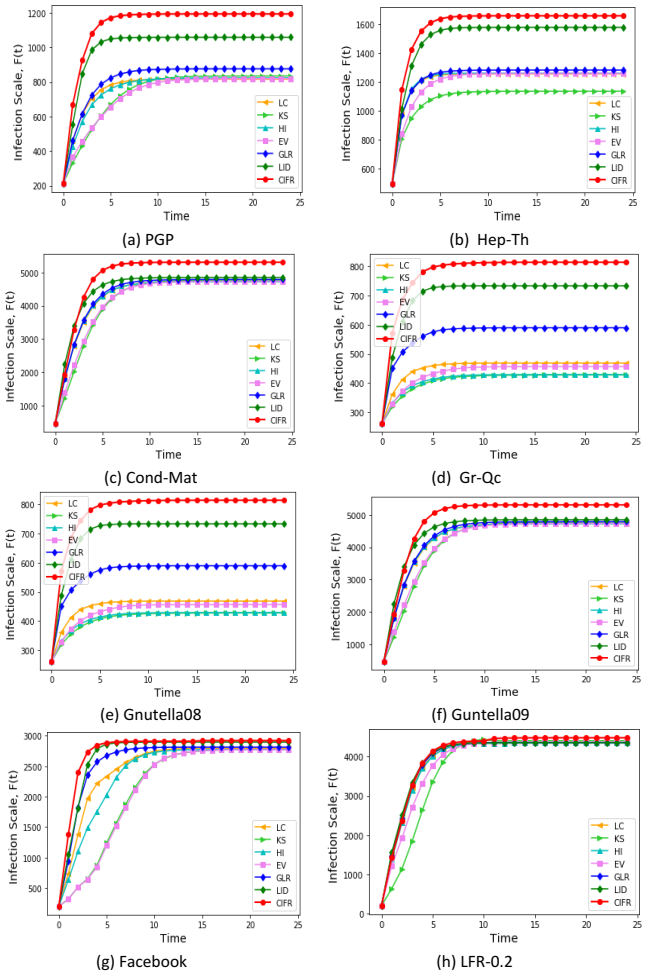



Fig. 1: (a)-(h): Infection scale ($F(t)$) Vs. time using the same rate of infection ($\beta$) $= 0.08$ for all networks. The number of initial spreaders is taken as $2\%$ for PGP, Cond-Mat, LFR-0.2 datasets, and $5\%$ for Hep-Th, Gr-Qc, Facebook, Gnutella08, and Guntella09 datasets. The results are obtained over 200 independent simulations of the SIR model

are relatively dense datasets, the performances of all methods very similar. However, CIFR remains at the top. The

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:48:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Industrial Informatics. 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/TII.2022.3203059



obtained result depicts the superiority of the seed selection by our algorithm.

### 4.2 Influence spread Vs. number of seed influential nodes:

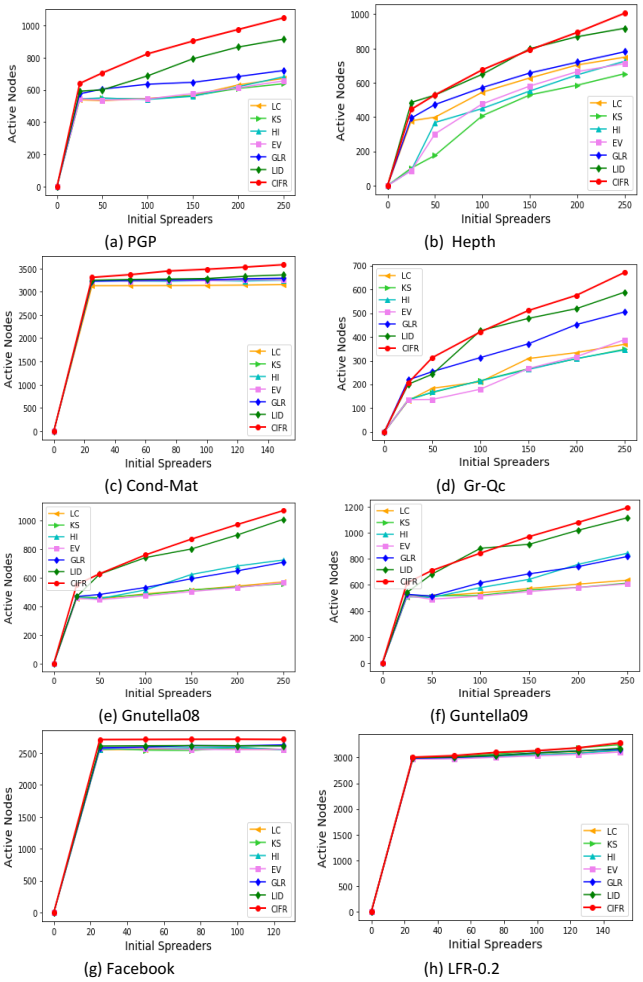



Fig. 2: (a)-(h): Number of influenced nodes at the end of information diffusion process with respect to different number of source influential nodes using the SIR model with infection rate ($\beta$) $= 0.08$ for all the datsets. The results are obtained from 200 independent simulations of SIR model

.

The influence spread refers to the extent or fraction of the network to which information originating from the set of seed nodes has finally spread at the end of the information diffusion process. To compute the number of influence spread at the end of the spreading process vs. number of seed influential nodes, we employed both the information spreading model SIR and IC for each dataset. We considered the different number of source nodes for various networks. For the datasets having more than $50,000$ edges like Cond-Mat, Facebook, and LFR-$0.2$, the number of source nodes were taken in the range of spreaders are taken in the range of $20$, $40$, $60$, $80$, $100$, and $120$. For remaining datasets with less then than $50,000$ edges ,i.e., PGP, Hep-Th, Gr-Qc, Gnutella08,and Gnutella09, we take the number of spreaders as $50$, $100$, $150$,$200$, and $250$. This variation in the considered source spreader nodes to plot influence spread for all the methods is because of the difference in the size of networks and to obtain the results in relatively less time.

Fig. 2 presents the effect of change in the number of initial spreaders, on the number of total influenced nodes using the SIR model with infection rate ($\beta$) as $0.08$ for all the datasets. The received results are averaged over $100$ simulations of the SIR model. For all datasets, we keep the same value of $\beta$ in the SIR model, to maintain consistency in the evaluation. From the results, it is clear that the proposed algorithm performs better than all other methods. From the results, it is evident that the proposed algorithm performs better than all other methods in the case of PGP, Hep-Th, Gr-Qc, Cond-Mat, Gnutella08, and Gnutella09 datasets. However, in the case of relatively dense datasets like Facebook and LFR-$0.2$ datasets, the performances of various methods are very close to each other, but CIFR manages to remains at the top. Similarly, Fig. 3 displays the effect of change in the number of seed influential nodes, on overall influence spread using the IC model with $p_{uv} = 0.12$. The received results are averaged over $100$ simulations IC model. For all datasets, we keep the same value of $p_{uv}$ in the IC model to maintain consistency in the evaluation. The results show that the proposed algorithm performs better than all other methods on the datasets PGP, Hep-Th, and Gr-Qc. However, in the case Cond-Mat, Gnutella08, Gnutella09, and LFR-$0.2$ datasets, the performances of CIFR and LID methods are very close to each other with CIFR is performing slightly better then LID and beating remaining all other techniques. From both the Fig. 2 and 3, it is visible that the proposed algorithm performed consistently well in all datasets. Hence, we can infer that the selection of influential nodes using the proposed CIFR algorithm maintains its performance with an increase in the number of source influential nodes.

### 4.3 Kendall Tau ($\tau$) Matrix:

We used Kendall's tau coefficient ($\tau$) [34] to assess the correctness of the ranking methods. This evaluation metric examines the correlation between the two rank lists, namely $X$ and $Y$. Here, $X$ is the value of the centrality score of nodes under various methods, and $Y$ is the actual value of influence spread for all nodes. If node $v_i$ appears before node $v_j$ in list $X$ and $Y$ both, then such pair are called concordant pair ($n_c$), otherwise discordant pairs ($n_d$). When $n_c > n_d$, the coefficient has a positive value, indicating similarity. Kendall's tau coefficient value ranges between $(-1, 1)$. The following formula represents Kendall's tau correlation values.

$$\tau = \frac{n_c - n_d}{\frac{n(n-1)}{2}} \tag{10}$$

where $n$ is the total number of vertices in each list. We computed Kendall's tau ($\tau$) correlation of the ranked nodes iden-

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:48:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Industrial Informatics. 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/TII.2022.3203059



TABLE 3: Kendall's Tau ($\tau$) corelation values of different real-life network using SIR information diffusion model and infection rate ($\beta$) $= 0.1$

| Dataset | $\tau(LC,\sigma)$ | $\tau(KS,\sigma)$ | $\tau(HI,\sigma)$ | $\tau(EV,\sigma)$ | $\tau(LID,\sigma)$ | $\tau(GLR,\sigma)$ | $\tau(CIFR,\sigma)$ |
|---|---|---|---|---|---|---|---|
| PGP | 0.059 | 0.014 | 0.035 | 0.050 | 0.031 | 0.014 | **0.138** |
| Hep-Th | 0.318 | 0.142 | 0.029 | 0.045 | 0.167 | 0.232 | **0.376** |
| CondMat | 0.136 | 0.149 | 0.185 | 0.258 | 0.444 | 0.553 | **0.581** |
| Gr-Qc | 0.294 | 342 | 0.348 | 0.520 | **0.591** | 0.197 | 0.455 |
| Gnutella08 | 0.120 | 0.334 | 0.502 | 0.500 | 0.363 | 0.305 | **0.511** |
| Gnutella09 | 0.492 | 0.241 | 0.40 | 0.552 | 0.363 | 0.541 | **0.569** |
| Facebook | **0.127** | 0.052 | 0.038 | 0.021 | 0.009 | 0.023 | 0.033 |
| LFR-0.2 | 0.160 | 0.203 | 0.020 | 0.285 | 0.340 | **0.514** | 0.453 |

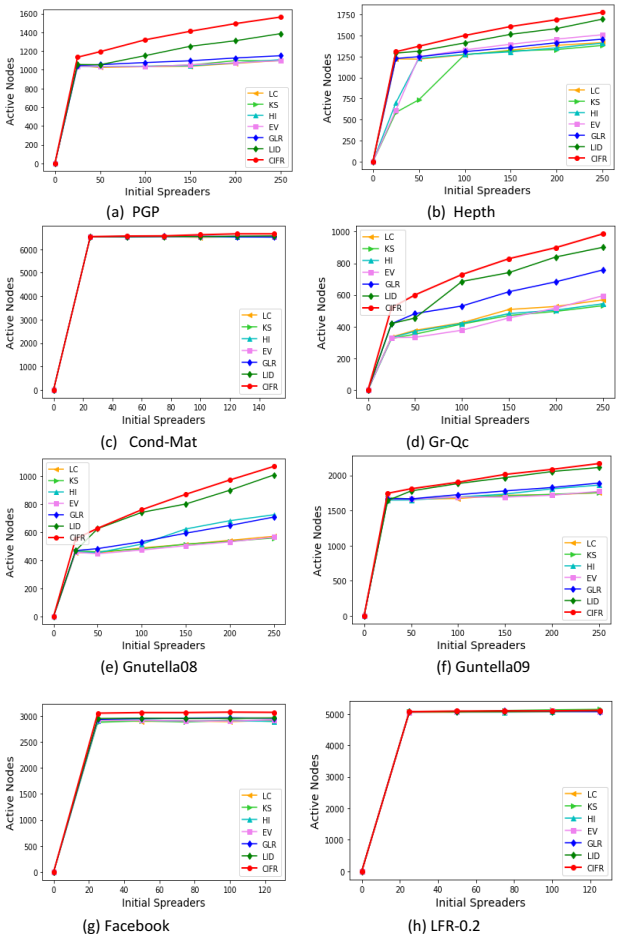



Fig. 3: (a)-(h): Number of influenced nodes at the end of information diffusion process with respect to different number of seed influential nodes using the IC model with $P_{uv} = 0.12$ for all the datasets. The results are obtained from 200 independent simulations of IC model

tified by different methods, namely local centrality (LC), K-shell, $h$-index, Eigenvector, gateway local rank (GLR), local information dimensionality (LID) and proposed algorithm (CIFR). For this calculation, we applied the SIR spreading model. To maintain consistency in the various performance measures, the value of the infection rate ($\beta$) is taken as $0.01$ for all the networks. Tab. 3 presents the values of $\tau$ obtained using all the algorithms. On the PGP, Hep-Th, Cond-Mat, Gnutella08, and Gnutella09 datasets, proposed algorithm produces the best $\tau$ values among all the approaches. However, local centrality produced the best result on Facebook, LID attains the best reslut on Gr-Qc and GLR obtains the best value on the LFR-0.2 dataset. Hence, on a majority of the networks, the proposed CIFR method produced best result.

## 5 Conclusion

Influential nodes are the nodes having high information spreading ability, and influence maximization can be achieved through these nodes in business networks of smart enterprises. In this paper, we proposed the Community structure with the Integrated Features Ranking (CIFR) algorithm for detecting influential nodes by utilizing the feature of community spread in the network. Initially, we divided the network into various communities for each community. We ranked all the nodes according to their local and gateway influences. Now, considering those communities as building blocks of the whole network, we rank all those communities. Hence, the importance of a node is weighed based on its influence in its community and community-specific global indicators like inter-connection, bridging roles, and relative importance of communities. Our algorithm utilizes the community spread selects top spreaders, which are both globally and locally optimal, leading to a high information propagation in their community and other connected communities as well. Our method, CIFR, is applicable in large-scale business networks that is for large as well as small-scale enterprises to locate the most desirable seed nodes with the high spreading capability, and it has more economical time complexity. We performed the experimental results on eight networks using various evaluation parameters. The proposed algorithm performed better against contemporary algorithms based on evaluation criteria like infection scale, influence spread, and Kendall tau correlation. The obtained results suggest the efficacy of the work and hence add value to smart enterprises. The proposed work may give rise to relatively inadequate performance if the underlying business network has unclear community structure. The future extension of the current

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:48:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Industrial Informatics. 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/TII.2022.3203059



work can accommodate more features independent of the network and community structure to identify influential nodes in the multi-layer business network.

## REFERENCES


[1] Y. C. H. Z. Zhang, Yongping and F. Tao, "Long/short-term preference based dynamic pricing and manufacturing service collaboration optimization," *IEEE Transactions on Industrial Informatics*, 2022.

[2] A. Kumar and A. Jaiswal, "A deep swarm-optimized model for leveraging industrial data analytics in cognitive manufacturing," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2938–2946, 2020.

[3] J. Tang, X. Tang, and J. Yuan, "Profit maximization for viral marketing in online social networks: Algorithms and analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 6, pp. 1095–1108, 2017.

[4] Y. Li, J. Fan, Y. Wang, and K.-L. Tan, "Influence maximization on social graphs: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 10, pp. 1852–1872, 2018.

[5] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 1029–1038.

[6] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 137–146.

[7] P. Basaras, G. Iosifidis, D. Katsaros, and L. Tassiulas, "Identifying influential spreaders in complex multilayer networks: A centrality perspective," *IEEE Transactions on Network Science and Engineering*, vol. 6, no. 1, pp. 31–45, 2017.

[8] G. Song, Y. Li, X. Chen, X. He, and J. Tang, "Influential node tracking on dynamic social network: An interchange greedy approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 2, pp. 359–372, 2016.

[9] Y.-C. Chen, W.-Y. Zhu, W.-C. Peng, W.-C. Lee, and S.-Y. Lee, "Cim: Community-based influence maximization in social networks," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 2, pp. 1–31, 2014.

[10] F. Cai, L. Qiu, X. Kuai, and H. Zhao, "Cbim-rsrw: An community-based method for influence maximization in social network," *IEEE Access*, vol. 7, pp. 152 115–152 125, 2019.

[11] S. Kumar, A. Mallik, A. Khetarpal, and B. Panda, "Influence maximization in social networks using graph embedding and graph neural network," *Information Sciences*, 2022.

[12] W. Luo, D. Zhang, L. Ni, and N. Lu, "Multiscale local community detection in social networks," *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[13] Z. Liu and B. Hu, "Epidemic spreading in community networks," *EPL (Europhysics Letters)*, vol. 72, no. 2, p. 315, 2005.

[14] V. A. Traag, L. Waltman, and N. J. Van Eck, "From louvain to leiden: guaranteeing well-connected communities," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[15] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature physics*, vol. 6, no. 11, pp. 888–893, 2010.

[16] G. Sabidussi, "The centrality index of a graph," *Psychometrika*, vol. 31, no. 4, pp. 581–603, 1966.

[17] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," *Physica a: Statistical mechanics and its applications*, vol. 391, no. 4, pp. 1777–1787, 2012.

[18] A. N. Langville and C. D. Meyer, "A survey of eigenvector methods for web information retrieval," *SIAM review*, vol. 47, no. 1, pp. 135–161, 2005.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[20] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National academy of Sciences*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[21] K. Berahmand, A. Bouyer, and N. Samadi, "A new local and multidimensional ranking measure to detect spreaders in social networks," *Computing*, vol. 101, no. 11, pp. 1711–1733, 2019.

[22] X. Rui, F. Meng, Z. Wang, and G. Yuan, "A reversed node ranking approach for influence maximization in social networks," *Applied Intelligence*, vol. 49, no. 7, pp. 2684–2698, 2019.

[23] T. Wen and Y. Deng, "Identification of influencers in complex networks by local information dimensionality," *Information Sciences*, vol. 512, pp. 549–562, 2020.

[24] H. Huang, H. Shen, Z. Meng, H. Chang, and H. He, "Community-based influence maximization for viral marketing," *Applied Intelligence*, vol. 49, no. 6, pp. 2137–2150, 2019.

[25] S. Kumar, A. Gupta, and I. Khatri, "Csr: A community based spreaders ranking algorithm for influence maximization in social networks," *World Wide Web*, pp. 1–20, 2022.

[26] C. Salavati, A. Abdollahpouri, and Z. Manbari, "Ranking nodes in complex networks based on local structure and improving closeness centrality," *Neurocomputing*, vol. 336, pp. 36–45, 2019.

[27] J. S. Mitchell, "Shortest paths and networks," in *Handbook of discrete and computational geometry*. Chapman and Hall/CRC, 2017, pp. 811–848.

[28] M. A. Serrano, M. Boguná, R. Pastor-Satorras, and A. Vespignani, "Correlations in complex networks," *Large scale structure and dynamics of complex networks: From information technology to finance and natural sciences*, pp. 35–66, 2007.

[29] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 2–es, 2007.

[30] M. Ripeanu and I. Foster, "Mapping the gnutella network: Macroscopic properties of large-scale peer-to-peer systems," in *international workshop on peer-to-peer systems*. Springer, 2002, pp. 85–93.

[31] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in neural information processing systems*, 2012, pp. 539–547.

[32] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Physical review E*, vol. 78, no. 4, p. 046110, 2008.

[33] C. Xia, Z. Liu, Z. Chen, S. Sun, and Z. Yuan, "Epidemic spreading behavior in local-world evolving networks," *Progress in Natural Science*, vol. 18, no. 6, pp. 763–768, 2008.

[34] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

**Sanjay Kumar** is currently an Assistant Professor in the Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India. He did Ph.D. degree in Computer Applications, Indian Institute of Technology (IIT) Delhi. He has completed M.Tech in Computer Application from Indian Institute of Technology (IIT) Delhi, India. Previously, he has worked with National Informatics Centre, Govt. of India as Scientist-B. His research interests include AI, Social Network Analysis, Graph Algorithm, Machine Learning, and Design and Analysis of Algorithms.

**Akshi Kumar,** SMIEEE, is a Post-doc from Federal Institute of Education, Science and Technology of Ceará, Fortaleza, Brazil and PhD from Faculty of Technology, University of Delhi, New Delhi, India. She is currently a Senior Lecturer-Data Science in the Department of Computing & Mathematics, Faculty of Science & Engineering, Manchester Metropolitan University, United Kingdom. She has worked as an Associate Professor at the Department of Information Technology, Netaji Subhas University of Technology (NSUT), New Delhi, India and as an Assistant Professor at the Delhi Technological University (DTU), formerly, Delhi College of Engineering, New Delhi, India.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:48:53 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Industrial Informatics. 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/TII.2022.3203059

**B. S. Panda** is currently a Professor in Computer Science and Application group of the Department of Mathematics, Indian Institute of Technology (IIT) Delhi, India. He did his Ph.D. in the Department of Mathematics, IIT Kanpur in 1994 in Algorithmic Graph Theory. He has served as an Assistant Professor in the Department of Computer and Information Sciences, University of Hyderabad. He was also in the Department of Computer Science and Engineering, University of Texas at Arlington, USA as a visiting researcher. His research interests are Graph Algorithms, Social Network Analysis, and Design and Analysis of Algorithms.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:48:53 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/363672115

# Impact Analysis of Institutional Credit on Agricultural Output in India

**Article** *in* Indian Journal of Economics and Development · September 2022

DOI: 10.35716/IJED/22213

CITATIONS
0

READS
17

**3 authors**, including:

Ghanshyam Pandey
SRM University-AP
**28** PUBLICATIONS **91** CITATIONS
SEE PROFILE

Seema Kumari
University of Delhi
**4** PUBLICATIONS **10** CITATIONS
SEE PROFILE

**Some of the authors of this publication are also working on these related projects:**




Fallow land in India View project

All content following this page was uploaded by Ghanshyam Pandey on 21 September 2022.

The user has requested enhancement of the downloaded file.

*Indian Journal of Economics and Development*
*Volume 18 No. 3, 2022, 700-706*
*DOI: https://doi.org/10.35716/IJED/22213*
*Indexed in ESCI (Clarivate Analytics: WoS)*

**Manuscript No. MS-22213**
**NAAS Score: 5.15**
**Indexed in Scopus**

# Impact Analysis of Institutional Credit on Agricultural Output in India

*Ghanshyam Pandey[1*], Jagriti Jaiswal[2], and Seema Kumari[3]*

*[1]Assistant Professor, and [2]Post Doc Fellow (NABARD), Department of Economics, SLASS, SRM University, Neerukonda, Mangalagiri Mandal, Mangalagiri-522240 Guntur District, (Andhra Pradesh), and [3]Assistant Professor (Guest) Department of Humanities, Delhi Technological University, Shahbad Daulatpur Village, Rohini, New Delhi-110042*

*[*]Corresponding author's email: ghanshyampndy@yahoo.com*

*Received: June 13, 2022* *Revision Submitted: August 23, 2022* *Revision Accepted: September 10, 2022*

**ABSTRACT**

This paper examined the major contribution of institutional credit for agriculture development after the green-revolution period in India. The total institutional credit for agriculture registered a significant growth rate at 8.74 per cent. However, credit from RRBs grew at the highest growth rate at around 14 per cent in the last 50 years. The results showed that the share of Scheduled Commercial Banks had emerged as the dominant source of agricultural credit in India. In contrast, the share of commercial banks in the total credit declined over the period. The result of the regression model also showed that the overall institutional credit for agricultural and other institutional sources of agricultural credit (cooperatives, RRBs, and SCBs) was statistically significant for the overall agricultural output growth in India.

**Keywords**
Agriculture, agricultural growth, institutional credit, workforce.

**JEL Codes**
Q10, Q12, Q14.

## INTRODUCTION

Capital accumulation is essential in practically all growth and development models because it enhances the sectors' productive capability. The rate of investment, which is determined by the rate of savings, determines the rate of capital accumulation. Financial institutions have a prominent role in mobilizing savings and channelise those funds into productive economic activities (Sidhu & Gill, 2006). Agricultural credits are essential in making production more efficient and productive, especially in developing countries (Bahsi & Cetin, 2020). It is an integral part of the process of commercialization of agricultural modernization and the rural economy for alleviation of poverty (Pandey, 2016). Easy and cheap credit represents the fastest way to improve agricultural production. In 1954, the All-India Survey stressed the need for the financial institution to emerge as active contributors to the country's agricultural development. Nevertheless, there is still a perpetual gap between the demand and supply of agriculture credit. To fill this gap, the government nationalized fourteen leading Indian commercial banks in 1969 and tempted them to enter agriculture credit disbursement (Nasar, 2015).

After post-reforms, India has made significant progress in physical infrastructure (roads, electricity, banking/credit, and markets) and communication and information networks and towards dismantling the regulatory barriers to inter-regional trade and free flow of resources, technologies, and information (Pandey & Kumari, 2021a; Hazrana et al., 2019; Pandey, 2016). This reform kept on the economic requirement, and in 1991, the government emphasized the implementation of potential norms. It included asset classification and capital risk-weighted asset ratio, which were supposed to compel the commercial bank to concentrate on financial efficiency and economic viability through rationalising

*Copyright ©2022 The Society of Economics and Development, except certain content provided by third parties.*

their operating system (Karmakar, 2008).

Along with other inputs, credit is essential for establishing sustainable farming systems. It is also an important tool to use their working capital, fixed capital, and consumer goods (Siddiqui & Bauch, 2009). Institutional credit has been conceived to play a pivotal role in the agricultural development of India (Kumar et al., 2010). The availability of timely and adequate credit helps farmers by reducing transaction costs, particularly when input-related expenditure rises, and gives them the capital to undertake new investments and/or adopt new technologies (Saboor et al., 2009). However, the additional support was always managed by the informal sources of finances incorporating local money lenders, landlords, traders, etc., in the rural areas. However, it contains a 20 per cent more interest rate than institutional sources, often keeps lands as collateral against the loan, and has a very high recovery rate (Gulati & Bathla, 2002). Henceforth, the rural branches of the banking system or SHGs seem to be promising in reducing debt trap and transportation costs.

Agriculture is a state subject, and the state government is responsible for formulating agricultural growth and development policies. It is responsible for presenting prospective plans for the respective state and ensuring the effective implementation of the scheme. These state governments can have a base from centers to execute the programs according to their needs and resources. Institutional credit has always been perceived as a critical factor for agricultural development in India. Therefore, it is pertinent to establish the causality of credit with agricultural output. The main objectives of the study were to examine the growth of institutional credit for agriculture and identify the factors which are responsible for the growth of Agricultural GDP.

## METHODOLOGY

The present study explored the relationship between the Agriculture Gross Domestic Product (AgGDP) and the various explanatory variables in India from 1971 to 2021. The study was based on secondary data compiled from various published sources. The data on the gross cropped area (GCA), total food grain production, cropping intensity, irrigated area, and agricultural gross domestic product (AgGDP) was compiled from the Agricultural Statistics at a Glance, published by the Department of Agriculture and Co-operation, Ministry of Agriculture, Government of India (GoI). The data on credit were collected from the Reports on Currency and Finance, published by the Reserve Bank of India (RBI).

### Econometric Model

Agriculture drives economic change and development in India through its causal links with factor and product markets (Pandey & Kumari, 2021b; Radhakrishna, 2020). Previous studies have also used a simple regression model (Kannan, 2011; Rehman et al., 2017a & 2017b) to identify the factors responsible for agricultural GDP. Therefore, this study also used the same methodology given by these authors. Using the Johansen cointegration test, we checked the data with the ADF unit root test and long-run relationship. We have developed two models due to the multicollinearity between total credit and sub-sectors of total credit, which means we cannot use both in the same equation. Model-1 comprised the regression of Agriculture Gross Domestic Product (AgGDP) with various agencies of loan supply, whereas the other Model-2 included the aggregate institutional loan supply regression with AgGDP. Hence, the mathematical form of this assumed framework of the study is as follows:

$$AgGDP = a_0 TFP^{a_1} CI_2^{a_2} IRR_3^{a_3} AgWF_4^{a_4} CRRB_5^{a_5} CCB_6^{a_6} CSCB_7^{a_7} CTAg_8^{a_8} \quad (1)$$

Where,

*AgGDP* = Agricultural Gross Domestic Product per cultivated hectare; *TFP*= total food-grain production/ha; *CI*= cropping intensity; *IRR*= Irrigated area; *AgWF*= Agricultural workforce/ha; *CRRB*= credit from Regional Rural Banks/ha; *CCB*= credit from Cooperative Banks/ha; *CSCB*= credit from Scheduled Commercial Banks/ha; *CTAg*= total institutional credit/ha.

So, by taking the natural log of Equation-1, it became

$$logAgGDPY = \log a_0 + a_1 logTFP + a_2 \log CI + a_3 \log IRR + a_4 logAgWF + a_5 logCRRB + a_6 logCRCB + a_7 logCSB + a_8 logCTAg \quad (2)$$

$a_0$, $a_1$, $a_2$ ..... $a_8$ are estimators to define the contribution of factors to Indian agriculture GDP.

Here, two logarithmic regression models are presented. The first model included the institutional credit from the different credit agencies on AgGDP, whereas the second included aggregate (sum of all three institutional loans) impact on AgGDP. This regression showed in two parts to avoid biasedness and was determined by the coefficient of the estimates.

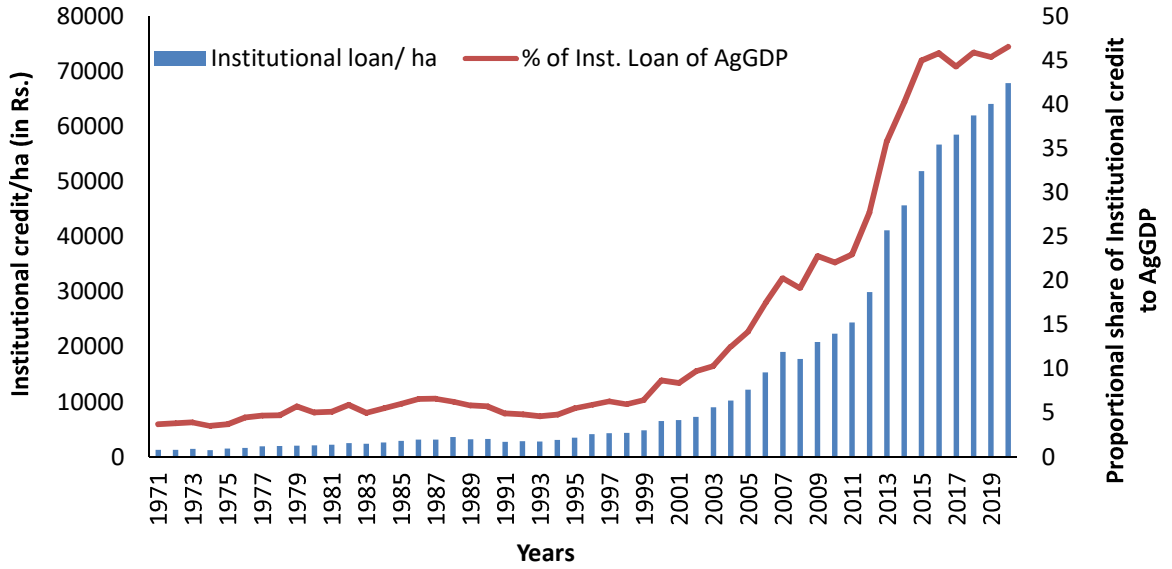
## RESULTS AND DISCUSSION

The investment policy always goes with the historical trend and respective influence. Figure 1 represents the institutional credit trend for the cultivated land. The figure indicated that institutional credit had a steep trend slope after 2003. It may be due to many industrial and economic

reforms that happened during the time. It was required to be proportionate to the agriculture output for efficient allocation. The ratio of institutional credit to agricultural GDP (AgGDP) is expressed in percentages from 1971 to 2020 (Figure 1). The share of institutional, agricultural credit as a proportion of AgGDP rose continuously from 1971 until 2016. Still, recent years have again been characterized by a concern over the falling share of institutional, agricultural credit to AgGDP. It grew from 3.7 per cent in 1971 to 45.8 per cent in 2016. The trend of a proportional share reflected a quantum jump after 2003. These indicators suggested that the agricultural credit system was geared toward agricultural growth, and credit availability to rural cultivators has increased substantially (Kumar et al., 2010).

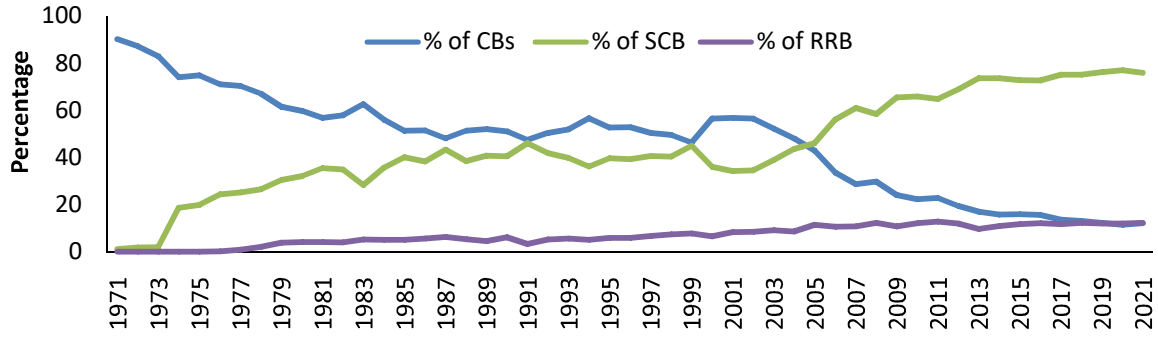
The perusal of Figure 2 showed the ratio of different institutional agencies in the total agriculture credit in India from 1971 to 2020. Both the structure of agriculture and the nature of institutional credit has been undergoing a rapid transformation since independence (Reddy & Dutta, 2018). However, credit supply was not an issue instead, its state of efficiency was a matter of concern. The number of small holders' farmers were excluded from getting the benefits of credit, which restricted the holistic development of the agriculture sector and capped the efficiency of credit (Mishra et al., 2018) & (Hanan & Abdallah, 2016). Indian agriculture also faced challenges after the 1990s, and traditional crop sectors had declined in profitability (Narayan, 2015).

The share of scheduled commercial banks (SCBs) increased from 1 per cent in 1971 to 76 per cent in 2021. Before nationalization, the commercial banks were virtually not lending credit to the agricultural sector




*Source(s): Estimated by authors.*

**Figure 1. Institutional credit for agriculture and share in Agricultural GDP**




*Source(s): Estimated by authors.*

**Figure 2. Percentage share of institutional credit to the agricultural sector**

(Kumar et al., 2010). At the same time, it was the opposite in the case of cooperative banks. The share declined from around 90 per cent in 1971 to 12 per cent in 2021. The proportion of loan disbursement from scheduled commercial banks and RRBs get expanded with time, but scheduled commercial banks had a high proportion of loan disbursement. However, the share of RRBs in institutional sources increased over the period but at a creeping rate.

The compound growth of institutional credit from the different agencies is presented in Table 1. The total institutional credit for agriculture registered a significant positive growth rate of 8.74 per cent. However, credit from RRBs had grown at the highest growth rate at around 14 per cent in the last 50 years. It was attributed to the nationalization of commercial banks in 1969 and mandatory lending for the priority sector of agriculture (Kumar et al., 2010). At the same time, cooperative banks witnessed the lowest growth at around 5 per cent during the same period. It was clear that during the 2000s, the growth of institutional credit was higher from all the sources compared to the other periods. It was mainly due to the government policy of doubling agricultural credit in 2004 for the next three years to boost agricultural production (Shivaswamy et al., 2020).

**Results of Augmented Dickey-Fuller test (ADF)**

Since we have used time series data for our analysis and major limitation of time series data was their assumption of stationarity (Birthal et al., 2019). Most time series tend to be non-stationary, which may lead to spurious regressions and inconsistent and indistinct short-run and long-run elasticities (Hallam & Zanoli 1993; Townsend 1997, McKay et al., 1999). To check for stationarity, we applied the Augmented Dicky Fuller test both with and without trend. These tests in Table 2 did not reject the null hypothesis of unit roots in the data series at the level. Still, at the different differences, the series became stationary; a series was integrated of order two or I (2).

The parametric of the Augmented Dickey-Fuller test is presented in Table 2. There were two ways to accept a hypothesis based on parameters (p-values and test

**Table 1. The annual growth rate of institutional credit to agricultural[1]**

| Period | CBS | SCBs | RRBs | Total |
|---|---|---|---|---|
| 1971-1980 | 2.58 | 36.40 | 46.37 | 7.03 |
| 1981-1990 | 1.27 | 4.23 | 4.17 | 2.58 |
| 1991-2000 | 8.66 | 6.08 | 14.54 | 7.76 |
| 2001-2010 | 5.48 | 24.38 | 20.36 | 16.55 |
| 2011-2020 | 6.25 | 17.02 | 17.26 | 13.06 |
| 1971-2020 | 4.89 | 12.46 | 14.02 | 8.74 |

***Source(s): Estimated by authors.***

**Table 2. Results of Augmented Dickey-Fuller test**

| Variables | At level | | | At second differences | | | |
|---|---|---|---|---|---|---|---|
| | Z-value | Critical value at 5 per cent | p-value | t-value | Critical value at 5 per cent | p-value | Order of integration |
| AgGDP /ha | -1.40 | -3.00 | 0.55 | -1.70 | -1.67 | 0.05 | 1(2) |
| TFP/ha | -1.20 | -3.00 | 0.66 | -1.1 | -1.67 | 0.10 | 1(2) |
| CTAg /ha | -1.22 | -3.00 | 0.66 | -1.31 | -1.70 | 0.09 | 1(2) |
| CI | -0.98 | -3.00 | 0.76 | -0.94 | -1.68 | 0.10 | 1(2) |
| IRR | -2.07 | -3.00 | 0.25 | -2.60 | -1.68 | 0.00 | 1(2) |
| AgWF /ha | -0.87 | -3.00 | 0.79 | -1.00 | -1.67 | 0.10 | 1(2) |
| CCB /ha | -0.28 | -3.00 | 0.92 | -2.02 | -1.68 | 0.00 | 1(2) |
| CSCB /ha | -3.95 | -3.00 | 0.00 | -2.03 | -1.67 | 0.02 | 1(0) |
| CRB /ha | -1.59 | -3.00 | 0.48 | -1.49 | -1.67 | 0.07 | 1(2) |

***Source(s): Estimated by authors.***

[1]***Calculation is based on the real term value of credit from the different institutions.***

statistics). If the p-value was less than 0.05 or 0.10, and in another way, if the test statistics value was greater than the critical values, the null hypothesis was rejected, meaning that no unit root and data presence was stationary. Commercial Schedule Banks were stationary at a level while other variables were stationary at second differences.

**Results of the Johansen Co-integration**

The test for the order of integration was the next step in the long-run relationship between the variables. For co-integration, individual time series should be integrated of the same order, or some linear combination of the series must have a lower order of integration than the original series (Engle & Granger 1987). Unlike the Engle-Granger test, the Johansen test allowed for more than one cointegrating association. It is, nevertheless, susceptible to asymptotic characteristics (large sample size) since a small sample size would provide erroneous findings. Using the test to determine the cointegration of many time series eliminated the problems that arise when mistakes are carried on to the next stage (Hjalmarsson & Österholm, 2007). The Johansen Co-integration test was based on two tests: Trace statistics and maximum eigenvalue statistics. The results of the long-run relationship are presented in Table 3.

The results of the Johnson integration Test, trace statistics value (637.65) was greater than the critical value, which showed a long-term relationship among the variables, are presented in Table 3. And we can go next to analyse the cause-effect relationship.

**Regression Analysis**

The results of the regression analysis are presented in Table 4. The large value of F-statistics shows that the explanatory variables included in the model collectively significantly influenced agricultural GDP. The overall significance of the model could be seen from the value of the coefficient of multiple determinates ($R^2$). The high adjusted $R^2$ values in both models suggested that over 99 per cent of the variation in agriculture production was explained by the explanatory variables included in the model. The results of the study showed that the total food production, cropping intensity, and irrigated land not have a significant impact on Agriculture GDP. In contrast, the sources of institutional loans inferred a significant impact on agriculture GDP. Loan distribution from RRB and Cooperative bank showed a statistically significant positive sign. The coefficient of these variables was 0.37 and 0.47, respectively, which indicated that a 1 per cent increase in these variables would increase the agriculture GDP by 0.37 and 0.43 per cent, respectively. In comparison, scheduled commercial banks showed a similar picture but higher probability. It does have a positive impact on agriculture GDP with the rate of 0.55. Apart from credit supply, the labour force also positively impacts agriculture GDP, with the model supported by the study by Zuberi (1989).

The details of the coefficient when the model contained five explanatory variables, including aggregate loan supplies from all institutional systems, are presented in Table 4. This regression model showed that total food grain production and crop intensity were not having a significant impact, but cropping intensity and labour also had a positive and statistically significant impact on agricultural GDP. The workforce had a similar influence on agriculture GDP. Moreover, the aggregate institutional loan had a significant impact on agriculture GDP. Increasing Institutional loans can increase agriculture GDP growth rate.

**Table 3. Johnson Co-integration test using trace statistic**

| Maximum rank | Parms | LL | Eigen value | Trace statistics | 5% critical value | Parms | LL | Eigen value | Trace statistics | 5% critical value |
|---|---|---|---|---|---|---|---|---|---|---|
| | Model I | | | | | Model II | | | | |
| 0 | 56 | -384.212 | . | 196.374 | 124.24 | 30 | -118.228 | . | 74.934 | 68.52 |
| 1 | 69 | -348.760 | 0.75782 | 125.470 | 94.15 | 39 | -100.489 | 0.508 | 39.456* | 47.21 |
| 2 | 80 | -328.892 | 0.54828 | 85.735 | 68.52 | 46 | -92.178 | 0.282 | 22.836 | 29.68 |
| 3 | 89 | -312.280 | 0.48546 | 52.511 | 47.21 | 51 | -84.844 | 0.254 | 8.168 | 15.41 |
| 4 | 96 | -298.166 | 0.43139 | 24.283* | 29.68 | 54 | -80.883 | 0.146 | 0.244 | 3.76 |
| 5 | 101 | -290.633 | 0.26016 | 9.218 | 15.41 | 55 | -80.760 | 0.004 | . | . |
| 6 | 104 | -286.193 | 0.16275 | 0.336 | 3.76 | . | . | . | . | . |
| 7 | 105 | -286.025 | 0.00670 | . | . | . | . | . | . | . |

***Source(s): Estimated by authors.***

**Table 4. Regression results, 1971-2020**

**(Included observations: 50)**

| Variables | Model-1 | | Model-2 | |
|---|---|---|---|---|
| | **Coefficient** | **Standard error** | **Coefficient** | **Standard error** |
| Ln (TFP/ha) | 0.49 | 0.40 | 0.43 | 0.44 |
| Ln (CTAg/ha) | -- | -- | 0.26** | 0.04 |
| Ln (CI | 3.05 | 2.90 | -3.87 | 3.04 |
| Ln (IRR | 0.43 | 0.68 | 3.06** | 0.49 |
| Ln (AgWF | 0.59** | 0.12 | 0.59** | 0.13 |
| Ln (CRRB/ha | 0.37** | 0.17 | -- | -- |
| Ln (CCB/ha) | 0.43** | 0.17 | -- | -- |
| Ln (CSCB/ha) | 0.55** | 0.15 | -- | -- |
| Constant | -12.72 | 14.89 | | |
| F statistics P value | 1873.99 | 0.996 | 1378.73 | 0.992 |
| Adjusted R square | | | | |

***Source(s): Estimated by authors.***
***** Significant at 5 per cent level.***

## CONCLUSIONS

It is clear from the above discussion that institutional credit per hectare for the agriculture sector and the proportional share of agricultural GDP increased over the period. Though early phases witnessed a major contribution by cooperative banks for the agriculture sector after that, scheduled commercial banks (SCBs) took place in India. However, the share of commercial banks declined sharply from around 90 per cent in 1971 to 12 per cent in 2021. This study confirmed that along with the institutional credit, other factors, that is, labour force and irrigation, also emerged as supportive factors to enhance the agricultural GDP in India. However, cropping intensity had no significant impact on it. The results also clearly indicated a need to formulate and initiate new policies and funding agencies for the development and improvement of the agriculture sector in general and smallholders and tenants' farmers in particular. The government can also take the initiative to increase institutional credit outreach by simplifying the procedure for loan disbursement in India.

## REFERENCES


Bahsi, N., & Ceti, E. (2020). Determining agricultural credit impact on agricultural production value in Turkey. *Ciencia Rural*, *50*(11), 1-13.

Birthal, P.S., Pandey, G., Jumrani, J., & Jaweriah, N. (2019). Supply response in Indian dairying. *Indian Journal of Animal Sciences, 89*(4), 459-465.

Engle, R.F., & Granger C.W. (1987). Co-Integration and error correction: Representation, estimation, and testing. *Econometrica* , *55*(2), 251-276.

Gulati, A., & Bathla, S. (2002). *Institutional credit to Indian agriculture: Defaults and policy options.* Occasional Paper. National Bank for Agriculture and Rural Development (NABARD), Mumbai (India).

Hallam, D., & Zanoli, R. (1993). Error correction models and agricultural supply response. *European Review of Agricultural Economics, 2*(1), 151–166.

Hanan, A., & Abdallah. (2016). Agricultural credit and technical efficiency in Ghana: Is there a nexus? *Agricultural Finance Review, 76*(2), 309-324.

Hazrana, J., Birthal, P.S., Negi, D.S., Mani, G., & Pandey, G. (2019). Spatial spillovers, structural transformation, and economic growth in India. *Agricultural Economics Research Review, 32*(2), 145-158.

Hjalmarsson, E., & Österholm, P. (2007). *Testing for cointegration using the Johansen methodology when variables are near-integrated.* IMF Working Paper No. WP/07/141. Washington DC, USA: Western Hemisphere Division, International Monetary Fund. Retrieved from https://www.imf.org/external/pubs/ft/wp/2007/wp07141.pdf.

Kannan, E. (2011). Trends in India's agricultural growth and its determinants. *Asian Journal of Agriculture and Development*, *8*(2), 79-99.

Karmakar, K. (2008). Trends in rural finance. *Indian Journal of Agricultural Economics*, *63*(1), 5-18.

Kumar, A., Singh, K.M., & Sinha, S. (2010). Institutional credit to the agriculture sector in India: Status, performance, and determinants. *Agricultural Economics Research Review, 23*(2), 253-264.

McKay, A., Morissey, O., & Vaillant C. (1999). Aggregate supply response in Tanzanian agriculture. *Journal of International Trade and Economic Development, 8*(1), 107-124.

Mishra, R., Chavan, P., & Verma, R. (2018). Credit and efficiency in Indian agriculture: Evidence from household-level data.

View publication stats



*Indian Journal of Agricultural Economics, 73*(1), 88-108.

Narayan, S. (2015, January). *The productivity of agricultural credit in India* (Working paper). Indira Gandhi Institute of Development Research (IGIDR), Mumbai.

Nasar, A. (2015). *Trends and patterns of agricultural credit In India during the post-reforms period* Doctoral dissertation). Cochin University of Science and Technology, Cochin and Kerala University, Thiruvananthapuram.

Pandey, G., & Kumari, S. (2021a). Dynamics of agricultural growth and diversification in Eastern India. *Journal of Agribusiness in Developing and Emerging Economies, 11* (2), 105-120.

Pandey, G., & Kumari, S. (2021b). Understanding agricultural growth and performance in Bihar, India. *SN Business and Economics*, *1*(10), 1-17.

Pandey, G.K. (2016). Extent, magnitude, and determinants of indebtedness among farmers in eastern India: A survey-based study. *Indian Journal of Agricultural Economics, 71*(4), 450-462.

Radhakrishna, R. (2020). Towards inclusive agricultural development: Growth performance, welfare challenges, and policy innovations. *Indian Journal of Agricultural Economics*, 75(1), 1-28.

Reddy, T.K., & Dutta, M. (2018). Impact of agricultural inputs on agricultural GDP in the Indian economy. *Theoretical Economics Letters*, *8*(10), 1840-1853.

Rehman, A., Chandio, A.A., Hussain, I., & Jingdong, L. (2017a). Is credit the devil in agriculture? The role of credit in Pakistan's agricultural sector. *The Journal of Finance and Data Science, 3*(1-4), 38-44.

Rehman, A., Jingdong, L., Chandio, A.A., & Hussain, I. (2017b). Livestock production and population census in Pakistan: Determining their relationship with agricultural GDP using econometric analysis. *Information Processing in Agriculture, 4*(2), 168-177.

Saboor, A., Hussain, M., & Munir, M (2009). Impact of microcredit in alleviating poverty: An insight from rural Rawalpindi. *Pakistan Journal of Life and Social Science*, 7 (1), 90-97.

Shivaswamy, G. P., Raghavendra, K.J., Anuja, A.R., Singh, K.N., Rajesh, T., & Kumar, H.H.V. (2020). Impact of institutional credit on agricultural productivity in India*:* A time series analysis*. Indian Journal of Agricultural Science*, *90*(2), 412-419.

Siddiqi, M. W., & Baluch, K.N. (2009). Institutional credit: A policy tool for enhancement of agricultural income of Pakistan. *International Research Journal of Arts & Humanities, 37*(3), 157-174.

Sidhu, R., & Gill, S.S. (2006). Agricultural credit and indebtedness in India: Some issues. *Indian Journal of Agricultural Economics*, *61*(1), 11-34.

Townsend R. (1997). *Policy distortions and agricultural performance in the South African economy* (Discussion Paper No 138). Information Business Unit, University of Pretoria, Hatfield (South Africa).

Zuberi, H. A. (1989). Production function, institutional credit and agricultural development in Pakistan. *The Pakistan Development Review, 28*(1), 43-55.

Proc. of the International Conference on Electrical, Computer and Energy Technologies (ICECET 2022)
20-22 July 2022, Prague-Czech Republic

# Impact of temperature on negative capacitance based Junctionless Accumulation Mode Ferroelectric FET(JAM-FE-FET)

Snehlata Yadav
*Department of Electronics & Communication Engineering*
*Delhi Technological University*
Delhi, India
snehlata_2k19phdec03@dtu.ac.in

Sonam Rewari
*Department of Electronics & Communication Engineering*
*Delhi Technological University*
Delhi, India
rewarisonam@gmail.com

Rajeshwari Pandey
*Department of Electronics & Communication Engineering*
*Delhi Technological University*
Delhi, India
rpandey@dce.ac.in

2022 International Conference on Electrical, Computer and Energy Technologies (ICECET) | 978-1-6654-7087-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECET55527.2022.9872837

***Abstract*—This paper examines the temperature impact on the characteristic parameters of a negative capacitance based device structure Junctionless Accumulation Mode Ferroelectric Field Effect Transistor (JAM-FE-FET). Here, Zr-doped $HfO_2$ (HZO) ferroelectric material is employed as gate stack along with Titanium Nitride (TiN) and Silicon dioxide (SiO2). HZO ferroelectric material generates negative capacitance (NC) effect during the device operation. The main focus of this work is to investigate the dc characteristics of the device in terms of threshold voltage ($V_{TH}$), subthreshold swing (SS), on current (Ion) and switching ratio ($I_{on}/I_{off}$) at different temperatures. Further, the reported device shows better performance in comparison to the Junctionless Ferroelectric Field Effect Transistor (JL-FE-FET).**

***Keywords—Negative Capacitance, ferroelectric material, Junctionless transistor, Junctionless Accumulation Mode, HZO***

## I. INTRODUCTION

The ability of negative capacitance (NC) based ferroelectric field-effect transistor (FE-FET) to attain subthreshold swing (SS) of 60 mV/dec at ambient temperature, has caught researchers' interest. [1]. In the past few decades, numerous researches on NCFET have been done on various ferroelectric materials in order to obtain steeper SS of the device. Various ferroelectric materials have been reported so far that exhibits NC effect, such as $BaTiO_3/BiFeO_3$ [2], PVDF [3-4], PZT [5], and HZO [6]. Among these materials, HZO has been found to be advantageous over other ferroelectric materials due to its compatibility with CMOS technology. In addition to this, the downscaling of device dimensions to nanometres range makes the fabrication process very difficult [7-8]. Therefore, to ease the fabrication complexity, a new device structure called Junctionless Transistor (JLT) was introduced. However, some major limitations in this device include the larger work function of gate metal to make sure that the device turns off. To overcome such issue, a novel improved structure known as the JAM FET was proposed [9], with strongly doped Source/Drain regions and moderately doped channel region, which reduces the mobility deterioration. A new device structure, which combines the benefits of JAM as well as NC effect was proposed and studied [10]. Since the device's performance degrades at high temperatures and at Subthreshold Swing (SS) values, it's important to investigate the effect of temperature variation on the device's performance. This paper investigates the effect of temperature on dc performance metrics such as threshold voltage ($V_{TH}$), subthreshold swing (SS), on current (Ion), and switching ratio ($I_{on}/I_{off}$). It is revealed from the simulation results that the JAM-FE-FET shows better performance in comparison to JL-FE-FET at various temperatures.

## II. DEVICE DESIGN & SIMULATION METHOD

The schematic representation of the proposed device is shown in Fig. 1. The gate stack includes HZO as ferroelectric layer and $SiO_2$ as the insulating layer of thickness ($t_{IL}$) 1 nm. The doping with n-type dopants in the channel is 1e17 $cm^{-3}$, but the drain and source regions with n-type dopants are heavily doped with concentration 1e19 $cm^{-3}$. The gate material used is Titanium Nitride (TiN) metal whose work function ($\phi_m$) is 4.65eV. The simulations have been performed using the Silvaco ATLAS TCAD simulator. The simulations are performed on Silvaco Atlas TCAD and the models used are Shockley-Read-Hall (SRH) recombination, Lombardi CVT , fermi, ferro, and Landau-Khalatnikov (LK) model [11]. After then, numerical methods like Gummel and Newton were utilized to enhance

978-1-6654-7087-2/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2022 at 05:04:40 UTC from IEEE Xplore. Restrictions apply.

convergence. Moreover, the list of simulation parameters of the device structure are given in Table 1. The simulations were performed at different temperatures varying from 300 K-500 K using the given models. The material parameters of ferroelectric material such as spontaneous polarization($P_s$), remanent polarization ($P_r$) and coercive field ($E_c$) for HZO have been obtained from the literature [12], and have been mentioned in Table 2.

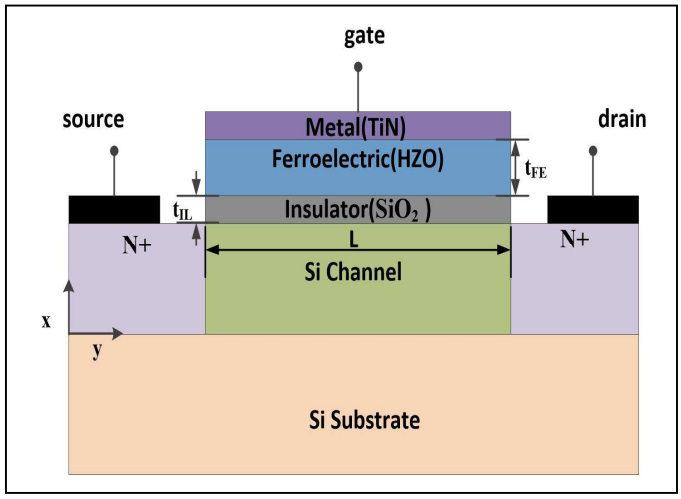



Fig. 1. Device structure of the JAM-FE-FET

TABLE I. DEVICE PARAMETERS AND THEIR VALUES

| Parameters | Symbol (unit) | Proposed | Conventional |
|---|---|---|---|
| Si channel length | L (nm) | 90 | 90 |
| Ferroelectric thickness | $t_{FE}$ (nm) | 10 | 10 |
| Insulator thickness | $t_{IL}$ (nm) | 1 | 1 |
| Source/Drain doping | N+ ( $cm^{-3}$) | $1x10^{19}$ | $1x10^{19}$ |
| Channel doping | N ($cm^{-3}$) | $1x10^{17}$ | $1x10^{19}$ |
| Gate metal work function | $\phi_m$ (eV) | 4.6 | 5.1 |

TABLE II. MATERIAL (FERROELECTRICS) PARAMETERS

| Parameters | Symbol (unit) | Values |
|---|---|---|
| Coercive Field | Ec (MV/cm) | 1-2 |
| Remnant Polarization | Pr ($\mu C/cm^2$) | 1-40 |
| Spontaneous Polarization | Ps ($\mu C/cm^2$) | 10-40 |
| Dielectric constant | ε | 30 |

## III. RESULTS & DISCUSSION

This section reveals the influence of temperature fluctuation from 300K to 500K on HZO ferroelectric gate stack in both JAM and JL Ferroelectric FET alongwith SiO2 dielectric and silicon channel. Fig. 2 shows the influence of temperature on input characteristic of both the devices. It can be seen in this diagram that when the temperature rises, the drain current decreases. Large variation at lower gate voltages has been observed with temperature as compared to the super threshold region.

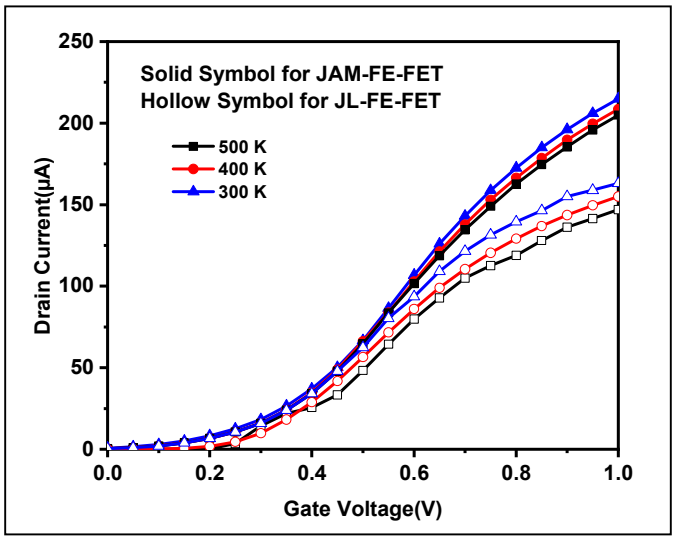



Fig. 2. Input characteristics at different temperatures

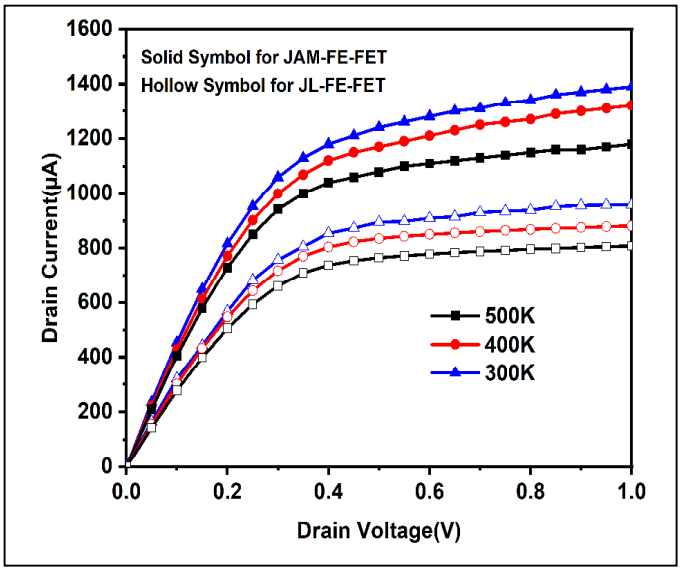



Fig. 3. Output characteristics at different temperatures

This happens due to the SRH recombination as it has an exponential dependence on temperature. Fig. 3 illustrate the output characteristics of both the device configuration at various temperatures. The figure clearly shows that the rise in drain current with drain voltage at lower temperatures is significant. Additionally, the detailed dc performance characteristics such as threshold voltage, SS, and $I_{on}/I_{off}$ ratio with respect to the temperature is obtained and compared for

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2022 at 05:04:40 UTC from IEEE Xplore. Restrictions apply.

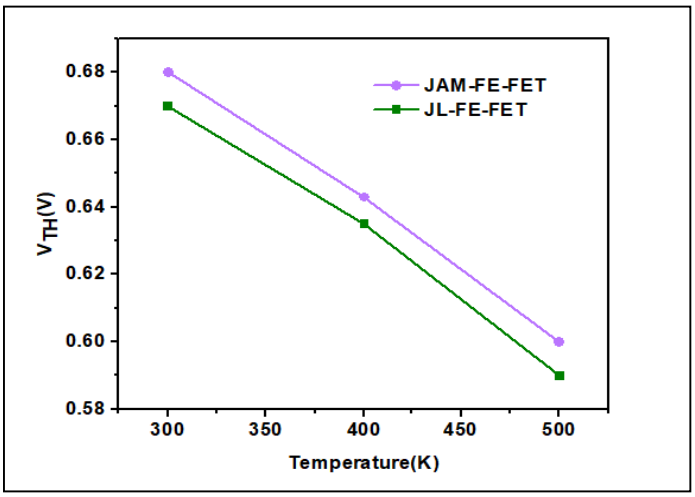



Fig. 4. $V_{TH}$ variation at different temperatures

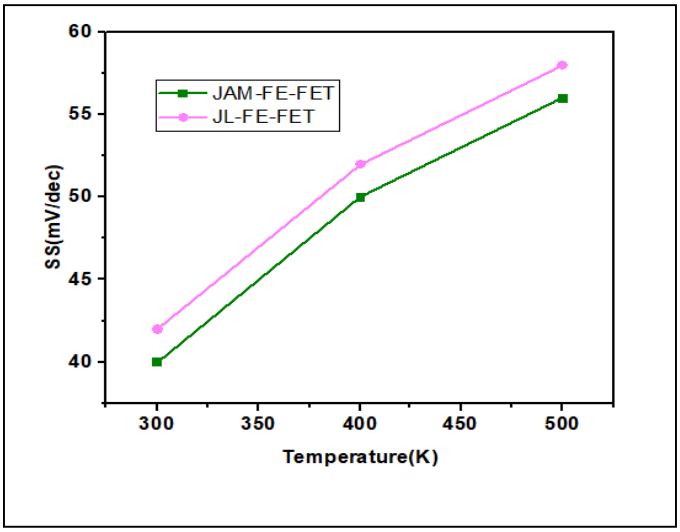



Fig. 5. SS variation at various temperatures

JAM and JL ferroelectric FET device configurations as in Fig. 4, Fig. 5, and Fig. 6. Fig. 4 shows the threshold voltage variation at different temperatures. It can be assessed from the figure that as the temperature increases, there is a decrease in threshold voltage, thereby showing a positive temperature coefficient due to the variation in the bandgap and fermi level. The change in threshold voltage is around 60mV when the temperature is varied from 300K to 500K. It can also be assessed that the threshold voltage is higher for JAM configuration when compared to JL ferroelectric configuration. Fig. 5 shows the subthreshold swing (SS) variation with respect to temperature. It can be assessed from the figure that SS increases with rise in temperature since SS is directly proportional to temperature as *(SS=kT/q)*. Therefore, this plot verifies the steeper SS at low temperatures. Moreover, the SS is much steeper for JAM configuration than JL FE-FET configuration. Fig. 6 depicts the variation in on current at various temperatures. It can be evident from Fig. 6 that Ion increases with rise in temperature.

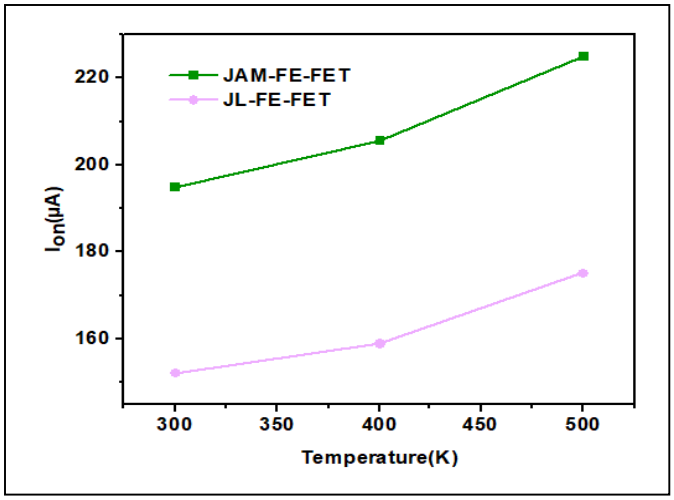



Fig. 6. Ion variation at different temperatures

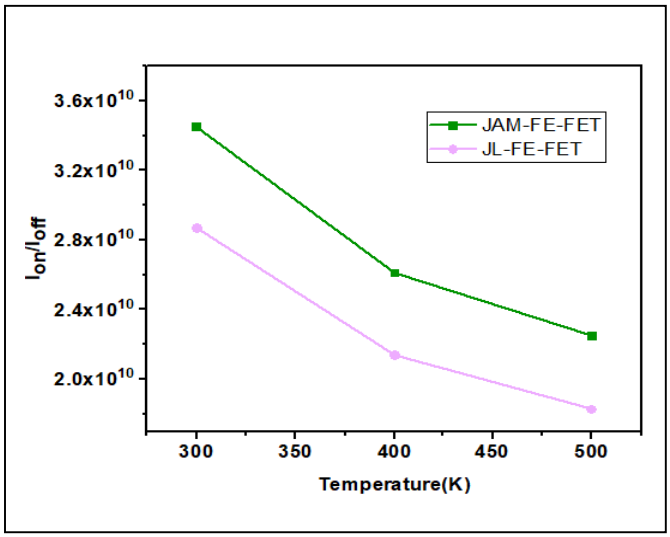



Fig. 7. $I_{on}/I_{off}$ variation at different temperatures

Fig. 7 represents the variation in $I_{on}/I_{off}$ with respect to temperature. As observed from Fig. 7, the $I_{on}/I_{off}$ ratio reduces with an increase in temperature. It is also evident from the figure that the $I_{on}/I_{off}$ ratio is improved for JAM ferroelectric FET when compared to JL Ferroelectric FET.

## IV. Conclusion

A comprehensive analysis of the dc performance dependence on temperature has been carried out for both JAM and JL ferroelectric FET. In the proposed structure, the inclusion of the ferroelectric layer (HZO) as the gate stack improves the device performance by achieving the NC effect during its operation. This research reveals that as temperature rises, $I_{on}$ current rises as well, and attained 225μA at 500K. Moreover, the highest $I_{on}/I_{off}$ ratio and minimum SS as $3.45\times10^{10}$ and 40 mV/dec has been obtained at 300K for JAM ferroelectric FET respectively. The threshold voltage variation is around 60 mV with respect to the change in temperature, which increases the device immunity towards temperature variations. As a result of this investigation, temperature has a substantial impact on the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2022 at 05:04:40 UTC from IEEE Xplore. Restrictions apply.

device's dc properties, making the JAM-FE-FET a viable choice for ultra-low power digital applications.

## REFERENCES


[1] S. Salahuddin, S. Datta, "Use of negative capacitance to provide voltage amplification for low power nanoscale devices," Nano Letters, vol. 8, no. 2, pp. 405-410, 2008.

[2] Ya-Fei Hou, Wei-Li Li, Tian-Dong Zhang, Yang Yu, Ren-Lu Han, and Wei-Dong Fei, "Negative capacitance in BaTiO3/BiFeO3 Bilayer Capacitors," ACS Applied Materials & Interfaces vol. 8 (34), pp. 22354-22360, 2016.

[3] M.A. Alam, M. Si, P.D. Ye, "A critical review of recent progress on negative capacitance field-effect transistor," Appl. Phys. Letters 114, 090401, 2019.

[4] S. Bakaul, C. Serrao, M. Lee et. al. , "Single crystal functional oxides on silicon," Nat. Commun., vol.7, pp.10547, 2016.

[5] S. Dasgupta, A. Rajashekhar, K. Majumdar, N. Agrawal et.al., "Sub-kT/q switching in strong inversion in PbZr0.52Ti0.48O3 gated negative capacitance FETs," IEEE J on Exploratory Solid-State Computational Devices and Circuits, vol. 1, pp.43-48, 2015.

[6] C. Wang, J. Wu, H. Yu, G. Han et. al., "Effects of Temperature on the Performance of Hf0.5Zr0.5O2-based Negative Capacitance FETs," IEEE Electron Device Letters pp(99):1-1, 2020.

[7] H. Amrouch, M. Rapp, S.Salamin, J. Henkel, "Impact of negative capcitance field effect transistor (NCFET) on many core system," In: Chen JJ. (eds) A Journey of Embedded and Cyber-Physical Systems,2021.

[8] M. Kobayashi, "A perspective on steep sub-threshold-slope negative capacitance field effect transistor," Appl. Phys. Express 11, 2018.

[9] TK. Kim, et.al., "First demonstration of junctionless accumulation mode bulk FinFETs with robust junction isolation," IEEE Electron Device Letters 34(12), 2013.

[10] S. Yadav, S. Rewari, and R. Pandey, "Junctionless Accumulation Mode Ferroelectric FET (JAM-FE-FET) for High Frequency Digital and Analog Applications," *Silicon*, 2022.

[11] ATLAS, (2019), User's manual. SILVACO International, CA, Santa Clara, US.

[12] M. Hoffmann *et al*., "Demonstration of High-speed Hysteresis-free Negative Capacitance in Ferroelectric Hf0.5Zr0.5O2," Technical Digest - International Electron Devices Meeting, IEDM, pp. 31.6.1-31.6.4, 2019.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 26,2022 at 05:04:40 UTC from IEEE Xplore. Restrictions apply.

*2022 2nd International Conference on Intelligent Technologies (CONIT)*
*Karnataka, India. June 24-26, 2022*

# Improved Edge Detection Approach to Tackle Edge Thickness and Better Edge Connectivity

Yatharth Saxena
*Department of Electronics and Communication Engineering*
*Delhi Technological University*
New Delhi, 110042, India
yatharthsaxena99@gmail.com

Nirdesh Mishra
*Department of Electronics and Communication Engineering*
*Delhi Technological University*
New Delhi, 110042, India
nirdeshmishra2k18ec113@dtu.ac.in

Mohd Sameer
*Department of Electronics and Communication Engineering*
*Delhi Technological University*
New Delhi, 110042, India
mohd.sameer2k18ec104@dtu.ac.in

Pankaj Dahiya
*Department of Electronics and Communication Engineering*
*Delhi Technological University*
New Delhi, 110042, India
pankajdahiya@dtu.ac.in

2022 International Conference on Intelligent Technologies (CONIT) | 978-1-6654-8407-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CONIT55038.2022.9848285

**Abstract—Edge detection is substantial in helping us to pre-process any image for various applications from helping us to detect objects to detecting various medical conditions. The paper tackled one major shortcoming with the currently present system which is edge thickness. To improve there is an implementation of multiple thresholds instead of two thresholds generally used by techniques like that in Canny. The selected method solves multiple problems perfecting the handling of errors and more real to truth results. Our aim of refining the method helps us in better edge detection in images with low contrast as well as medical images like MRIs and X-rays.**

***Keywords— Edge Thickness, Edge Detection, Multiple Thresholding***

## I. INTRODUCTION

In Computer Vision or Image Processing, edge detection has been the primary pivot for all the processing. One needs to determine the edges accurately to infer details out of them. The major avenues to it are either through thresholding or by edge fitting. In the thresholding method, the discontinuity is being analysed on image characteristics and improved by different operators while in other fitting is being done of ideal edge pixels. The historical importance of studying the challenge is in the fact that edge detection suffers from extremely high volatility and hence the inability of the standard model to suffer with variations with a lot of scope available for improvement. Ideally, any edge detector for an image may be following connected adjacent curves that bespeak the edges of the object we want to detect the edge of. Labelling as well as curves corresponding to discontinuities and variations mentioned above.

### *A. Research History*

Over time Researchers have done multiple efforts in forming improved operators for edge detection. The boundaries of an object in a picture are bound to have localized changes or intensity changes in image. This does combine with our commonly held belief that in image segmentation, our first step is bound to be edge detection. Over the years this is the reason or fuel behind making advanced edge detection algorithms and have formed a major part of research in low-level vision and led to several edge detection algorithms published in journals over the last ten to twenty years.

The gradient of the signal, i.e. one-dimensional - there's maxima located at the centre of the original signal and is a fundamental characteristic of the gradient filter. If the first derivative is maxima, then the second derivative depicts zero. Hence, finding zeros in the second derivative is one of the methods [1]. Rosenfeld et al. formed a method based on quality evaluation although it lacked uniformity in width [2]. Improvements made using partial derivative [3] break down local information around the bounding. In 1998 there were improvements for noise removal [4] using Sobel operator with threshold wavelet.

Lots of improvements were suggested on the Sobel operator but the lack of acceptable results made Dollar et al. [5] make a method for boundary detection using a Probalistic Tree that chooses and combines of varied scale. In the year 2016, Goel [6] discussed BFA getting improved results than canny as it uses a different swarm filter instead of a gaussian filter. Later proposals discussed the problems with attenuation as the higher signal is seldom one-dimensional. To get around the discussed problem Hong [7] introduced differential operator on histograms overcoming the shortcomings of canny which also was a little resistant to noise. In the year 2018, Beşdok et al. [8] developed a backtracking clustering algorithm for images with a high grain of noise resulting in better statistical outcomes and truer to real-life detection. SAR images de-noising principle when put together was capable of generating visual edges on SAR images [9]. Recently Otsu model has been used to develop MapReduce parallel programming-based Otsu-canny operator aiming to better double thresholding and performance of canny [10]. Later many methods based on contour detection have been suggested [11][12] while they can also be used for performance analysis [13] but they are computationally extensive. Earlier gradient based methods like Prewitt [14] and Sobel [15] are easy on load but unreliable. Old canny [16] is being improved by many changing characteristics developing many models of



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:59:29 UTC from IEEE Xplore. Restrictions apply.

Improved Canny [17] while always suffering from double edges.

The literature review informed about the research on this topic in the last 50 years. Progress in techniques and developed operators over the years while the basis of most of them remains the same. That shows that there's still lack in a few major aspects of edge detection mostly with edge thickness and edge connectivity in low contrast or medical images. The paper suggests new way to tackle the problem using multiple thresholding while complying with following points:

- Model should tackle edge thickness with uniform edge width
- Complete and fuller real edge connectivity
- Results must be within reasonable MSE (Mean Square Error)

## II. PROBLEM CONSTRAINT

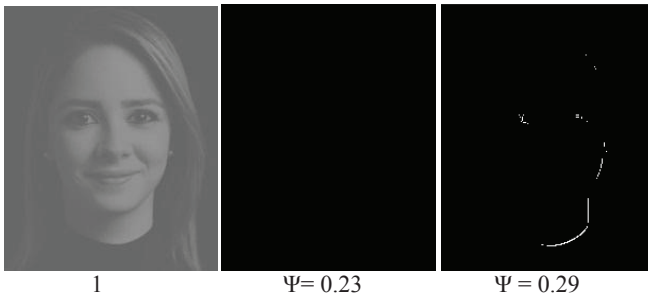
Frequency domain or Spatial domain are two major categories of edge detection. The latter accomplishes using operators over the pixels in the image. They are further categorized into first-order and second-order techniques based on the first and second derivatives respectively. The second-order methods are that of Laplacian and Canny. The gradient based techniques of first order were relatively easy to implement but lacked continuity in edges.

While that could be converted to the frequency domain performing Fourier Transform and various processing could be done on it later. Hence, data are taken from low-frequency while image contour could be extracted from high frequencies. They do lack huge space for search.

TABLE I. CHARACTERISTICS OF PHYSICAL MEASURES OF AN EDGE

| Edge Character | Equation | Annotations |
|---|---|---|
| Noise | $n = -1\,(1 + xC)^{-1}$ | n: noise index<br>x: distance from noisy pixels to true pixel. |
| Smooth Index | $s = (1 + R_E)^{-1}$ | s: index of smoothness<br>$R_E$: Roughness of edge |
| Local Index | $l = (E_L + 1)^{-1}$ | l: localization index<br>$E_L$: Error Location Average |
| Quality | $Q = C_W C + S_W S + L_W L$ | $C_W$: Continuity Weight<br>$S_W$: Smoothness Weight |
| Thickness | $T = W_A$ | T: Thickness Index<br>$W_A$: Average Width |

Based on the above quality measures in Table 1 and general behaviour there is a clear understanding that Canny technique basically uses non-maxima suppression varies as the result. Certain shortcomings of Canny are as follows: 1) High need of connecting edges of the obtained result to reach true edges. 2) Gaussian Filter either lack smoothness or makes it too smooth. 3) Open edges are missed as do the junctions. 4) Blur pixels at missing junctions, open ends and corners. Low contrast images often inhibit true detection as shown in Fig 1Suitable gray threshold needs to be calculated as many a times image is not suitable for it and hence by testing various images it is being analysed that gray threshold of 0.2 is required in many a case. Hence, increasing contrast fulfils 0.2 threshold condition and edges are being detected.



1 Ψ= 0.23 Ψ = 0.29

Fig. 1. Lower Contrast can make models totally unable to detect any edge information

## III. ARCHITECTURE (ALGO1)

In this section there is a description of the proposed algorithm based on multiple thresholding. The Flowchart in Fig 2 depicts the same and the algorithm is as follows:

1. Image is taken from the database.
2. Gray threshold for individual pixels are computed by (1)

$$\Psi = \frac{(\sum_{i=1}^{y}\sum_{j=1}^{x} ara)}{x.y} \quad (1)$$

x and y show the dimensions of the pixel while ara depicting array of input image. Also, every image has unique gray threshold which provides mean for those individual images.

3. Adjustment of Intensity: Another threshold for the image is then calculated by (2)

$$\phi = \frac{\Psi.20}{8.23} \quad (2)$$

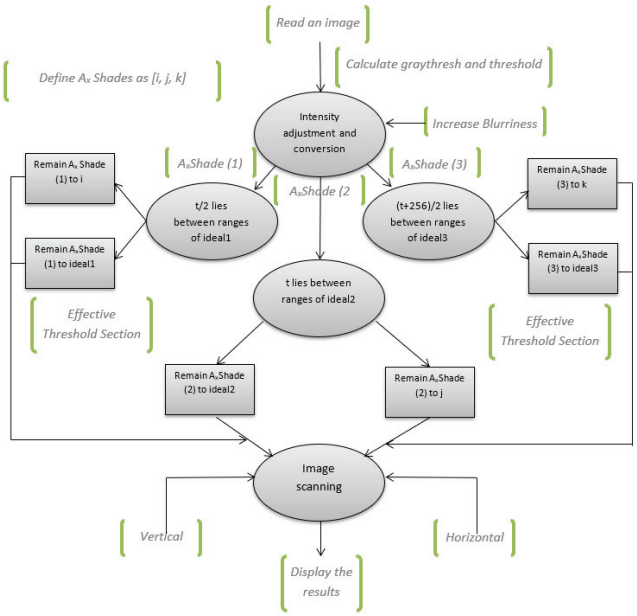



Fig. 2. Flowchart depicting the Algorithm

the purpose is to step up the contrast of the image by multiplying graythresh with constant 20/8.23 as under the fact that simulations suggested that bigger values will result in increased ϕ values. This will induce blurriness bringing in high smoothness reducing edge data. While lower value would bring in double edges and lower desired results. From the



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:59:29 UTC from IEEE Xplore. Restrictions apply.

experiments, it is being found that there's no edge detection in images with contrast less than 0.2. We generated the function to increase image contrast on needed images. Hence, we solved the problem with 'Edge Thickness'.

4. Conversion of phase in grayscale: Intensity varies from 0 to 255 and from the Literature review we know that selection of threshold is basically an art. Therefore, considering faster compute abilities there is a use of triple thresholding. The values being i, j, k aiming an overall coverage. As in audio equipment, there is a need of wider coverage of highs, lows and mids similarly here we needed such intensity values that provides us best possible intensities for wider coverage of the space are taken as 64, 84 and 169. The conversion done by (3)

$$ara(i,j) = \frac{\sum(ara(i,j,:))}{3} \quad (3)$$

where, i∈ (1, 2, 3, …, n) and j∈ (1, 2, 3, …, n)). Here the intensities are made to satisfy the requirements of low intensities, also 84 lies in grayscale mid-range and as to avoid double edges there is also an intensity measure of 169. Three intensities provide us with complete coverage of range and hence better edge connectivity.

5. Selection of solution: As all three thresholds are calculated from the above methods than the horizontal arrays of pixels have been taken and will add the data in an empty array generating final edges for the direction.
If (ara (i-k, j-k) > ϕ then update empty matrix, where k∈ {-1, 0, 1}

6. Repeating same as 5 but for vertical sets of pixels.

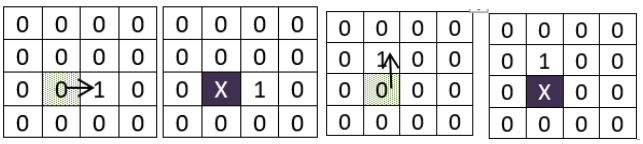



Fig. 3. Matrix depicting the edge pixel comparing it with the threshold

The processing of pixels in both horizontal and vertical direction as shown in Fig 3 he shown pixel is considered as an edge pixel if the pixel adjacent right/left to it has a larger value than the calculated threshold and similarly for vertical the adjacent above/below pixel depicts larger value than the threshold.

## IV. Results

Using Cornell's contour performance model [11] and BSDS500 that is a modified BSDS300 been used as measure. Tested on Canny, Improved Canny and proposed method to make performance comparison. There are 300 images for training and 170 for testing [12][13] for contours with true reality as shown in Fig. 4., Fig. 5. and Fig 6. Major benchmarked techniques as that of Prewitt [14] and Sobel [15] is not included because of non-reasonable MSE in the experiments setting Canny [16] and ICA [17] as the measure of performance analysis.

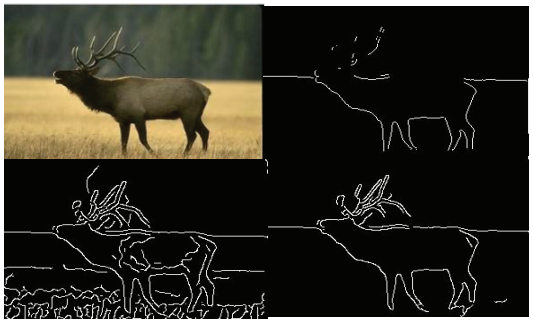


Fig. 4. One of the test image from database. Original Image, Canny [10], ICA [17], Algo1 respectively

Fig. 5. Low Contrast Image on test. Original image, canny [10], ICA [17], Algo1 respectively

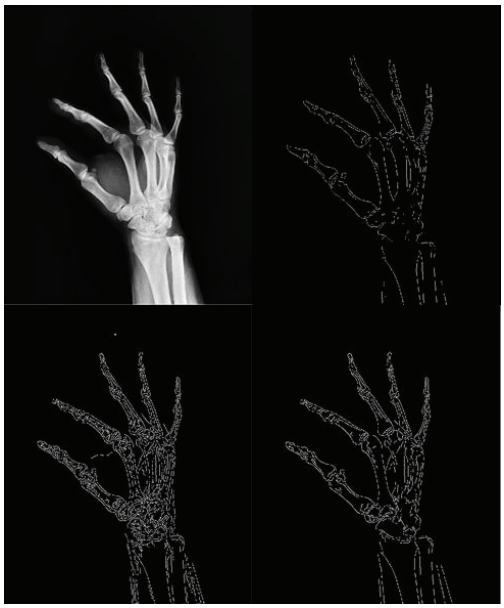


Fig. 6. Medical Images like that of an x-ray on test. Original Image, Canny [10], ICA [17], Algo1 respectively

Although through naked eye it might seem that the proposed method was unable to detect background edges successfully but as presented in true image the key object edges were detected as required without any successive discontinuity as observed in other methods. In Medical images [18][19] there is always been high error rate because of complexity and low contrast. The medical images in our model have shown great potential.

## V. Performance Analysis

Quantitative Analysis based on Mean Square Errors [20][21] (MSE) has been performed for previously available



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:59:29 UTC from IEEE Xplore. Restrictions apply.

methods and proposed method while comparing with true ground reality data available from the dataset.

### *A. Mean Square Error (MSE)*

MSE for individual and respective result is calculated using (4)

$$E_r = \frac{1}{a.b}\sum_0^{a-1}\sum_0^{b-1}\|O(u,v) - P(u,v)\|^2 \qquad (4)$$

The source image data in a matrix is shown by 'O' while the result image data in a matrix is shown by 'P'. While 'a' and 'b' represent total count of rows and columns respectively. Index value of rows being depicted by 'u' while same depicted for columns by 'v'. High MSE corresponds to non-acceptable results.

The calculated MSE for the resulted images as well as comparison with canny and ICA are depicted in Table II.

TABLE II. CALCULATED MSE VALUES FOR THE RESULTS

| Image No. | Canny [10] | ICA [17] | Algo 1 |
|---|---|---|---|
| 1 | 0.1807 | 0.3184 | **0.1564** |
| 2 | 0.1658 | 0.2758 | **0.1158** |
| 3 | 0.1253 | 0.3115 | **0.0985** |
| 4 | **0.1145** | 0.3154 | 0.1152 |
| 5 | 0.0715 | 0.1358 | **0.0704** |
| 6 | 0.1424 | **0.1340** | 0.1415 |
| 7 | 0.1433 | 0.1419 | **0.1379** |
| 8 | 0.0892 | 0.1400 | **0.0461** |
| 9 | 0.1514 | 0.1520 | **0.1381** |
| 10 | 0.9821 | 0.1450 | **0.0852** |
| 11 | 0.1278 | 0.1924 | **0.0659** |
| 12 | 0.1064 | 0.1181 | **0.0553** |
| 13 | 0.0784 | 0.1371 | **0.0377** |
| 14 | 0.2432 | 0.2083 | **0.1981** |
| 15 | 0.1135 | 0.1172 | **0.0983** |
| 16 | 0.1782 | 0.1979 | **0.1462** |
| 17 | 0.1576 | 0.1590 | **0.1298** |
| 18 | 0.1235 | 0.1731 | **0.0987** |
| 19 | 0.1513 | 0.1735 | **0.0448** |
| 20 | 0.1187 | 0.1437 | **0.0669** |

The values as shown in Table II depicts the calculated MSE and we can observe that our proposed method gives lower MSE values consistently and is very close in image 4 and 6 as well. To depict it more clearly, we've converted the table into a line graph as shown in Fig.7..

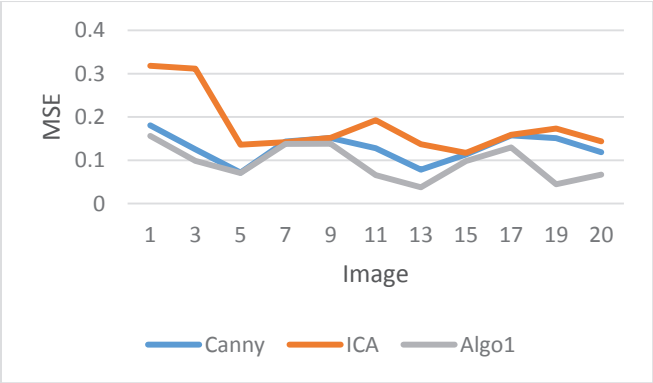



Fig. 7. MSE of different methods

## VI. CONCLUSION

The Proposed Method success on medical images were exemplary and was consistently performing at-par or better than the traditional models. The ability to match short contours for its better overall performance. The following conclusions can be deciphered from this:

- Robust method to detect even thin edges with little to no background noise edge detected.
- Edge continuity problem is better tackled than most of the current techniques.
- Both horizontal and vertical direction best solution matrix has been taken hence direction improvement.
- The method provides with better MSE values than all the traditional techniques used for the same.
- Although computational time is an area which needs a little improvement but considering modern computational power the implementation for the same is a no big task.
- Neural networks can be trained for large datasets with proposed method to witness exponential improvement.

The researchers in future might test and develop this method for further more complex and challenging databases. Improvement in detection of edges and further that of objects, faces and places will result in better application at places like medical equipment, robots, safety and surveillance, warfare and many more. Various devices with inbuilt processing power can make use of technique to further implement real time solutions to various problems.

### *A. Future Work*

Further Parameters like Signal to Noise Ratio, PSNR and Structural Similarity Index could be used in place of MSE to performance analysis the given algorithm and tweaking decimal intensities to reach more improved performance. Also, Deep learning can be used to train neural networks for low contrast image and can super sample images for significant more probability of true edge detection. While that is computationally demanding and requires lots of pre training to reach even acceptable results but with ever increasing computing power it will be better improvement to it.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:59:29 UTC from IEEE Xplore. Restrictions apply.

## REFERENCES


[1] Wechsler, H., & Kidode, M. (1977). A new edge detection technique and its implementation. *IEEE Transactions on Systems, Man, and Cybernetics*, *7*(12), 827-836

[2] Kitchen, L., & Rosenfeld, A. (1981). Edge evaluation using local edge coherence. IEEE Transactions on Systems, Man, and Cybernetics, 11(9), 597-605.

[3] Abdou, I. E., & Pratt, W. K. (1979). Quantitative design and evaluation of enhancement/thresholding edge detectors. Proceedings of the IEEE, 67(5), 753-763.

[4] Russo, F. (1998, May). Edge detection in noisy images using fuzzy reasoning. In IMTC/98 Conference Proceedings. IEEE Instrumentation and Measurement Technology Conference. Where Instrumentation is Going (Cat. No. 98CH36222) (Vol. 1, pp. 369-372). IEEE.

[5] Dollar, P., Tu, Z., & Belongie, S. (2006, June). Supervised learning of edges and object boundaries. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) (Vol. 2, pp. 1964-1971). IEEE.

[6] Agarwal, A., & Goel, K. (2016, February). Comparative analysis of digital image for edge detection by using bacterial foraging & canny edge detector. In 2016 Second International Conference on Computational Intelligence & Communication Technology (CICT) (pp. 125-129). IEEE.

[7] Xuan, L., & Hong, Z. (2017, November). An improved canny edge detection algorithm. In 2017 8th IEEE international conference on software engineering and service science (ICSESS) (pp. 275-278). IEEE.

[8] Günen, M. A., Atasever, Ü. H., & Beşdok, E. (2017, May). A novel edge detection approach based on backtracking search optimization algorithm (BSA) clustering. In 2017 8th International Conference on Information Technology (ICIT) (pp. 116-122). IEEE.

[9] Ma, X., Liu, S., Hu, S., Geng, P., Liu, M., & Zhao, J. (2018). SAR image edge detection via sparse representation. Soft Computing, 22(8), 2507-2515.

[10] Cao, J., Chen, L., Wang, M., & Tian, Y. (2018). Implementing a parallel image edge detection algorithm based on the otsu-canny operator on the hadoop platform. Computational intelligence and neuroscience, 2018.

[11] Yang, J., Price, B., Cohen, S., Lee, H., & Yang, M. H. (2016). Object contour detection with a fully convolutional encoder-decoder network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 193-202).

[12] Arbelaez, P., Maire, M., Fowlkes, C., & Malik, J. (2010). Contour detection and hierarchical image segmentation. IEEE transactions on pattern analysis and machine intelligence, 33(5), 898-916.

[13] Lim, J. J., Zitnick, C. L., & Dollár, P. (2013). Sketch tokens: A learned mid-level representation for contour and object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3158-3165).

[14] Cherri, A. K., & Karim, M. A. (1989). Optical symbolic substitution: edge detection using Prewitt, Sobel, and Roberts operators. Applied Optics, 28(21), 4644-4648.

[15] El-Khamy, S. E., Lotfy, M., & El-Yamany, N. (2000, February). A modified fuzzy Sobel edge detector. In Proceedings of the Seventeenth National Radio Science Conference. 17th NRSC'2000 (IEEE Cat. No. 00EX396) (pp. C32-1). IEEE.

[16] Canny, J. (1986). A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence, (6), 679-698.

[17] Bustacara-Medina, C., Florez-Valencia, L., & Diaz, L. C. (2020). Improved Canny Edge Detector Using Principal Curvatures. Journal of Electrical and Electronic Engineering, 8(4), 109-116.

[18] Dhruv, B., Mittal, N., & Modi, M. (2017, October). Comparative analysis of edge detection techniques for medical images of different body parts. In International Conference on Recent Developments in Science, Engineering and Technology (pp. 164-176). Springer, Singapore.

[19] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.

[20] Laishram, R., Kumar, W. K., Gupta, A., & Prakash, K. V. (2014, January). A novel MRI brain edge detection using PSOFCM segmentation and canny algorithm. In 2014 International Conference on Electronic Systems, Signal Processing and Computing Technologies (pp. 398-401). IEEE.

[21] Salman, M. S., Kukrer, O., & Hocanin, A. (2017). Recursive inverse algorithm: Mean-square-error analysis. Digital Signal Processing, 66, 10-17.




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:59:29 UTC from IEEE Xplore. Restrictions apply.

WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Latest Advancements in Electric Vehicles

***Arhant Tewari, Aniket Jha, Hammad, Dr. JP Kesari***

Delhi Technological University

**ABSTRACT**

This paper provides an overview of the latest work of electric vehicles in the world. The paper outlines the development, evolution, and the comparison ofvarious different part of components involved in the design and manufacturing of EVs. The major components involved include battery technology, charger design, motor, steering,etc. The paper at lengthhighlights some electric vehicle prototype as a conclusion to the paper.

Keywords – Electric vehicle, AF, steering system, braking system, ABS, battery management systems, BMS, Inverter

## 1.INTRODUCTION

Electrical vehicles (EVs) are based on electric propulsion systems,no internal combustion (IC) engine is used. All the power is rooted on electric power as the primary energy source. The main advantage is the high efficiency of power conversion through its proposition system of the electric motor. Recently there has been great research and developmental work reported in both academia and industry. Commercial vehiclesare also available, and many countries have already provided variousmotive to users and manufacturers by lowering lower tax or providing tax exemption, free parking, and subsidized charging facilities. On the other hand, the hybrid electric vehicle (HEV) is another possiblealternative. It has been used widely in the past few years. Nowadays, almost all the car manufacturers have at least one model in hybrid electric vehicle. The questions that arise here is: Which vehicle will dominate the market, andis suitable for future? This paper is to examine the recent development of electric vehicle and suggest the future development in this regard.

## EV, SERIES AND PARALLEL HYBRIDS

HEV has been fostered extensively in the last decennary. It is meant to rescue the battery energy storage problems at that time. Using hybrid vehicles allows electric power to be acquired from engine. HEVs are classified into two categories: series hybrid and parallel hybrid. The engine power of the series hybrid completelydependson the battery as,all the motor power is derived from the battery. As for the parallel hybrid, both the engine and motor contributeto the total propulsion power output. The torque derived is the net sum of both the motor and engine. Also note that, motor can also used as a generator to absorb the power from engine through the battery transmission. Both series andparallel hybrids can absorb power through regeneration during braking or deceleration.

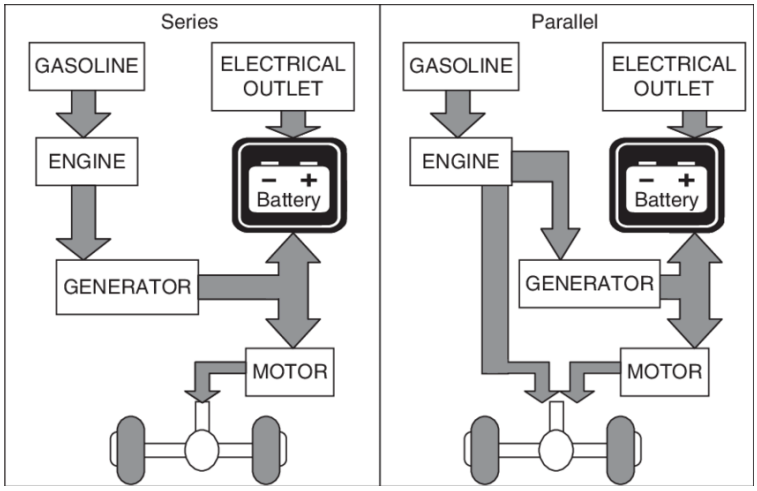



Figure 1 Series or Parallel path of an EV

Nonetheless, HEV still has emission issues which needs to be taken care of. The latest developments in the field of plug-in HEV attempts to solve some of these problems. It accepts electric power to battery via plug in from the mains. Therefore, when convenient, users may charge the battery using AC from the mains.

## KEY COMPONENTS IN EV

The electric vehicle has a rather straightforward structure. The key components are the propulsion parts shown in detail in Fig 2, the battery is the main energy storage, battery chargersare used to convert the electricity directly from mains to charge the battery. The battery voltage is DC, and I is reversed into switched-mode signal through power electronic inverter to drive the motor. The other electronic components in a vehicle can be supplied to the battery through AC-DC converter that step downs the voltage from the battery pack to voltages as low as 20-50 V.

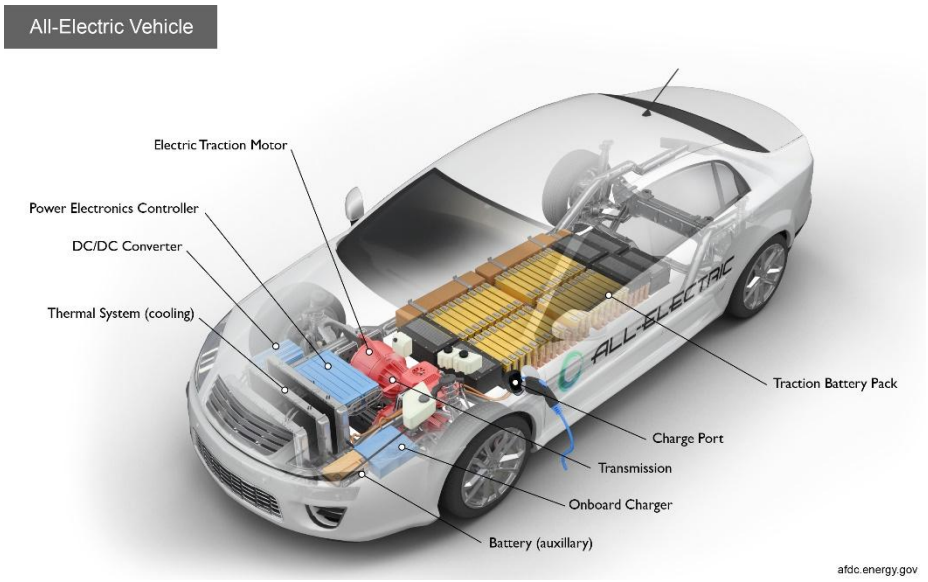



Figure 2 Key components in EV

### 1. MOTORS

There are various types of motors available for electric vehicles: DC motors, Induction motor, DC brushless motor, Permanent magnetic synchronous motor, and Switched reluctance motor.

1. **DC motors**: It is a classical and most used motor and has been used in motor control for a long time and continues to be relevant. All the power involved in the electromechanical conversions is moved to the rotor through non dynamic brushes which are in frictional contact with the copper segments of the commutator. However,it requires regular maintenance and has a shorter lifetime. Generally, it is most suitable for low power applications. It has found applications in electric wheelchairs, transporter, and micro-car. Today, almost all golf-carts are using DC motors where power levels are as low as 4KW.

2. **Induction motor:** It is one of the most popular AC motors. Moreover,it has a large market share in fluctuating speed drive application such as air-conditioning, elevator,and escalator. Many of the higher power electric vehicles, with power level requirements of more than 5kW use induction motor. Usually, a vector drive is used to provide torque and monitors the speed.

3. **Direct Drive And In Wheel Motor**

One of the major advantages of the Direct drive is that it reduces the losses incurred in the mechanical units to the drive train. The motor is directly connected to the shaft in a straight line to decreaserequirements of transmission, clutch, and gear box. The in-wheel motor is there to turn the rotor inside out and is connected to the wheel's rim and the tire,however,there is no gear box and drive shaft. Fig 3 shows the in-wheel motoralso called the wheel-hub motor. Its main advantage over others isthat is independent of the control of each wheel. Fig 4 shows the 4-wheele drive vehicle. Each of the wheels can work at any speed and direction which not only aids in parallel parking but alsothe Anti-lock braking system (ABS) can be implemented effortlessly by this technology. Moreover, it has shown that it can successfully prevent spinouts. Many different types of motor can be utilized for in-wheel motor. The prominent one being switched-reluctance types. Its phase-winding is independent and therefore the fault clemency is much more advanced than any other. There is no permanent magnet in the motor, which reduces any interference by permanent field and the fluctuation of the permanent magnetic materials.

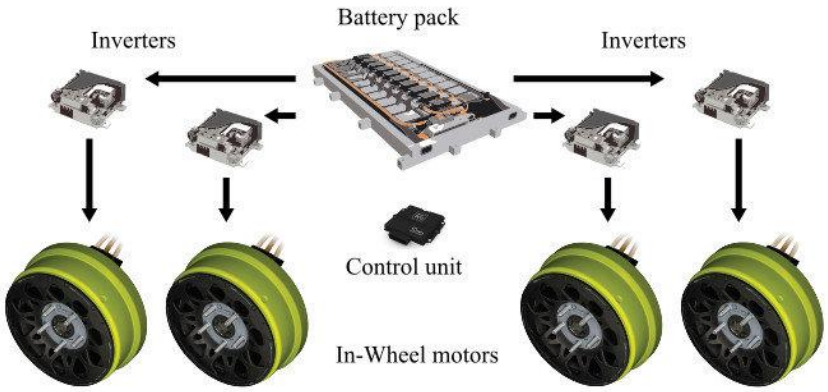



Figure 3 The in- wheel motor

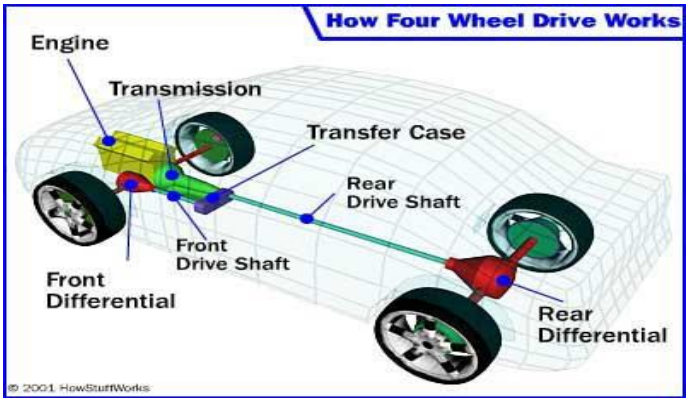



Figure 4Four wheel drive

***BATTERY CHARGING SYSTEMS***

**1. General charger:**

The charger required for the battery system for slow charging or fast charging are both required to handle high power levels. Themost used converter is the H-Bridge convertor shown in Fig 5.The converter is desirable for its efficiency and has found application in charger and DC-DC converters.

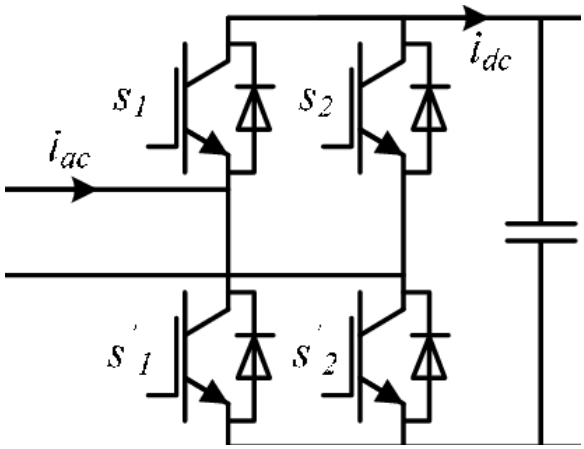



Figure 5 H bridge convertor for charger

**2. Ultra-capacitor charger:**

The voltage on the ultra-capacitor charger various from the full voltage to zero nits' and the energy storage varies from full to zero. It is important to note that, this is different from the battery as its voltage varies only within 25% limit. The capacitor voltage is internal point and is not in contact with the users. A tapped converter must be used as it will exhibit higher efficiency for power conversion. The efficiency of the power converter is higher than the transformer-isolated version and the structure is straightforward.

**3. Battery management systems:**

It is also referred to as BMS. The battery system is formed by several battery cells. They are connected in parallel or series according to the design. Each cell should be monitored and regulated. The conditionedobservations include the voltage, current and temperature checks. The measured parameters are used to provide the decision framework for the system control and protection. Two limits are provided- the state of charge (SoC) and the State of Health (SoH). SoC is like the oil tank meter that provides the battery a charging condition and is measured by the information of voltage and current. The SoH is to account for the health or aging condition. There are a few definitions but the prominent one is:

SoH= (Nominal Capacity-Loss of Capacity)/ (Nominal Capacity)

Cell balancing is to ensure that each cell is operating under the same condition, or a regulation is used to charge or discharge each cell by the balancing control. This avoids the overloading of a particular cell. Energy management systems for ultra-capacitor systems,is made by a fewcapacitors in a combination with other energy storage devices such as battery. The same conditioning monitoring and management system will be used.

## CONCLUSION

This paper discusses the latest developments in electric vehicles. The paper first describes general structure and discusses the energy storage. It then extends to the future vehicle components. The paper provides an overview of the recent EV work.

**REFERENCES**

[1] Jones, W.D., "Hybrids to the rescue [hybrid electric vehicles]", IEEE Spectrum, Vol. 40(1),

[2] Jones, W.D., "Take this car and plug it [plug-in hybrid vehicles]", Spectrum, IEEE, Vol. 42, Issue 7, July 2005,

[3] HyunjaeYoo; Seung-Ki Sul; Yongh Park; JongcanJeong, "System Integration and Power-Flow Management for a Series Hybrid Electric Vehicle Using Supercapacitors and Batteries", IEEE Trans. on Industry Applications, Vol. 44, Issue 1, Jan.-Feb. 2008, pp. 108 – 114.

[4] Haddoun, A.; Benbouzid, M. E. H.; Diallo, D.; Abdessmed, R.; Ghouili, J.; Srairi, K., "A Loss-Minimization DT Scheme for EV Induction Motors", IEEE Trans on Vehicular Technology, Vol. 56(1),

[5] Jinyun Gan; Chau, K.T.; Chan, C.C.; Jiang, J.Z., "A new surface-inset, permanent-magnet, brushless DC motor drive for electric vehicles", IEEE Transactions on Magnetics, Vol. 36, Issue 5, Part 2, Sept 2000, pp. 3810 – 3818.

[6] Chau, K.T.; Chan, C.C.; Chunhua Liu, "Overview of Permanent-Magnet Brushless Drives for Electric and Hybrid Electric Vehicles", IEEE Trans. on Industrial Electronics, Vol. 55, Issue 6, June 2008, pp. 2246 – 2257.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/363145370

# Linearity analysis of T-gate HEMT with graded back-barrier for wireless applications

**Conference Paper** · July 2022

DOI: 10.1109/CONECCT55679.2022.9865687

CITATIONS
0

READS
21

**3 authors:**

Megha Sharma
Delhi Technological University
**8** PUBLICATIONS **18** CITATIONS
SEE PROFILE

Bhavya Kumar
Delhi Technological University
**13** PUBLICATIONS **73** CITATIONS
SEE PROFILE

Rishu Chaujar
Delhi Technological University
**247** PUBLICATIONS **1,885** CITATIONS
SEE PROFILE

**Some of the authors of this publication are also working on these related projects:**



Novel Mosfet Design for Digital Applications View project



Modelling and Simulation of GAA FinFET View project

All content following this page was uploaded by Bhavya Kumar on 01 September 2022.

The user has requested enhancement of the downloaded file.

# Linearity analysis of T-gate HEMT with graded back-barrier for wireless applications

Megha Sharma
Department of Applied Physics
Delhi Technological University
Delhi, India
meghasharma_phd2k18@dtu.ac.in

Bhavya Kumar
Department of Applied Physics
Delhi Technological University
Delhi, India
bhavyakumar_phd2k18@dtu.ac.in

Rishu Chaujar
Department of Applied Physics
Delhi Technological University
Delhi, India
chaujar.rishu@dtu.ac.in

2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) | 978-1-6654-9781-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CONECCT55679.2022.9865687

***Abstract* — This paper studied the linearity performance of double heterostructure T-gate HEMT with graded back-barrier using the Silvaco Atlas simulator. The proposed device has the graded back-barrier feature, which improves the electron confinement at the GaN/AlGaN junction by increasing the conduction band discontinuity. A precise comparison has been made of the linear and nonlinear parameters of the proposed structure with the conventional one. The linear change of Al composition in the back-barrier layer effectively improved the linearity parameters. All the linear and nonlinear parameters such as $g_{m2}$, $g_{m3}$, VIP2, VIP3, IIP3, IMD3, and 1-dB compression point were investigated. Besides, the effect of channel length variation on the device's performance has also been investigated. The simulated results reveal that 30 nm channel length is the optimum value to uplift the linearity performance.**

***Keywords — linearity, T-gate, heterostructure, polarization.***

## I. Introduction

The expansion of the market for wireless communication systems and fourth-generation (4G) mobile communication is coming closer to its limitation. There is an extensive requirement for high-power RF devices with commendable linearity and low intermodulation distortion. Various power amplifier technologies compete in the market like GaAs MESFETs, SiC MESFETs, GaAs HBT, Si-LDMOS, and GaN HEMTs. However, for high-power applications, GaN-based devices are the best candidates. The GaN material has remarkable properties compared to the competing materials, i.e., high electron mobility, high saturation velocity, and large bandgap [1-3]. Although various approaches are available for improving linearity at the system level, the sophisticated circuitry required by these methods makes it difficult to use them. Therefore, the increase of linearity at the transistor level is more suited for wireless applications. This also applies to the research of linearity performance at the device level regarding critical design parameters. Therefore, to obtain greater linearity performance and increased distortion characteristics in future wireless applications, new FET designs are necessary to be developed. Several research works have already been done to upgrade the efficiency of AlGaN/GaN HEMT [1-8]. However, these devices have a strain issue at AlGaN/GaN junction, degrading the device's performance. Recently InAlN/AlN HEMT resolved these issues and obtained extensive attention in the submillimeter-wave application. Although the InAlN/AlN T-Gate HEMTs have numerous advantages over AlGaN/GaN HEMT, they still have one major problem of two-dimensional electron gas (2DEG) confinement in the channel region, degrading the device mobility and reliability [9]. The back-barrier (BB) engineering is the most acceptable way for the 2DEG confinement, which blocks the carrier's flow in the buffer region and forms the double heterojunction structure. Several researchers have published the back-barrier effect on HEMT structure [10-11]. The insertion of the BB layer enhances the 2DEG confinement to some extent. Still, it produces the unwanted channel at the AlGaN/GaN heterojunction, adversely affecting the device's performance. A graded back-barrier (GBB) layer can solve the above problem. B. Du et al. reported the experimental paper on the influence of graded buffer on the performance of InGaAs photodetector [12], D. Gonzalez et al. studied the graded buffer for strain relaxation by transmission electron microscopy [13], S. Li et al. exhibited the experimental work on graded AlGaN layer by molecular beam epitaxy [14]. V. Krishnamoorthy et al. reported the stack of different composition InGaAs on GaAs using the molecular beam epitaxy technique [15]. C. Yao et al. reported the N-face light-emitting diode with a graded AlGaN electron blocking layer [16]. Y. Gu et al. studied the graded AlGaN layer's effect on the device's RF performance [17]. To the author's best knowledge, no work has been published on linearity analysis of T-gate HEMT with a graded back barrier.

This paper reported the linearity parameters of double heterostructure T-gate with graded back-barrier HEMT. The linear reduction of Aluminium in the BB layer effectively improves the linearity parameters. The linear characteristics of the proposed structure are matched with the fixed composition AlGaN Back barrier HEMT. The outcome of channel length deviation on the device's linearity performance has been studied.

## II. Device Structure and Experimental Calibration

Fig. 1 depicts the structure of the proposed device. The structure comprises an InAlN barrier layer with 10 nm thickness, 1 nm AlN spacer layer, and 30 nm GaN channel layer. In conventional HEMT, fixed composition (x = 0.1) AlGaN layer is used, whereas in the proposed structure, a 10 nm graded back-barrier layer is considered in which Al composition changes linearly from 0.1 near the main channel to 0 at the bottom of the layer. As a result, the unwanted channel at the AlGaN/GaN buffer interface has been reduced. The 2 µm unintentionally doped (UID) GaN was used as a buffer layer, grown over a sapphire substrate. A 70 nm T-shaped gate formed the Schottky contact. The drain and source regions are doped with silicon to achieve

978-1-6654-9781-7/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on September 01,2022 at 05:44:33 UTC from IEEE Xplore. Restrictions apply.

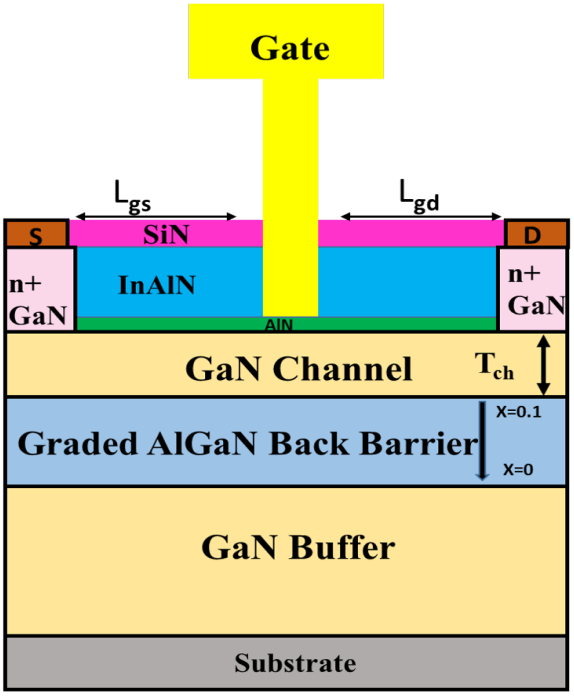



Fig. 1. Proposed device structure.

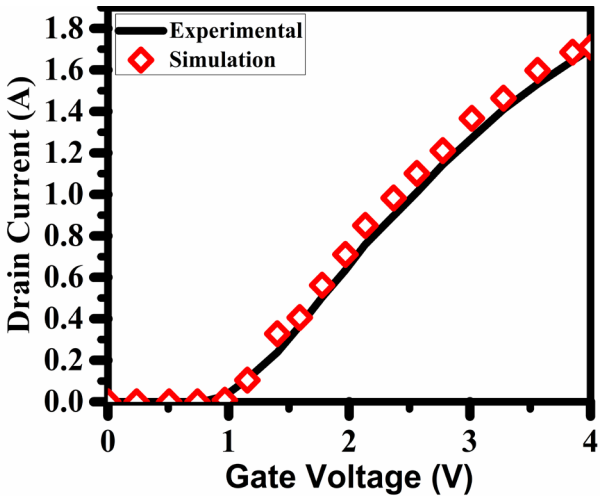



Fig. 2. Comparison of experimental and simulation result.

smaller contact resistance. The device is passivated by the SiN layer [18], and the length of the gate to drain and gate to the source is kept at 1.5 µm and 400 nm.

The suggested tool has been modeled in a computer simulation with the aid of the Atlas-Silvaco simulator [19]. During the simulation, several different physical models, including the Fermi-Dirac, Shockley-Read-Hall, low field mobility, and polarization models, were considered. When calculating the electron trapping mechanism for each trap level, the SRH model is the one that is used. To investigate the scattering mechanism, [19] we use the high field mobility model and the low field mobility model. Those above mathematical and physical models have been fine-tuned using the experimentation carried out by R. Wang et al. [20]. The computed and experimental $I_d$-$V_g$ curves of a 150 nm recessed T-gate HEMT is shown in Fig. 2. It has been noted that the results of the experiments show a high degree of unity with the results of the simulations.

## III. Results and Discussion

### A. Comparison of performance between GBB-HEMT and BB-HEMT

In this section linearity performance of GBB-HEMT and BB-HEMT has been investigated. The linearity assessment is the primary requirement in the RF system, which ensures the higher-order harmonics and intermodulation are lower at the output. The low noise and linearity are the feasible solutions for obtaining superior device performance. Thus, there is an essential requirement for innovative design with high linearity and transconductance. Fig. 3(a) exhibits the drain current characteristics of both devices. The drain current of the GBB-HEMT at $V_{gs}$ = 1.5 V is 0.10 A which is 25% greater than the BB-HEMT. The BB layer produces the negative polarization field, which significantly confines the 2DEG in the channel. The GBB layer reduces the negative polarization field and weaker the 2DEG confinement, which slightly shifts the threshold voltage in the negative direction and significantly enhances the drain current density. Fig. 3(b) displays the transconductance characteristics of both devices. A higher transconductance peak for GBB-HEMT has been observed at $V_{gs}$ = 1.2 V. The GBB layer reduces the adverse effect of the parasitic channel and significantly improves transconductance characteristics.

The study of nonlinear parameters such as second-order and third-order transconductance ($g_{m2}$ and $g_{m3}$) is essential to reduce device distortion and enhance the device's linearity. The nonlinear parameters are the major source of distortion shown in Eq. (1) and (2) [21]. These nonlinear parameters should be as small as possible to obtain high linearity.

$$g_{m2} = \delta^2 I_d / \delta V^2_{gs} \quad (1)$$

$$g_{m3} = \delta^3 I_d / \delta V^3_{gs} \quad (2)$$

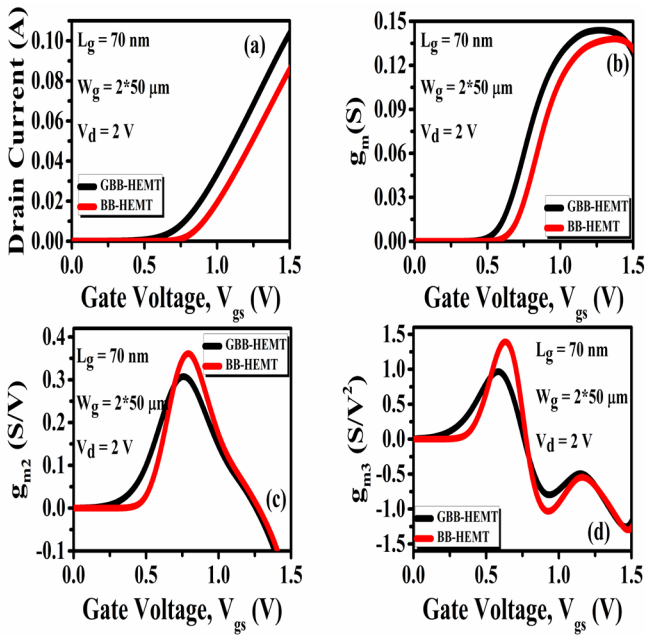



Fig. 3. (a) Drain current (b) transconductance (c) nonlinear parameters $g_{m2}$ (d) $g_{m3}$ for BB and GBB HEMT.

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on September 01,2022 at 05:44:33 UTC from IEEE Xplore. Restrictions apply.

Fig. 3(c) and (d) show the effect of GBB on the amplitude of nonlinear parameters. The linear reduction of Al composition in the back-barrier reduces the amplitude of nonlinear parameters ($g_{m2}$ and $g_{m3}$) and improves the device's linearity. In Fig. 3(c), when the $V_{gs}$ is between 0.5 V to 1 V, there is an increase in $g_{m2}$, and after that, there is a downward swing in $g_{m2}$, which effectively enhances the linearity parameter by improving the VIP2 value. At higher $V_{gs}$, $g_{m2}$ becomes more negative. This is because, at higher $V_{gs}$, 2DEG reaches the equilibrium value, and the drain current becomes constant, leading to a decrease in transconductance at higher $V_{gs}$. Fig. 3(d) depicted the variation of $g_{m3}$ with gate voltage. This parameter determines the DC singularity bias point for better device performance. Fig. 3(d) exhibits that a graded back-barrier effectively reduces the amplitude of $g_{m3}$ and enhances the proposed device's linearity performance.

Linearity plays a vital role in wireless applications, so it is important to analyze the device's linearity parameters in depth. The linearity parameters investigated in this paper are expressed using Eq. (3-6) [22]. The third-order voltage intercept point (VIP3) and second-order voltage intercept point (VIP2) denote the extrapolated input voltage amplitude at which 3rd and 2nd order harmonics equal the fundamental tone in device current ($I_{ds}$). The above figure of merits (FOMs) should be as high as possible to achieve high linearity.

$$\mathrm{VIP2} = 4 \times (g_m/g_{m2}) \tag{3}$$

$$\mathrm{VIP3} = \sqrt{24 \times (g_m/g_{m3})} \tag{4}$$

$$\mathrm{IIP3} = (2/3) \times \left(g_m/(g_{m3}R_s)\right) \tag{5}$$

$$1-\mathrm{dB\ compression\ point} = 0.22 \times \sqrt{(g_m/g_{m3})} \tag{6}$$

The VIP2 and VIP3 linearity parameters for GBB-HEMT and BB HEMT are depicted in Fig. 4(a-b). VIP2 and VIP3 parameters are contrariwise dependent on the $g_{m2}$ and $g_{m3}$, i.e., the inferior the value of $g_{m2}$ and $g_{m3}$, the higher the linearity performance will be. The higher values of VIP2 (11.5 V) and VIP3 (5.2 V) are obtained at gate voltage $V_{gs}$ = 1.2 V for GBB-HEMT, which is two times greater than the BB-HEMT (7.6 V and 1.3 V). The third-order input intercept point (IIP3) and 1-dB compression points are also investigated for the proposed structure. Fig. 4(c) depicted the graph of IIP3 as a function of gate voltage. In Eq. (5), the IIP3 parameter depends on the transconductance and, inversely, on the third-order transconductance. As we have already discussed for GBB-HEMT, the higher value of $g_m$ and lower value of $g_{m3}$ were achieved, which effectively enhanced the IIP3 parameter. The higher value of IIP3 is obtained 46.1 dBm at $V_{gs}$ = 0.8 V for GBB-HEMT, which is 15% higher than the BB-HEMT (40 dBm). Fig. 4(d) exhibits both devices' bar graphs of the 1-dB compression point. A greater value of 1-dB point is observed for GBB-HEMT, which is 31% higher than the BB-HEMT and emphasizes the superiority in the linearity performance of GBB-HEMT for set-up in microwave applications. The simulated results of both the structures are tabulated in Table I.

TABLE I
COMPARISON OF GBB-HEMT AND BB- HEMT

| Performance Parameter | BB-HEMT (conventional) | GBB-HEMT (Proposed) |
|---|---|---|
| **Drain current (A)** | 0.11 | 0.13 |
| **$g_m$ (mS/mm)** | 0.13 | 0.14 |
| **$g_{m2}$ (S/V)** | 0.4 | 0.3 |
| **$g_{m3}$ (S/V$^2$)** | 1.5 | 1 |
| **VIP2 (V)** | 11.2 | 14.5 |
| **VIP3 (V)** | 1.3 | 7.5 |
| **IIP3 (dBm)** | 41 | 46 |
| **1-dB compression point (dB)** | -14.5 | -11 |

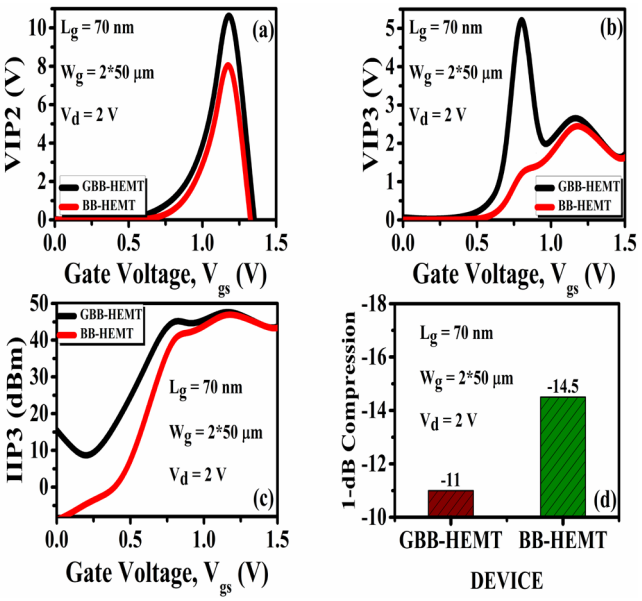



Fig. 4. (a) VIP2 (b) VIP3 (c) IIP3 (d) 1-dB compression point for both the devices.

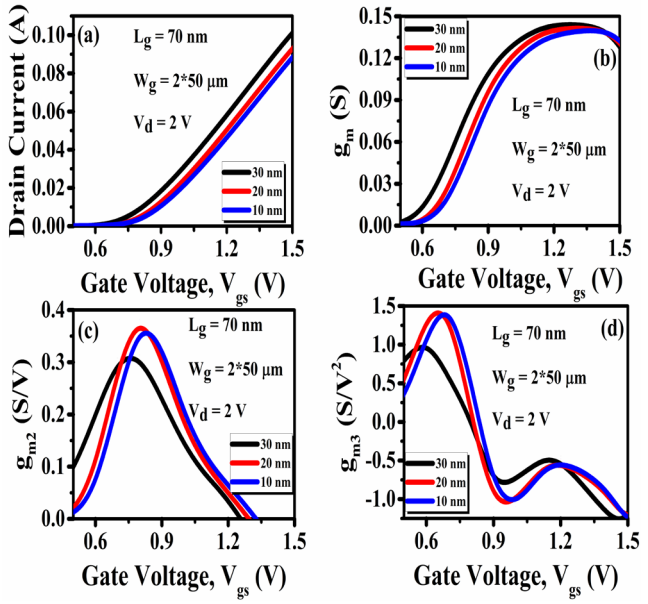



Fig. 5. (a) Drain current (b) $g_m$ (c) nonlinear parameters $g_{m2}$ (d) $g_{m3}$ for different channel length.

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on September 01,2022 at 05:44:33 UTC from IEEE Xplore. Restrictions apply.

### B. *Impact of channel length variation on linearity performance of GBB-HEMT*

In this section, the consequence of channel length ($T_{ch}$) on the linearity matrices of GBB-HEMT has been discussed. The drain current and transconductance performance for various channel lengths have been shown in Fig. 5(a-b). It is witnessed that increasing the $T_{ch}$ enhances the drain current density (1.0 A/mm) because the depth of the quantum well at the heterojunction increases, which further increases the electron concentration. The increase in $g_m$ parameter on increasing the channel length is observed in Fig. 5(b). This improvement is owing to the low scattering of 2DEG and high mobility. As we have already discussed, nonlinear parameters adversely affect the linearity of the device. The higher the nonlinear parameters, the higher the distortion, which reduces the device's RF performance. Fig. 5(c-d) shows that the amplitude of nonlinear parameters is reduced by increasing the channel length from 10 nm to 30 nm. The peak of $g_{m2}$ decreases for 30 nm channel length, which enhances the linearity of the device. Third-order transconductance is essential in reducing the device's linearity, as it is responsible for wasting useful powers. Fig. 5(d) shows that the third-order transconductance reduces on increasing the channel length, which effectively enhances the device's linearity performance.

Fig. 6(a-b) shows the second-order voltage intercept point and third-order voltage intercept point with gate bias for different channel lengths. It is witnessed that on increasing the $T_{ch}$, the linearity parameters increase. The peak value of VIP2 and VIP3 increase by 62% and 50%, respectively, for 30 nm $T_{ch}$ (11.5 V and 5.2 V) compared to 10 nm $T_{ch}$ (8.9 V and 2.6 V). The VIP2 and VIP3 parameters directly depend upon $g_m$ and inversely upon $g_{m2}$ and $g_{m3}$, due to the high mobility and low scattering disorder of 2DEG, higher $g_m$ and lower $g_{m2}$ and $g_{m3}$ have been obtained for 30 nm $T_{ch}$.

Fig. 6(c) exhibits a third-order input intercept point variation for variable $T_{ch}$. The increment in the IIP3 parameter is observed with increasing the channel length, the higher peak of IIP3 observed for 30 nm $T_{ch}$. Fig. 6(d) exhibits the bar graph of the 1-dB compression point for different channel lengths. For 30 nm channel length, a greater value of 1-dB point (-11 dB) at lower gate bias $V_{gs}$ = 0.7 V has been observed, which is 18% higher than the 10 nm (-13 dB). This highlights how superior performance is achieved for GBB-HEMT with a channel length of 30 nm for operating in microwave circuits in terms of its linearity performance.

## IV. Conclusion

This work investigated the GBB effect on the linearity performance of the proposed structure. The GBB layer reduces the parasitic concentration and improves the drain current density by 25% and transconductance by 10%. The linearity parameters like VIP3 increased twice, whereas VIP2 increased by 50%, IIP3 by 15%, and 1-dB point by 42% for GBB-HEMT. The consequence of channel length thickness on the proposed device linearity performance is also studied. The simulated results showed that the 30 nm channel length effectively improved the linearity performance of the GBB-HEMT.

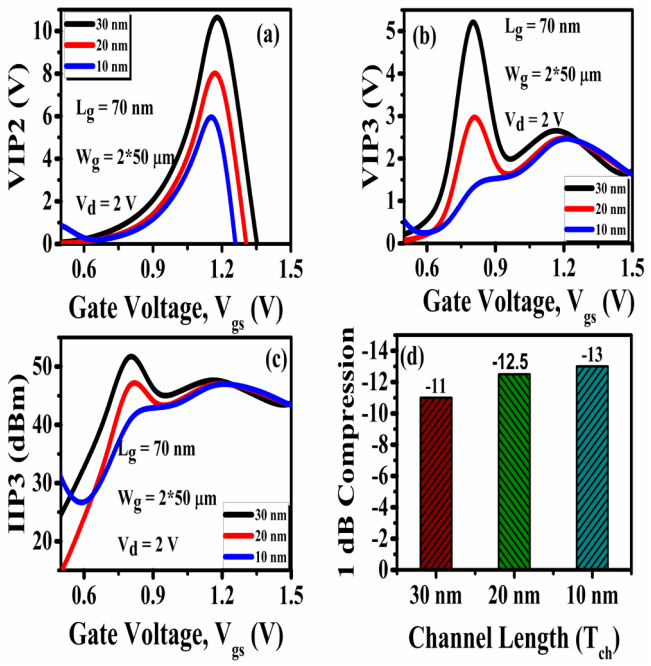



Fig. 6. (a) VIP2 (b) VIP3 (c) IIP3 (d) 1-dB compression point for different channel lengths.

## References


[1] U. K. Mishra, P. Parikh, and Y. F. Wu, "AlGaN/GaN HEMTs - An overview of device operation and applications," Proc. IEEE, vol. 90, no. 6, pp. 1022-1031, 2002, DOI: 10.1109/JPROC.2002.1021567.

[2] M. Sharma and R. Chaujar, "Ultrascaled 10 nm T-gate E-mode InAlN/AlN HEMT with polarized doped buffer for high power microwave applications," Int. J. RF Microw. Comput. Aided Eng., vol. 32, no. 4, pp. 23057, 2022, DOI:10.1002/mmce.23057.

[3] M. Sharma and R. Chaujar, "Design and Investigation of Recessed-T-Gate Double Channel HEMT with InGaN Back Barrier for Enhanced Performance," Arab. J. Sci. Eng., vol. 47, pp. 1109-1116, 2022, DOI: 10.1007/s13369-021-06157-7.

[4] Z. H. Feng, Y. G. Zhou, S. J. Cai, and K. M. Lau, "Enhanced thermal stability of the two-dimensional electron gas in GaN/AlGaN/GaN heterostructures by $Si_3N_4$ surface-passivation-induced strain solidification," Appl. Phys. Lett., vol. 85, no. 22, pp. 5248–5250, 2004, DOI: 10.1063/1.1828231.

[5] Z. Lin, H. Kim, J. Lee, and W. Lu, "Thermal stability of Schottky contacts on strained AlGaN/GaN heterostructures," Appl. Phys. Lett., vol. 84, no. 9, pp. 1585–1587, 2004, DOI: 10.1063/1.1650875.

[6] S. P. Kumar, A. Agarwal, R. Chaujar, and R. S Gupta, "Device linearity and intermodulation distortion comparison of dual material gate and conventional AlGaN/GaN high electron mobility transistor," Microelectron., vol. 51, pp. 587–596, 2011, DOI: 10.1016/j.microrel.2010.09.033.

[7] Shaveta, H. M. M. Ahmed, and R. Chaujar, "Rapid detection of biomolecules in a dielectric modulated GaN MOSHEMT," J. Mater. Sci. Mater. Electron., vol. 31, pp. 16609–166152, 2020, DOI: 10.1007/s10854-020-04216-7.

[8] S. Sano, K. Ebihara, T. Yamamoto, T. Sato, and N. Miyazawa, "GaN HEMTs for wireless communication," SEI Tech. Rev., no. 86, pp. 65–70, 2018.

[9] M. A. Khan, J. N. Kuznia, J. M. Van Hove, N. Pan, and J. Carter, "Observation of a two-dimensional electron gas in low-pressure metalorganic chemical vapor deposited GaN-$Al_xGa_{1-x}N$ heterojunctions," Appl. Phys. Lett., vol. 60, no. 24, pp. 3027–3029, 1992, DOI: 10.1063/1.106798.

[10] P. Murugapandiyan, S. Ravimaran, and J. William, "DC and microwave characteristics of $L_g$ 50 nm T-gate InAlN/AlN/GaN HEMT for future high power RF applications," AEU - Int. J. Electron. Commun., vol. 77, pp. 163–168, 2017, DOI: 10.1016/j.aeue.2017.05.004.


Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on September 01,2022 at 05:44:33 UTC from IEEE Xplore. Restrictions apply.

[11] P. Murugapandiyan, S. Ravimaran, J. William, and K. M. Sundaram, "Design and analysis of 30 nm T-gate InAlN/GaN HEMT with AlGaN back-barrier for high power microwave applications," Superlattices Microstruct., vol. 111, pp. 1050–1057, 2017, DOI: 10.1016/j.spmi.2017.08.002.

[12] B. Du, Y. Gu, Y. G. Zhang, X. Y. Chen, S. P. Xi, Y. J. Ma, W. Y. Ji, Y. H. Shi, X. Li, and H. M. Gong, "Effects of continuously or step-continuously graded buffer on the performance of wavelength extended InGaAs photodetectors," J. Cryst. Growth, vol. 440, pp. 1–5, 2016, DOI: 10.1016/j.jcrysgro.2016.01.016.

[13] D. González, D. Araújo, S. I. Molina, A. Sacedón, E. Calleja, and R. García, "Step-graded buffer layer study of the strain relaxation by transmission electron microscopy," Mater. Sci. Eng. B, vol. 28, no. 1–3, pp. 497–501, 1994, DOI: 10.1016/0921-5107(94)90114-7.

[14] S. Li, M. Ware, J. Wu, P. Minor, Z. Wang, Z. Wu, Y. Jiang, and G. J. Salamo, "Polarization induced pn-junction without dopant in graded AlGaN coherently strained on GaN," Appl. Phys. Lett., vol. 101, no. 12, 2012, DOI: 10.1063/1.4753993.

[15] V. Krishnamoorthy, Y. W. Lin, and R. M. Park, "Application of ''critical compositional difference'' concept to the growth of low dislocation density ($<10^4/cm^2$) $In_xGa_{1-x}As$ ($x \leqslant 0.5$) on GaAs," J. Appl. Phys., vol. 72, no. 5, pp. 1752–1757, 1992, DOI: 10.1063/1.351699.

[16] C. Yao, G. Yang, Y. Li, R. Sun, Q. Zhang, J. Wang, and S. M. Gao, "Investigation of N-face AlGaN ultraviolet light-emitting diodes with composition-varying AlGaN electron blocking layer," Opt. Quantum Electron., vol. 48, no. 1, pp. 1–9, 2016, DOI: 10.1007/s11082-015-0345-5.

[17] Y. Gu, D. Chang, H. Sun, J. Zhao, G. Yang, Z. Dai, and Y. Ding, "Theoretical Study of InAlN/GaN High Electron Mobility Transistor (HEMT) with a Polarization-Graded AlGaN Back-Barrier Layer," Electronics, vol. 8, pp. 1–9, 2019, DOI: 10.3390/electronics8080885.

[18] D. S. Lee, X. Gao, S. Guo, D. Kopp, P. Fay, and T. Palacios, "300-GHz InAlN/GaN HEMTs With InGaN Back Barrier," IEEE Electron Device Lett., vol. 32, no. 11, pp. 1525–1527, 2011, DOI: 10.1109/LED.2011.2164613.

[19] ATLAS User's Manual, SILVACO International, CA, USA, 2016.

[20] R. Wang, P. Saunier, X. Xing, C. Lian, X. Gao, S. Guo, G. Snider, P. Fay, D. Jena, and H. Xing, "Gate-recessed enhancement-mode InAlN/AlN/GaN HEMTs with 1.9-A/mm drain current density and 800-ms/mm transconductance," IEEE Electron Device Lett., vol. 31, no. 12, pp. 1383–1385, 2010, DOI: 10.1109/LED.2010.2072771.

[21] B. Kumar and R. Chaujar, "Numerical Study of JAM-GS-GAA FinFET: A Fin Aspect Ratio Optimization for Upgraded Analog and Intermodulation Distortion Performance," Silicon, vol. 14, pp. 309-321, DOI:10.1007/s12633-021-01395-8.

[22] B. Kumar and R.Chaujar, "TCAD Temperature Analysis of Gate Stack Gate All Around (GS-GAA) FinFET for Improved RF and Wireless Performance," Silicon, vol. 13, pp. 3741-3753, 2021, DOI: 10.1007/s12633-021-01040-4.

Authorized licensed use limited to: Jaypee Insituite of Information Technology-Noida Sec 128 (L3). Downloaded on September 01,2022 at 05:44:33 UTC from IEEE Xplore. Restrictions apply.
View publication stats

The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/0368-492X.htm



# Nexus of knowledge management and organizational performance: a cross-country study of China and Pakistan higher educational institutes

Asha Thomas
*Department of Management, Jagan Institute of Management Studies, Delhi, India and*
*Delhi Technological University, Delhi, India*
M. Saleem Ullah Khan Sumbal
*Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, China*
Mehwish Malik
*School of Engineering and Built Environment, Griffith University, Brisbane, Australia, and*
Umar Farooq Sahibzada
*Asia-Australia Business College, Liaoning University, Shenyang, China*

Received 27 August 2021
Revised 18 December 2021
6 April 2022
13 July 2022
8 August 2022
Accepted 11 August 2022

## Abstract

**Purpose** – The study explores the impact of knowledge management (KM) enablers, i.e. trust and organizational climate, on KM processes. The study further observes the indirect relationship of KM processes on organizational performance via the mediating role of knowledge workers' satisfaction in cross-cultural settings.

**Design/methodology/approach** – This research used a survey of 784 educational and administrative personnel from higher education institutions (HEIs) in Pakistan and China. Smart Partial Least Square (PLS) 3.2.9 was employed to perform the analysis.

**Findings** – The result shows that trust and organizational climate influences KM processes, and these KM processes, in turn, impact organizational performance via the partial mediating effect of knowledge worker satisfaction (KWS) in Pakistan. The multi-group analysis confirmed the substantial differential effect of KM processes on KWS in culturally different HEIs. At the same time, the study's overall sample substantiated full mediation in China. Furthermore, the impact of KM processes on organizational performance did not substantiate in China.

**Practical implications** – Outcomes of this research affirm KM university practice and recommend how higher education academics and administrators prioritize trust, organizational climate, KM processes, and KWS while strengthening organizational performance in a culturally different environment.

**Originality/value** – A lack of research ascertains the inter-relationship between trust, organizational climate, KM processes, KWS, and organizational performance in culturally different environments. This is one of the initial studies that examine the relationship between trust, organizational climate, KM processes, KWS, and organizational performance in HEIs. The study empirically examines the inter-relationships among these variables and enlightens insights into the current literature by immediately investigating the mediating role of KWS in culturally different environments.

**Keywords** Trust, Organizational climate, KM processes, Knowledge worker satisfaction, Organizational performance, Cross-culture

**Paper type** Research paper

Kybernetes
© Emerald Publishing Limited
0368-492X
DOI 10.1108/K-08-2021-0768

## 1. Introduction

Higher education institutions (HEIs) are persuaded to create, acquire, store, share, and apply knowledge (Laal, 2011; Sahibzada *et al.*, 2020b). The metamorphic and reframing shift in higher education has influenced and surpassed the old, stale, and decomposed teaching customs (Ramjeawon and Rowley, 2018; Veer-Ramjeawon and Rowley, 2019). As a result, HEIs scramble to extract and disseminate novel knowledge (Bano and Taylor, 2015). HEIs are important for knowledge creation, acquisition, storage, distribution, and utilization, contributing to social, economic, and technological advancement (Trivella and Dimitrios, 2015; Iqbal *et al.*, 2019). The administration and faculty of HEIs will improve their progression and amenities with the help of knowledge, i.e. education, knowledge, study, curriculum growth, administration, and planned expansion (Ahmad *et al.*, 2017; Zutshi *et al.*, 2021).

Knowledge management (KM) is critical for improving collaboration and research that drives organizational performance (Ramjeawon and Rowley, 2018; Sahibzada *et al.*, 2020a; Thomas and Gupta, 2021). Despite the importance of knowledge and KM in university provision, there has been very little recognition that universities are knowledge-intensive institutions (Schmitz *et al.*, 2014; Zutshi *et al.*, 2021), and a lack of KM and research in this area is a significant deficiency in this regard (Iqbal *et al.*, 2019).

Several researchers have defined and examined organizations' KM skills regarding KM processes and enablers (Iqbal *et al.*, 2019; Rehman and Iqbal, 2020; Rios-Ballesteros and Fuerst, 2021). KMs are actions related to KM processes that improve the HEIs reasonable benefits (Barley *et al.*, 2018; Iqbal *et al.*, 2019). Despite the existence of such a significant indication of the role of KM in the current literature, insufficient research has verified the association between KM enablers and the successful implementation of KM processes in HEIs (Fullwood and Rowley, 2017; Iqbal *et al.*, 2019; Shafait *et al.*, 2021), particularly in the context of emerging and developing countries in the higher education sector (Fullwood and Rowley, 2017; Iqbal *et al.*, 2019). Furthermore, KM is more likely to lead to more radical progress in KM processes, such as assurance in the form of trust (Koohang *et al.*, 2017; Lei *et al.*, 2019) and organizational climate (Al-Abdullat and Dababneh, 2018), both of which improve KM processes (Thani and Mirkamali, 2018; Sahibzada *et al.*, 2020c). On a fundamental level, trust is directly related to KM (Mcneish and Mann, 2010). Trust is evident as a factor and an outcome of interpersonal relationships (Yasir and Majid, 2017; Sahibzada *et al.*, 2020c; Thomas and Gupta, 2021). Trust strengthens people's relationships. These above-stated associations provide additional evidence to trust each other, which is the foundation for refining individuals' KM activities (Holste and Fields, 2010; Thomas and Paul, 2019). Consequently, scholars urge an empirical study to enhance the role of organizational characteristics (i.e. trust and organizational climate) in the successful implementation of KM processes in HEIs (Muqadas *et al.*, 2017). Thus, the first main objective of this present study is to fill the gap by including trust and organizational climate as important enablers of KM processes.

Second, the direct effect mechanism of KM processes on OP remains unclear (Iqbal *et al.*, 2019; Delshab *et al.*, 2022). Preliminary empirical studies have studied the mechanism of KM processes impact on HEIs outcomes (Ahmad *et al.*, 2017; Iqbal *et al.*, 2019). Furthermore, Iqbal *et al.* (2019) stressed the need to investigate the mediating variables. Current research on the KM processes and OP relationship in knowledge-intensive industries has suggested examining knowledge workers' satisfaction as an intervening variable among KM processes and OP (Shujahat *et al.*, 2018; Sahibzada *et al.*, 2020a, b). Knowledge workers' satisfaction as an intervening variable between KM and organizational performance is mostly neglected because the authors could not identify any evidence linking knowledge workers' satisfaction to OP improvements, specifically in cross-culture studies. Despite increased studies on KWS, its significance in expanding OP is rarely discussed in educational cross-country settings (Bratianu and Bejinaru, 2017; Chatterji and Kiran, 2017; Shujahat *et al.*, 2018).

Moreover, knowledge is the primary contribution of knowledge workers; hence, KM's optimal providing of knowledge-to-knowledge workers boosts their satisfaction without question (Kianto *et al.*, 2016; Sahibzada *et al.*, 2020b). Therefore, KM increases the satisfaction of knowledge workers, which can contribute to enhanced HEI outcomes (Shujahat *et al.*, 2018; Sahibzada *et al.*, 2020b). This study utilizes knowledge workers' satisfaction as a mediator to examine the interaction between KM processes and OP. Therefore, in the second objective of the present study, KWS was considered a mediating variable in the relationship between KM processes and organizational performance. HEIs' KM processes research are either insufficient or unpredictable (Fullwood and Rowley, 2017), and there is a need to investigate factors that contribute to improved KM (Iqbal *et al.*, 2019), particularly in emerging economies such as China and developing countries like Pakistan (Peng *et al.*, 2010; Turner and Acker, 2017). Preliminary research has been conducted on Chinese and Pakistani HEIs (Lo, 2016; Sahibzada *et al.*, 2020a, d).

Pakistan and China have lower-than-average static KM processes (Sahibzada *et al.*, 2021). To effectively improve organizational performance in cross-cultural settings among academics and administration, KM processes should be a significant factor in improving organizational efficiency (Iqbal *et al.*, 2019; Sahibzada *et al.*, 2020a, d). Because of the differences between the (emerging) Chinese and (developing) Pakistani economies, which are becoming more vital globally, the current study selected Pakistan and China to evaluate the inter-relationships between trust, organizational climate, KM processes, KWS, and organizational (HEI) performance (Munir *et al.*, 2019).

According to cultural analysts, Chinese culture has a complex framework that must be understood. Furthermore, according to Lewis (2010), a low-trust cultural group is exemplified in a country like China as people trust only those who are like their family and one or two close lifetime friends. On the other hand, Pakistan is defined by studies as a collectivist, high-power distance society (Hofstede, 2001; Lewis, 2010). It is widely assumed that the underlying relevance of social culture in this country is based on authority figures who promise to harmonize the entire population. They stress the significance of power and politics in Pakistani society. It is still widely assumed that Pakistanis think, feel, and act in a limited manner. It is more acquainted with regional identities and subcultures such as castes, local communities, and language groups. Furthermore, many countries, including Pakistan, benefit directly or indirectly from China's rapid economic growth (Munir *et al.*, 2019; Sahibzada *et al.*, 2021). Because both countries work in a variety of fields and employ students for work and research, they have a shared understanding of academic and administrative workloads (Iqbal *et al.*, 2019; Sahibzada *et al.*, 2020a, d, e) and their impact on organizational efficiency can provide significant results from which both countries can benefit. This study assumes that the known knowledge gaps of KM in HEIs are filled to bridge limitations, and thus the current research makes many theoretical contributions to the current literature by evaluating the inter-relationship between trust, organizational climate, KM processes, KWS, and organizational performance. The research examines whether KWS intervenes in the association between the KM processes and organizational performance in cross-cultural circumstances. The study will offer a structure for insight into the mediating device by which the processes of the KM impact the result of ethically different HEIs. This study will considerably help accept and emphasize the KM role in Chinese and Pakistani HEIs, with a considerably restricted study on the position of the KM in Pakistan and Chinese HEIs. This research helps in adding to the area of the KM in the higher education sector and explains the following:

(1) Herzberg theory (Herzberg, 1974) and

(2) Knowledge-based view (KBV) (Grant, 1996).

Based on the literature gaps, the following research questions (RQs) are proposed:

*RQ1.* Do trust and the organizational climate at universities in China and Pakistan affect KM processes?

*RQ2.* Do KM processes have a direct effect on organizational performance in the universities of Pakistan and China?

*RQ3.* Does KWS mediate the relationship in the universities of Pakistan and China between KM processes and organizational performance?

*RQ4.* Are there any differences in the significance of the relationship between HEIs of Pakistan and China?

## 2. Hypotheses development and research framework

### *2.1 Theoretical underpinnings*

To explain the interlinkages of KM enablers (i.e. trust, organizational climate), KM processes, KWS, and organizational performance with a research framework that is constructed on (1) KBV theory (Grant, 1996) and the KM capability model (Gold *et al.*, 2001), (2) Herzberg theory (Herzberg, 1966, 1974).

Gold *et al.* (2001) have divided KM Capability Model into Knowledge Process Capability and Knowledge Infrastructure Capability. Knowledge Process Capability represents KM processes (i.e. creation, acquisition, storage, sharing, and utilization), and the Knowledge Infrastructure Capability signifies the enablers of KM processes, i.e. organizational climate and trust (Butt *et al.*, 2018; Shujahat *et al.*, 2018; Sahibzada *et al.*, 2020b). The KM Capability Model promotes that an organization's efficiency depends on Knowledge Infrastructure Capability, ensuring that KM processes facilitate the environment. This helps in enhancing organizational performance persistently.

Knowledge-based view relies on an organization's resource-based view and includes significant, exceptional, and limited knowledge benefits and traits (Grant, 1996). Furthermore, KBV asserts that knowledge is a critical advantage in organizations as it significantly benefits a "sustainable competitive edge" (Grant, 1996). Thus, organizations can function effectively and efficiently by administering and employing their knowledge and understanding (Seleim and Khalil, 2007; Zack *et al.*, 2009). Based on the theoretical views, the present research suggests an integrated model that investigates the interconnection between trust, organizational climate, KM processes, KWS, and organizational performance.

This study has a strong theoretical connection with Herzberg's theory. Herzberg's approach has two main conclusions: (a) hygienic considerations and (b) motivational elements (Herzberg, 1966, 1974).

(1) Hygiene factors (also known as job dissatisfiers) are extrinsic work environment components.

(2) Motivating factors are intrinsic work elements that lead to satisfaction. KM processes ensure the withdrawal of dissatisfaction while motivating factors signify reinforcement (Shujahat *et al.*, 2018; Sahibzada *et al.*, 2020a, d).

### *2.2 Trust and knowledge management processes*

As per Mayer *et al.* (1995), trust means a trustor's readiness to accept other trustees' actions based on the expectation that a trustor's particular important action can be achieved by the trustee irrespective of the ability to scrutinize or control the other party. Trust needs a partner's reliability and intent to execute on that reliability (Moorman *et al.*, 1992; Sahibzada

*et al.*, 2020c). There is a connection between trust and KM at a basic level (Mcneish and Mann, 2010; Yasir and Majid, 2017; Sahibzada *et al.*, 2020c). Trust is the driving force resulting from interpersonal connections (Ford, 2004). There is an improved connection between different people's associations because trust and more features help people have faith in each other (Holste and Fields, 2010). This forms the initial foundation for improving a person's KM behavior in HEIs (Yasir *et al.*, 2017). Trust inspires the growth of connections (Whisnant and Khasawneh, 2014; Koohang *et al.*, 2017). The research identifies that personal trust positively impacts KM activities in an organization (Dirks and Ferrin, 2002). Trustworthy individuals will be simple and keen to impart their specific knowledge and expertise (Lee *et al.*, 2010). Trust in leaders and trust between co-workers significantly impact the KM processes (Lee *et al.*, 2010). The management literature proved the considerable impact of trust on the KM processes (sharing knowledge in HEIs) (Yasir and Majid, 2017; Yasir *et al.*, 2017). This is because trust between workers produces an emotional connection (Chowdhury, 2005), makes the baseline stronger for the worker to support and attract information between other members (Levin *et al.*, 2006), and consent to create, acquire, store, share, and utilize knowledge (Zand, 1972). According to Zand (1972), people or groups with more trust amongst members are reasonable and distribute added information than people or groups experiencing less confidence amongst associates. Lewis (2010) stated that "low and high-trust cultural group is exemplified by country, for example, people in China trust completely only those they best like their family and one or two close lifetime friends compared to Pakistan. Thus, based on the earlier mentioned discussion, the following hypothesis is suggested:

*H1a.* Trust has a positive and significant effect on KM processes.

### *2.3 Organizational climate and knowledge management processes*

Janz *et al.* (1997) explained that organizational climate is the general practice of shared faith and values followed by an organization. The organizational climate can alter KM performance (Jones *et al.*, 2006; Sahibzada *et al.*, 2020c). The impact of organizational climate on the KM processes is a strong factor found in the literature on KM (Lin and Lee, 2006; Jain *et al.*, 2015). Organizations with an enhanced organizational climate increase dealing between workers, and there is excessive knowledge sharing for creative opinions (Edmondson, 1999; Chen and Huang, 2007). When there are original ideas, collaboration between individuals is important in advancing those thoughts (Sveiby and Simons, 2002). Individual-produced new organizational knowledge is set up through group communication (Floyd and Lane, 2000).

Organizations can improve individuals' readiness to relate with people by developing an organizational climate (Chen and Huang, 2007; Sahibzada *et al.*, 2020c). Members work jointly and share the information that supports everyone's performance if the organizational climate is positive (Janz and Prasarnphanich, 2003). According to Davenport and Prusak (1998) and Dixon (2000), factors like an honest exchange of ideas and employees' anxiety to work are important for routing KM (acquisition, sharing, and utilization). Also, assurance from the management and the recognition of hazards are mentioned in the literature as important aspects of climate in an organization that allows the processes of KM (Pérez *et al.*, 2004; Lee *et al.*, 2006). Based on the earlier discussion, the following hypothesis is presented:

*H1b.* Organizational climate positively and significantly affects KM processes.

### *2.4 Knowledge management processes and organizational performance*

As organizations have unstable business surroundings, knowledge is a medium for reasonable, sustainable benefits (Masa'deh *et al.*, 2017; Shahzadi *et al.*, 2021). Organizational knowledge benefits worth can be measured when these benefits are utilized to produce or generate products, deliver services, sell, or do business for the price (Wiig, 1999). Thus, organizational

performance can be advanced by efficiently applying the previously created, stored, transferred, and applied knowledge (Alavi and Leidner, 2001). Similarly, KM is gaining prominence in HEIs due to its performance-driven nature (Masa'deh *et al.*, 2016). On the other hand, OP depends on the efficient management and utilization of available knowledge-based resources and the productive implementation of KM processes (Mahdavi Mazdeh and Hesamamiri, 2014; Shahzadi *et al.*, 2021). As per KBV, knowledge-related reserves are the important and planned reserves that elevate organizational performance (Grant, 1996; Donate and Guadamillas, 2015). Also, KBV supports the idea that influences and continues with the ability to produce, convey, and use the information to encourage organizational performance (Martelo-Landroguez and Cepeda-Carrión, 2016). Information-focused services rely on the effective execution of organizational knowledge (Obeidat *et al.*, 2016). Likewise, organizational combined knowledge, result, and output with the help of new creative ideas, goods, and services are made to the subsequent level with the help of acquisition, sharing, and utilization of knowledge (Chiu and Chen, 2016; Masa'deh *et al.*, 2017). Knowledge-sharing improves the research functioning in HEIs (Mahamed Ismail *et al.*, 2015). Based on the current literature, it has been observed and verified that there is an important and positive association between KM processes and organizational performance (Chiu and Chen, 2016; Ngah *et al.*, 2016; Shahzad *et al.*, 2016). Ahmad *et al.* (2017) and Iqbal *et al.* (2019) demonstrated a significant and direct association between KM processes and HEIs performance.

*H2.* KM processes significantly positively and directly affect organizational performance.

### *2.5 Knowledge management processes, knowledge worker satisfaction, and organizational performance*

The positive impact of the processes of KM on the fulfillment of KWS is supported by Herzberg's Two Factors Theory (Herzberg, 1966, 1974). The two important conclusions from Herzberg's Two Factors Theory are hygiene and motivators (Herzberg, 1966, 1974). Methods of KM verification, particularly KM processes, supply the hygiene elements to avoid discontent and reinforce the need for incentives (Shujahat *et al.*, 2018). The KM verifies that compensation, management help, and reasonable behavior are needed for employees' satisfaction along with the condition of hygiene elements (Drucker, 1998; Drucker, 1999; Kulkarni *et al.*, 2006, Donate and De Pablo, 2015). The KM processes also impact motivators and motivating elements with the help of the following motivation elements (Drucker, 1999; Turriago-Hoyos *et al.*, 2016; Liu *et al.*, 2017; Palvalin, 2017; Palvalin *et al.*, 2017; Shujahat *et al.*, 2018):

Providing knowledge workers with subsequent motivational underpinnings includes:

(1) Presenting employees with complete autonomy in the job;

(2) Producing a learning and information culture that impacts the KWS on KM processes;

(3) Permitting employees to participate in knowledge activities, thus, helps them with intrinsic motivation to generate and maintain knowledge and information (Nonaka and Takeuchi, 1995).

(4) Allocating the work to workers to generate, impart, and use the information;

(5) Ensuring the conditions for maintaining employment services.

Knowledge management can act as a catalyst for KWS (Sahibzada *et al.*, 2020a). The following reasoning supports the earlier-mentioned discussion (Razmerita *et al.*, 2016; Shujahat *et al.*, 2018; Sahibzada *et al.*, 2020):

First, the literature assessment postulates that subjective criteria like KWS should examine KM. This means that knowledge-worker satisfaction is important for KM effectiveness (Sahibzada *et al.*, 2020a, d, e, 2021g; Shahzadi *et al.*, 2021).

Furthermore, monetary incentives do not motivate knowledge workers (Sahibzada *et al.*, 2020a; Shahzadi *et al.*, 2021). They are encouraged when they are allowed to face challenges associated with knowledge-based activities (Sahibzada *et al.*, 2020e). KM provides knowledge workforces with this encounter and the possibility of solving knowledge-related problems. Through its policies, KM provides knowledge workers with the right amount and type of knowledge at the right time and place (Shujahat *et al.*, 2018; Kianto *et al.*, 2019).

The use of information helps in making the work easy. The present research maintains that knowledge creation, acquisition, storage, sharing, and utilization can increase the fulfillment of KWS as the new information can enhance the knowledge workers' output (Sahibzada *et al.*, 2020e). In addition, sharing knowledge ensures that employees' interpersonal requirements are fulfilled (Shahzadi *et al.*, 2021). Thus, KM processes to aid in the fulfillment of KWS.

Knowledge management processes and knowledge-worker satisfaction can be derived from the earlier research on KM and an employee's fulfillment. Prior experiential research relates to KM scope, particularly the sharing of information and work contentment acquired directly or not (Kianto *et al.*, 2016; Butt *et al.*, 2018; Shujahat *et al.*, 2018). This covers KM processes and fulfillment with inter-relations relating to KM (Chou *et al.*, 2005; Chatzoudes *et al.*, 2015). A few scholars established the impact of processes of KM on worker's fulfillment directly and indirectly (Lim *et al.*, 1999; Sharma, 2008; Bontis and Serenko, 2009; Singh and Sharma, 2011).

The previous studies elaborated on the impact of KM processes on employee satisfaction but were inconclusive because:

(1) No study is available that investigates the influence of second-order constructs of the procedures of the management of knowledge on workers' contentment in specific research, specifically in a cross-cultural context.

(2) These researches do not replicate the information attained from the knowledge of employees, particularly from the information-based service sector HEIs.

(3) The earlier research did not directly observe the impact of KM processes on the knowledge employees' fulfillment. Also, these studies observed the connection as a secondary hypothesis in an element of the broad model.

(4) Many of the earlier-mentioned studies reproduce the fulfillment in a job as an alternative to an employee's fulfillment because of knowledge management.

The present research investigates the impact of KM processes (creation, acquisition, storage, sharing, and utilization) on organizational performance with the mediating role of KWS.

A complete valuation of the literature depicts that KM promotes organizational performance (Pang and Lu, 2018; Viñas-Bardolet *et al.*, 2018). This augments the fulfillment of KWS (Shujahat *et al.*, 2018; Sahibzada *et al.*, 2020). Hence, observing the association between processes of KM, KWS, and HEIs performance is significant (see Table 1).

*H3.* Knowledge management processes have a significant and positive impact on KWS.

*H4.* KWS has a significant and positive impact on organizational performance.

*H5.* KWS mediates the association between the KM processes and organizational performance.

**Table 1.** Comparison of previous studies on KM enablers, knowledge management processes, and knowledge worker satisfaction

| Author (s), year | KM enablers | KM processes | Knowledge worker satisfaction | Industry | Country | Findings |
|---|---|---|---|---|---|---|
| *Relationship Among Enablers* | | | | | | |
| Bennett and Gabriel (1999) | Structure, Culture, Size, Environment, **KM Method** | N/A | N/A | Marketing Firms | United Kingdom | The substantial effect of change friendly culture on the number of KM methods employed |
| *Relationship Between Enablers and Processes* | | | | | | |
| Zander and Kogut (1995) | Characteristics of societal knowledge | **Transfer (time to transfer)** | N/A | Project Base Engineering Firms | Sweden | Modifiability, teachability, and parallel development significantly affect the time to knowledge transfer |
| Appleyard (1996) | Industry and national characteristics | **Transfer (number of times the respondents provide and receive knowledge in each period)** | N/A | Steel Industry | The United States and Japan | Public sources of knowledge are much more prevalent in knowledge transfer in semiconductors than in the steel industry; public sources of technical knowledge play a more significant role in knowledge transfer in Japan than in the United States |
| Szulanski (1996) | Characteristics of knowledge transferred source recipient source | **Transfer (Four-Stage transfer processes)** | N/A | Corporate Companies | United Kingdom | The recipient's lack of absorptive capacity, causal ambiguity, and arduousness of the relationship are the major impediments to knowledge transfer |
| Hansen (1999) | Weak ties (distant and infrequent relationships); knowledge characteristics | **Transfer (percentage of a project's total knowledge that comes from other divisions)** | N/A | Electronic and Computer Companies | United States | Weak ties impede the transfer of complex knowledge |

(*continued*)

Table 1.

| Author (s), year | KM enablers | KM processes | Knowledge worker satisfaction | Industry | Country | Findings |
|---|---|---|---|---|---|---|
| Iqbal *et al.* (2019) | Leadership, Culture, Incentives | **Knowledge Acquisition, Knowledge Sharing, Knowledge Utilization** | N/A | Higher Educational Institutes | Pakistan | Leadership support, organizational culture, and incentives are mandatory for successful implementation of KM processes |
| Sahibzada *et al.* (2020c) | Trust, Organizational Climate | **Knowledge, Creation, Knowledge Acquisition, Knowledge Storage, Knowledge Sharing, Knowledge Utilization** | N/A | Higher Educational Institutes | China | Trust and Organizational Climate are the key components to influence KM processes positively |
| *Relationship Between KM Processes and Knowledge Worker Satisfaction* | | | | | | |
| Shahzadi *et al.* (2021) | N/A | Knowledge Identification, Knowledge Creation, Knowledge Acquisition Knowledge Organizing Knowledge Storage Knowledge Sharing Knowledge Utilization | **Knowledge Worker Satisfaction** | Software Industry | China | Knowledge management processes influence project success via knowledge worker satisfaction as a mediator both directly and indirectly. Moreover, the study found partial mediation of knowledge worker satisfaction between knowledge management processes and project success |
| Sahibzada *et al.* (2020a) | N/A | Knowledge Acquisition, Knowledge Sharing, Knowledge Utilization | **Knowledge Worker Satisfaction** | Higher Education Institute | Pakistan | KM processes (i.e. acquisition, sharing and utilization) assist KWS via improved knowledge systems and learning, enhanced organizational operations, capabilities, and personnel's cognitive methods |
| Shujahat *et al.* (2018) | N/A | Knowledge Creation, Knowledge Sharing, Knowledge Utilization | **Knowledge Worker Satisfaction** | Software Houses | Pakistan | Knowledge creation and knowledge sharing have a positive impact on knowledge worker satisfaction |

**Note(s):** Boldface type indicates dependent variables

*2.6 Knowledge management enabler (S), knowledge management processes, knowledge worker satisfaction, and organizational performance: the contextual influence*

The World Economic Forum's Global Competitiveness Report defined three groups of countries according to the economic development stage as mentioned in Munir *et al.* (2019) "The World Economic Forum's Global Competitiveness Report defined three groups of countries according to their economic-development stage, based on their GDP and other related variables: factor-driven economies; efficiency-driven economies; and innovation-driven economies. Factor-driven economies are in the early stages of economic development and mainly depend on their agriculture sector. Efficiency-driven economies are characterized by increased productivity and the development of small-scale or medium-sized manufacturing sectors. Innovation-driven economies are characterized by their research, technology, innovation advancements, more sophisticated production processes, and improved products. Of the two countries in this study, Pakistan is factor-driven, and China is efficiency-driven".

Thus, Pakistan is a factor-driven economy, and China is an efficiency-driven economy (Sahibzada *et al.*, 2021), highlighting the importance of context. It also emphasized that knowledge management is context-dependent, i.e. the processes (Knowledge creation, sharing, and application) and enablers (e.g. technology) for a particular organization and environment might not work for other organizations with different contextual backgrounds and characteristics (Sumbal *et al.*, 2019). Hence, the background of KM processes is often seen as a specific culture and context (Iqbal *et al.*, 2019; Sahibzada *et al.*, 2020a, d), and various levels of country-based influence have been assigned to organizational performance through KM processes (Sahibzada *et al.*, 2020b, d). Therefore, to determine KM processes, the country's specific context should be considered (Sahibzada *et al.*, 2020b); for example, knowledge creation might not be applicable in certain developing countries' contexts in industries utilizing the existing knowledge through knowledge application and knowledge sharing, e.g. (Ahmed *et al.*, 2021). In addition, economies differ in terms of their economic circumstances, level of progress, environment, and background (Munir *et al.*, 2019), which may cause differences in the strength of KM enabler(s), KM processes, KM satisfaction, and organizational performance relationships.

*H6.* The impact of (a) trust on KM processes, (b) organizational climate on KM processes, (c) KM processes on organizational performance, (d) KM processes on KWS, and (e) KWS on organizational performance is likely to differ between Pakistan and China (Figure 1).

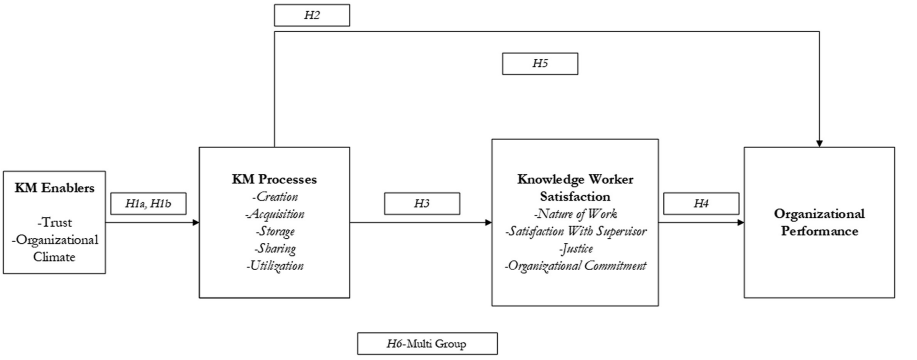



**Figure 1.**
Conceptual framework

## 3. Methodology

### *3.1 Sample, sampling technique, and data collection*

As per the developing and emerging economies, economies based on knowledge depend only on university functioning that can finally impact the growth and economy of a society (Sahibzada *et al.*, 2019, 2020b). Thus, research-related activities should be an element of universities in emerging economies like China and developing economies like Pakistan. The Ministry of Education of the People's Republic of China and the Higher Education Commission, Pakistan, encourage investigative surroundings that lead to higher education studies or university research activities (Lo, 2016; Iqbal *et al.*, 2019). The research proposals considered by the Ministry of Education of the People's Republic of China and the Higher Education Commission of Pakistan are still in the infancy stage in HEIs. This needs an active KM to meet the knowledge-based economy obstacles (Feiz *et al.*, 2019).

This research presents Pakistan and Chinese public and private universities' academic and administrative staff as the considered population. They are engaged in succeeding in "multi-disciplinary education and investigating nature and technology, engineering, and social sciences. The questionnaires were distributed in English in Pakistan, as English is the official language for communication. In China, the questionnaire was distributed in both English and Chinese. The questionnaire was translated into Chinese by a Chinese language expert and further validated by a senior professor in the field of Business and Management from Northwestern Polytechnical University, Xi'an. The study utilizes the convenience sampling technique. There were 1,120 questionnaires distributed in two major cities of China, namely, eight public universities in Xi'an and Chengdu. There were 609 questionnaires collected from sixteen Chinese universities and ticked a response rate of 54.3%. The number of discarded questionnaires was 73 (both unfinished and unacceptable). The remaining 536 questionnaires were kept for data analysis as these ticked a response rate of 47.8% for statistical examination from China. The Pakistan sample had 600 distributed questionnaires from sixteen universities in two cities, i.e. Peshawar and Islamabad. The collected response was 309, i.e. an overall response rate of 51.5%. The number of unacceptable and missing questionnaires was 61, which were discarded. In total, 248 questionnaires were used for data analysis, and the response rate of used questionnaires was 41.3%. Data were collected from August 2019 to January 2020. This is the correct sample dimension for applying the structural equation modeling to examine the complicated path model (Kline, 2011). (For respondent profiles, see Tables 2 and 3).

### *3.2 Instrumentation*

Fifty-five measurement items were used in this research. Constructs were adapted from previous studies, as given in Table 4. In the meantime, the language in the articles was modified to run parallel with university understandings (Sahibzada *et al.*, 2019). A five-point Likert scale was employed in the present research. This varies from "1", i.e. "strongly disagree," to "5" denoting "Strongly Agree" (For questionnaire, please see appendix).

### *3.3 Analysis of data*

The study is examined by Smart PLS 3.2.9 (Ringle *et al.*, 2005). Partial Least Square Structural Equation Modeling (PLS-SEM) was selected for examining "quantifiable information." PLS-SEM is a rising "knowledge-processing" method that is applied in business and social science studies to administer sample size and "non-normal evidence" effectively (Hair *et al.*, 2014). This method is appropriate for testing the accessible hypotheses and contains complicated structural models (Fernandes, 2012; Ringle *et al.*, 2018). Two approaches apply to the SEM method: Covariance-based (CB-SEM) and PLS-SEM (Hair *et al.*, 2012). PLS-SEM is applied to the present research instead of CB-SEM as the same is best suited for complicated

| Demographics variables | Frequency | Percentage |
|---|---|---|
| *City* | | |
| Xi'an | 325 | 60.6% |
| Chengdu | 211 | 39.4% |
| *Age* | | |
| 20–29 | 187 | 34.9% |
| 30–39 | 155 | 28.9% |
| 40–49 | 129 | 24.1% |
| 50–59 | 65 | 12.1% |
| *Gender* | | |
| Male | 345 | 64.4% |
| Female | 191 | 35.6% |
| *Education* | | |
| PhD | 349 | 65.1% |
| Masters | 187 | 34.9% |
| *Job Tenure* | | |
| 0–5 | 187 | 34.9% |
| 6–10 | 153 | 28.5% |
| 10–15 | 52 | 9.7% |
| 16–20 | 81 | 15.1% |
| 20–25 | 40 | 7.5% |
| 26–30 | 23 | 4.3% |
| *Area* | | |
| Administration | 147 | 27.4% |
| Academics | 389 | 72.6% |
| *Designation* | | |
| Professor | 65 | 12.1% |
| Associate Professor | 130 | 24.3% |
| Assistant Professor | 154 | 28.7% |
| Lecturer | 40 | 7.5% |
| Administrators | 147 | 27.4% |
| *Institutions* | | |
| Northwestern Polytechnical University | 59 | 11.0% |
| Shaanxi Normal University | 46 | 8.6% |
| Northwest University | 49 | 9.1% |
| Northwest Agriculture and Forestry University | 38 | 7.1% |
| Xi'an Jiaotong University | 39 | 7.3% |
| Changa'an University | 36 | 6.7% |
| Xidian University | 34 | 6.3% |
| Xi'an Architecture University | 25 | 4.7% |
| Sichuan University | 25 | 4.7% |
| Southwest Jiaotong University | 12 | 2.2% |
| Southwestern University of Finance and Economics | 40 | 7.5% |
| Sichuan Normal University | 22 | 4.1% |
| Chengdu University | 22 | 4.1% |
| Chengdu University of Technology | 32 | 6.0% |
| Chengdu University of Science and Technology | 23 | 4.3% |
| University of Electronic Science and Technology of China | 34 | 6.3% |

**Table 2.**
Profile of respondents from China

| Demographics variables | Frequency | Percentage |
|---|---|---|
| *City* | | |
| Islamabad | 111 | 44.8% |
| Peshawar | 137 | 55.2% |
| *Age* | | |
| 20–29 | 85 | 34.3% |
| 30–39 | 70 | 28.2% |
| 40–49 | 57 | 23.0% |
| 50–59 | 36 | 14.5% |
| *Gender* | | |
| Male | 139 | 56.0% |
| Female | 109 | 44.0% |
| *Education* | | |
| PhD | 164 | 66.1% |
| Masters | 84 | 33.9% |
| *Job Tenure* | | |
| 0–5 | 84 | 33.9% |
| 6–10 | 70 | 28.2% |
| 10–15 | 36 | 14.5% |
| 16–20 | 24 | 9.7% |
| 20–25 | 24 | 9.7% |
| 26–30 | 10 | 4.0% |
| *Area* | | |
| Administration | 74 | 29.8% |
| Academics | 174 | 70.2% |
| *Designation* | | |
| Professor | 35 | 14.1% |
| Associate Professor | 59 | 23.8% |
| Assistant Professor | 73 | 29.4% |
| Lecturer | 7 | 2.8% |
| Administrators | 74 | 29.8% |
| *Institutions* | | |
| Abasyn University | 19 | 7.7% |
| City University | 12 | 4.8% |
| Institute of Management Sciences | 14 | 5.6% |
| Islamia College University | 20 | 8.1% |
| University of Engineering and Technology | 14 | 5.6% |
| Shaheed Benazir Bhutto Women University | 15 | 6.0% |
| CECOS University | 14 | 5.6% |
| Qurtaba University | 17 | 6.9% |
| Sarhad University | 12 | 4.8% |
| Bahria University | 15 | 6.0% |
| Comsats University | 18 | 7.3% |
| Institute of Space Technology | 13 | 5.2% |
| Riphah University | 20 | 8.1% |
| Air University | 15 | 6.0% |
| International Islamic University | 18 | 7.3% |
| Shaheed Zulfikar Ali Bhutto Institute of Science and Technology | 12 | 4.8% |

**Table 3.**
Profile of respondents from Pakistan

| Variable | Dimension | No. of items | Source |
|---|---|---|---|
| Knowledge Management Enablers | Trust | 04 | Huff and Kelley (2003) |
| | Organizational Climate | 10 | Bock *et al.* (2005) |
| Knowledge Management Processes | Creation | 06 | Bryant and Terborg (2008) |
| | Acquisition | 03 | Huang and Li (2009) |
| | Storage | 04 | Masa'deh *et al.* (2017) |
| | Sharing | 04 | Bryant and Terborg (2008) |
| | Utilization | 05 | Lee *et al.* (2005), Huang and Li (2009) |
| Knowledge Worker Satisfaction | Nature of Work | 03 | Weiss *et al.* (1967) |
| | Satisfaction with Supervisor | 04 | Smith (1976) |
| | Justice | 04 | Magner *et al.* (1994) |
| | Organizational Commitment | 03 | Porter *et al.* (1974) |
| Organizational Performance | | 05 | Tseng (2010) |

**Table 4.** Sources of measurement instruments

associations among the dormant concepts and searching theoretical levels (Henseler *et al.*, 2009; Hair *et al.*, 2014). This method is applied in theory validation and for examination. It is best suited for examining whether there are many connections (Fornell and Larcker, 1981). A twofold analysis is used for PLS-SEM. These are Measurement Model Calculation (outer model) and Structural Model Analysis (inner model) (Wong, 2013; Ringle *et al.*, 2018). Measurement Model requirements guarantee that these viewpoints are suitable for pointer loading, convergent validity, composite reliability, and discriminant validity, expanding to the Structural Model. Measuring path coefficients are employed in the evaluation of the Structural Model. Also, the current information management analytical literature has applied the PLS-SEM data processing technique (Shujahat *et al.*, 2018; Sahibzada *et al.*, 2020a).

## 4. Data analysis

### *4.1 Assessment of Measurement Model*

The Measurement Model's first assessment phase was done to verify the construct's reliability and validity (Hair, 2006). There were 55 items in the procedure. In the Measurement Model, no items were dropped as all the factor loadings were more than the recommended value of 0.600 (Hair *et al.*, 2016). Thus, all questions were retained in the last measurement procedure (Table 5). The average variance extracted (AVE) and composite reliability of all the constructs are equivalent to or exceed the values of 0.50 and 0.70. Thus, convergent validity and reliability are verified for both countries and the complete sample. Also, Tables 6–8 depict the discriminant validity outcome via Fornell and Larcker criterion. Hence, there were no items removed, and discriminant validity was ascertained.

### *4.2 Structural model assessment*

The results revealed $R^2$ values in China, 0.89, 0.83, and 0.64 for KM Processes (KMPs), Knowledge Worker Satisfaction (KWS), and Organizational Performance (OP), respectively. In Pakistan, 0.80, 0.60, and 0.54 for KMPs, KWS, and OP, respectively, and in the overall sample, 0.86, 0.77, and 0.62 for KMPs, KWS, and OP, respectively. The $R^2$ values support the

| Constructs | China | | | | Pakistan | | | | Overall | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L | $\alpha$ | CR | AVE | L | $\alpha$ | CR | AVE | L | $\alpha$ | CR | AVE |
| *AQ* | | *0.86* | *0.91* | *0.78* | | *0.75* | *0.86* | *0.67* | | *0.84* | *0.90* | *0.75* |
| AQ1 | 0.83 | | | | 0.74 | | | | 0.80 | | | |
| AQ2 | 0.77 | | | | 0.65 | | | | 0.75 | | | |
| AQ3 | 0.80 | | | | 0.72 | | | | 0.77 | | | |
| *C* | | *0.91* | *0.93* | *0.70* | | *0.83* | *0.87* | *0.54* | | *0.90* | *0.92* | *0.66* |
| C1 | 0.81 | | | | 0.76 | | | | 0.79 | | | |
| *C2* | 0.74 | | | | 0.68 | | | | 0.72 | | | |
| *C3* | 0.77 | | | | 0.68 | | | | 0.74 | | | |
| C4 | 0.82 | | | | 0.66 | | | | 0.79 | | | |
| C5 | 0.81 | | | | 0.56 | | | | 0.77 | | | |
| C6 | 0.80 | | | | 0.73 | | | | 0.78 | | | |
| *NOW* | | *0.89* | *0.93* | *0.82* | | *0.73* | *0.85* | *0.65* | | *0.86* | *0.92* | *0.78* |
| NOW1 | 0.87 | | | | 0.74 | | | | 0.85 | | | |
| NOW2 | 0.83 | | | | 0.66 | | | | 0.80 | | | |
| NOW3 | 0.86 | | | | 0.74 | | | | 0.83 | | | |
| *OCC* | | *0.85* | *0.91* | *0.77* | | *0.65* | *0.81* | *0.59* | | *0.81* | *0.89* | *0.73* |
| OCC1 | 0.83 | | | | 0.69 | | | | 0.80 | | | |
| OCC2 | 0.84 | | | | 0.75 | | | | 0.82 | | | |
| OCC3 | 0.81 | | | | 0.62 | | | | 0.77 | | | |
| *OC* | | *0.94* | *0.95* | *0.67* | | *0.91* | *0.93* | *0.56* | | *0.94* | *0.95* | *0.64* |
| OC1 | 0.81 | | | | 0.71 | | | | 0.79 | | | |
| OC2 | 0.77 | | | | 0.69 | | | | 0.75 | | | |
| OC3 | 0.86 | | | | 0.77 | | | | 0.84 | | | |
| OC4 | 0.79 | | | | 0.76 | | | | 0.78 | | | |
| OC5 | 0.88 | | | | 0.80 | | | | 0.86 | | | |
| OC6 | 0.74 | | | | 0.67 | | | | 0.72 | | | |
| OC7 | 0.80 | | | | 0.76 | | | | 0.79 | | | |
| OC8 | 0.87 | | | | 0.83 | | | | 0.86 | | | |
| OC9 | 0.82 | | | | 0.72 | | | | 0.80 | | | |
| OC10 | 0.82 | | | | 0.76 | | | | 0.80 | | | |
| *PJ* | | *0.89* | *0.93* | *0.76* | | *0.80* | *0.87* | *0.62* | | *0.88* | *0.92* | *0.73* |
| PJ1 | 0.84 | | | | 0.77 | | | | 0.82 | | | |
| PJ2 | 0.80 | | | | 0.67 | | | | 0.77 | | | |
| PJ3 | 0.81 | | | | 0.67 | | | | 0.79 | | | |
| PJ4 | 0.84 | | | | 0.73 | | | | 0.82 | | | |
| *SH* | | *0.87* | *0.91* | *0.72* | | *0.71* | *0.82* | *0.54* | | *0.85* | *0.90* | *0.68* |
| SH1 | 0.77 | | | | 0.66 | | | | 0.74 | | | |
| SH2 | 0.87 | | | | 0.72 | | | | 0.83 | | | |
| SH3 | 0.79 | | | | 0.56 | | | | 0.74 | | | |
| SH4 | 0.80 | | | | 0.57 | | | | 0.75 | | | |
| *ST* | | *0.87* | *0.91* | *0.72* | | *0.73* | *0.83* | *0.56* | | *0.85* | *0.90* | *0.68* |
| ST1 | 0.81 | | | | 0.69 | | | | 0.78 | | | |
| ST2 | 0.80 | | | | 0.60 | | | | 0.75 | | | |
| ST3 | 0.77 | | | | 0.67 | | | | 0.75 | | | |
| ST4 | 0.83 | | | | 0.66 | | | | 0.79 | | | |
| *SWS* | | *0.91* | *0.94* | *0.78* | | *0.82* | *0.88* | *0.65* | | *0.89* | *0.92* | *0.75* |
| SWS1 | 0.83 | | | | 0.72 | | | | 0.81 | | | |
| SWS2 | 0.85 | | | | 0.72 | | | | 0.83 | | | |
| SWS3 | 0.83 | | | | 0.69 | | | | 0.80 | | | |
| SWS4 | 0.87 | | | | 0.78 | | | | 0.85 | | | |
| *T* | | *0.89* | *0.92* | *0.75* | | *0.75* | *0.84* | *0.57* | | *0.86* | *0.91* | *0.71* |
| T1 | 0.87 | | | | 0.75 | | | | 0.84 | | | |
| T2 | 0.84 | | | | 0.63 | | | | 0.80 | | | |

(*continued*)

**Table 5.**
Item loadings, reliability, and convergent validity

| Constructs | China L | China α | China CR | China AVE | Pakistan L | Pakistan α | Pakistan CR | Pakistan AVE | Overall L | Overall α | Overall CR | Overall AVE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T3 | 0.86 | | | | 0.78 | | | | 0.83 | | | |
| T4 | 0.90 | | | | 0.84 | | | | 0.88 | | | |
| *OP* | | *0.92* | *0.94* | *0.75* | | *0.80* | *0.86* | *0.56* | | *0.90* | *0.93* | *0.72* |
| OP1 | 0.86 | | | | 0.73 | | | | 0.84 | | | |
| OP2 | 0.84 | | | | 0.74 | | | | 0.82 | | | |
| OP3 | 0.86 | | | | 0.73 | | | | 0.84 | | | |
| OP4 | 0.89 | | | | 0.76 | | | | 0.87 | | | |
| OP5 | 0.89 | | | | 0.79 | | | | 0.87 | | | |
| *UT* | | *0.90* | *0.93* | *0.71* | | *0.90* | *0.92* | *0.91* | | *0.90* | *0.93* | *0.71* |
| UT1 | 0.78 | | | | 0.89 | | | | 0.67 | | | |
| UT2 | 0.81 | | | | 0.89 | | | | 0.70 | | | |
| UT3 | 0.71 | | | | 0.85 | | | | 0.61 | | | |
| UT4 | 0.87 | | | | 0.84 | | | | 0.74 | | | |
| UT5 | 0.78 | | | | 0.74 | | | | 0.66 | | | |

**Table 5.**

| | AQ | CT | J | NOW | OC | OCc | OP | SH | ST | SWS | T | UT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AQ | *0.88* | | | | | | | | | | | |
| CT | 0.87 | *0.84* | | | | | | | | | | |
| J | 0.77 | 0.83 | *0.87* | | | | | | | | | |
| NOW | 0.81 | 0.80 | 0.85 | *0.90* | | | | | | | | |
| OC | 0.85 | 0.81 | 0.81 | 0.83 | *0.82* | | | | | | | |
| OCc | 0.79 | 0.82 | 0.86 | 0.87 | 0.81 | *0.88* | | | | | | |
| OP | 0.69 | 0.72 | 0.78 | 0.76 | 0.69 | 0.75 | *0.87* | | | | | |
| SH | 0.83 | 0.81 | 0.81 | 0.82 | 0.80 | 0.81 | 0.70 | *0.85* | | | | |
| ST | 0.83 | 0.82 | 0.81 | 0.82 | 0.79 | 0.80 | 0.72 | 0.81 | *0.85* | | | |
| SWS | 0.78 | 0.80 | 0.86 | 0.87 | 0.80 | 0.87 | 0.75 | 0.83 | 0.85 | *0.88* | | |
| T | 0.81 | 0.81 | 0.85 | 0.85 | 0.81 | 0.84 | 0.74 | 0.81 | 0.84 | 0.86 | *0.87* | |
| UT | 0.79 | 0.83 | 0.80 | 0.80 | 0.77 | 0.81 | 0.69 | 0.92 | 0.86 | 0.84 | 0.83 | *0.84* |

**Table 6.**
Discriminant validity (Fornell and Larcker criterion) "China"

| | AQ | CT | J | NOW | OC | OCc | OP | SH | ST | SWS | T | UT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AQ | *0.82* | | | | | | | | | | | |
| CT | 0.79 | *0.73* | | | | | | | | | | |
| J | 0.56 | 0.66 | *0.79* | | | | | | | | | |
| NOW | 0.64 | 0.65 | 0.72 | *0.81* | | | | | | | | |
| OC | 0.73 | 0.77 | 0.62 | 0.66 | *0.75* | | | | | | | |
| OCc | 0.52 | 0.61 | 0.75 | 0.77 | 0.58 | *0.77* | | | | | | |
| OP | 0.56 | 0.60 | 0.64 | 0.67 | 0.55 | 0.62 | *0.75* | | | | | |
| SH | 0.65 | 0.71 | 0.68 | 0.71 | 0.75 | 0.65 | 0.62 | *0.73* | | | | |
| ST | 0.67 | 0.74 | 0.63 | 0.66 | 0.79 | 0.55 | 0.64 | 0.74 | *0.75* | | | |
| SWS | 0.53 | 0.59 | 0.73 | 0.72 | 0.63 | 0.76 | 0.64 | 0.68 | 0.67 | *0.80* | | |
| T | 0.66 | 0.74 | 0.74 | 0.68 | 0.72 | 0.66 | 0.59 | 0.70 | 0.70 | 0.64 | *0.76* | |
| UT | 0.60 | 0.78 | 0.66 | 0.63 | 0.79 | 0.62 | 0.63 | 0.64 | 0.65 | 0.66 | 0.72 | *0.84* |

**Table 7.**
Discriminant validity (Fornell and Larcker criterion) "Pakistan"

model's in-sample predictive power (Sarstedt *et al.*, 2014) since it is above the required level of 0.10 (Falk and Miller, 1992). Furthermore, effect sizes are calculated to assess the extent a predicting (exogenous) variable contributes to the $R^2$ value of an endogenous variable. In this study, KWS was predicted by KMPs, KWS and KMP predicted OP, and Trust and

| | AQ | CT | J | NOW | OC | OCc | OP | SH | ST | SWS | T | UT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AQ | *0.87* | | | | | | | | | | | |
| CT | 0.85 | *0.82* | | | | | | | | | | |
| J | 0.73 | 0.81 | *0.85* | | | | | | | | | |
| NOW | 0.78 | 0.78 | 0.83 | *0.89* | | | | | | | | |
| OC | 0.82 | 0.80 | 0.77 | 0.79 | *0.80* | | | | | | | |
| OCc | 0.74 | 0.79 | 0.84 | 0.85 | 0.77 | *0.85* | | | | | | |
| OP | 0.67 | 0.71 | 0.75 | 0.75 | 0.66 | 0.739 | *0.85* | | | | | |
| SH | 0.79 | 0.80 | 0.79 | 0.80 | 0.76 | 0.79 | 0.69 | *0.83* | | | | |
| ST | 0.80 | 0.78 | 0.78 | 0.79 | 0.77 | 0.76 | 0.71 | 0.81 | *0.83* | | | |
| SWS | 0.73 | 0.76 | 0.84 | 0.85 | 0.74 | 0.80 | 0.73 | 0.80 | 0.82 | *0.87* | | |
| T | 0.78 | 0.80 | 0.83 | 0.82 | 0.77 | 0.81 | 0.71 | 0.79 | 0.81 | 0.81 | *0.84* | |
| UT | 0.60 | 0.63 | 0.61 | 0.59 | 0.70 | 0.61 | 0.52 | 0.72 | 0.67 | 0.64 | 0.64 | *0.84* |

**Note(s):** The data on the diagonal (in italic) are the square root of the AVE of the construct, while the other values are the correlations with other constructs; **AQ: acquisition; CT: creation; J: justice; NOW: nature of work; OC: organizational climate; OCc: organizational commitment; OP: organizational performance; SH: sharing; ST: storage; SWS: satisfaction with supervisor; T: trust; UT: utilization

**Table 8.** Discriminant validity (Fornell and Larcker criterion) "overall"

organizational climate predicted KMPs. The relative effect sizes (f2) of the predicting (exogenous) constructs were calculated and show that the exogenous variable has a very big effect on your endogenous variable (>0.35) (Cohen, 1988) (see Table 9)

The hypotheses were investigated in a sequence. At the initial level, trust and organizational climate immediately impacted KM processes. KM processes directly affect KWS at the secondary level, establishing organizational performance. After the critical assessment of the Measurement Model, the assessment of the Structural Model test was completed in the second phase. Bootstrap evaluation of 5,000 resamples was used to analyze the importance of direct paths and test standard errors (Ringle *et al.*, 2005). Table 9 presents the strong link's evaluation findings between China, Pakistan, and the cumulative survey.

Additionally, the intervening effect of KWS among KM processes and organizational performance was verified in China, Pakistan, and the overall study. Lastly, the complete survey was confirmed for both countries. Table 10 describes the mediation analysis.

The results revealed a substantial positive influence of trust on KM processes in China ($\beta = 0.27$, $t = 7.09$, $p < 0.001$), Pakistan ($\beta = 0.36$, $t = 5.84$, $p < 0.001$) and the overall study ($\beta = 0.31$, $t = 9.22$, $p < 0.001$). Therefore, H1a was supported. Similarly, there is significant influence of organizational climate on KM processes in China ($\beta = 0.69$, $t = 18.80$, $p < 0.001$), Pakistan ($\beta = 0.60$, $t = 13.30$, $p < 0.001$) and the cumulative results ($\beta = 0.06$, $t = 20.78$, $p < 0.001$) thereby accepting H1b. However, the results revealed an insignificant direct influence of KM processes and organizational performance in China ($\beta = 0.10$, $t = 1.57$, $p = 0.11$) but the relationship was significant for Pakistan ($\beta = 0.30$, $t = 2.83$, $p < 0.001$) and overall survey ($\beta = 0.17$, $t = 3.28$, $p < 0.001$). Hence, H2 was partially substantiated. Additionally, the results also acknowledged that KM processes has a direct positive influence on KWS in China ($\beta = 0.91$, $t = 51.13$, $p < 0.001$), Pakistan ($\beta = 0.77$, $t = 6.44$, $p < 0.001$) and overall survey ($\beta = 0.88$, $t = 42.28$, $p < 0.001$), which lend support to accept the H3. Lastly, we found a significant positive direct impact of KWS on organizational performance in China ($\beta = 0.27$, $t = 7.09$, $p < 0.001$), Pakistan ($\beta = 0.48$, $t = 3.46$, $p < 0.001$) and overall survey ($\beta = 0.63$, $t = 11.16$, $p < 0.001$) accepting H4.

### *4.3 Mediation analysis*

Also, H5 assesses whether KWS mediates the link between KM processes and organizational performance. The result depicts that with the establishment of the mediator in the model, the

**Table 9.**
Results of structural model path coefficient (direct relationships)

| Hypotheses | Relationship | China $\beta$ | SD | $t$-value | $p$-value | Pakistan $\beta$ | SD | $t$-value | $p$-value | Overall $\beta$ | SD | $t$-value | $p$-value |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H1a | T → KMPs | 0.27 | 0.03 | 7.09 | 0.00 | 0.36 | 0.06 | 5.84 | 0.00 | 0.31 | 0.03 | 9.22 | 0.00 |
| H1b | OC → KMPs | 0.69 | 0.03 | 18.80 | 0.00 | 0.60 | 0.04 | 13.30 | 0.00 | 0.06 | 0.03 | 20.78 | 0.00 |
| H2 | KMPs → OP | 0.10 | 0.06 | 1.57 | 0.11 | 0.30 | 0.10 | 2.83 | 0.00 | 0.17 | 0.05 | 3.28 | 0.00 |
| H3 | KMPs → KWS | 0.91 | 0.01 | 51.13 | 0.00 | 0.77 | 0.12 | 6.44 | 0.00 | 0.88 | 0.02 | 42.28 | 0.00 |
| H4 | KWS → OP | 0.70 | 0.06 | 19.22 | 0.00 | 0.48 | 0.13 | 3.46 | 0.00 | 0.63 | 0.05 | 11.16 | 0.00 |

| China | $R^2$ | $f^2$ | Pakistan | $R^2$ | $f^2$ | Overall | $R^2$ | $f^2$ |
|---|---|---|---|---|---|---|---|---|
| KMP's | 0.89 | KMPs → OP–0.09 | KMP's | 0.80 | KMPs → OP–0.08 | KMP's | 0.86 | KMPs → OP–0.12 |
| KWS | 0.83 | KWS → OP–0.23 | KWS | 0.60 | KWS → OP–0.19 | KWS | 0.77 | KWS → OP–0.23 |
| OP | 0.64 | KMPs → KWS–4.9 | OP | 0.54 | KMPs → KWS–1.51 | OP | 0.62 | KMPs → KWS– 3.50 |
| | | T → KMPs–0.19 | | | T → KMPs–0.31 | | | T → KMPs–0.22 |
| | | OC → KMPs–1.20 | | | OC → KMPs– 0.89 | | | OC → KMPs–0.92 |

direct effect was positive but insignificant ($\beta = 0.10$, $t = 1.57$, $p = 0.11$) for China, while the indirect effect was found significant ($\beta = 0.64$, $t = 9.45$, $p < 0.001$). Hence, the outcome shows a full mediation.

The direct effect for Pakistan ($\beta = 0.30$, $t = 0.10$, $p < 0.001$) and overall survey ($\beta = 0.17$, $t = 42.28$, $p < 0.001$) was significant. Similarly, the indirect effect KWS for Pakistan ($\beta = 0.37$, $t = 3.11$, $p < 0.001$) and overall survey ($\beta = 0.56$, $t = 9.93$, $p < 0.001$) was also significant which shows a partial mediation effect. This further reveals that the influence of KM processes on organizational performance passes through KWS. Consequently, H5 is accepted. Results of mediation analysis are presented in Table 10.

### *4.4 Multi-group analysis*

The objective of H6 was to validate whether the association between trust, organizational climate, KM processes, KWS, and organizational performance differs diagonally between the two samples. To verify the variations between the two nations in evaluation, a multi-group analysis was conducted to calculate the statistical dissimilarities between the two nations. A non-parametric PLS-MGA was led by Henseler *et al.* (2009). PLS-MGA provides evidence of invariance between two countries (Hair *et al.*, 2016).

The results showed a substantially different effect of KM processes on KWS in China and Pakistan. Thus, H6d is substantiated. However, none of the other variations was important when comparing sample countries' effect of KM enablers on KM processes and the impact of KM processes on organizational performance. Hypothesized relationship variations were considered negligible. Consequently, our results do not support hypotheses H6a, H6b, H6c, and H6e. Thus, this demonstrates that the results of KM processes are equivalent in Pakistan and China (Table 11).

## 5. Discussion, conclusion, and implications

### *5.1 Discussion*

The present study investigates the impact of KM enablers (trust and organizational climate) on KM processes, as well as the direct impact of KM processes (creation, acquisition, storage, sharing, and utilization) on the functioning of an organization in Pakistani and Chinese research universities, with the mediating effect of KWS.

| | Total effect (KMP → OP) | | | Direct effect (KMP → OP) | | | Specific indirect effect (KMP → OP) | | |
|---|---|---|---|---|---|---|---|---|---|
| H5: KMP → KWS → OP | $\beta$ | *t*-value | *p*-value | *B* | *t*-value | *p*-value | $\beta$ | *t*-value | *p*-value |
| China | 0.74 | 17.63 | 0.00 | 0.10 | 1.57 | 0.11 | 0.64 | 9.45 | 0.00 |
| Pakistan | 0.67 | 4.414 | 0.00 | 0.30 | 0.10 | 0.00 | 0.37 | 3.11 | 0.00 |
| Overall | 0.73 | 0.041 | 0.00 | 0.17 | 42.28 | 0.00 | 0.56 | 9.93 | 0.00 |

**Table 10.** Mediation results

| | Path coefficients-diff (China–Pakistan) | *p*-Value new (China vs. Pakistan) |
|---|---|---|
| H6a: T → KMPs | 0.19 | 0.47 |
| H6b: OC → KMPs | 0.12 | 0.33 |
| H6c: KMPs → OP | 0.17 | 0.58 |
| H6d: KMPs → KWS | 0.10 | *0.02* |
| H6e: KWS → OP | 0.13 | 0.39 |

**Note(s):** Italic value shows the significant differences

**Table 11.** Multi-group comparison (Hypothesis H6a, b, c, d, e)

Trust has a significant impact on KM processes. The trust establishes a critical position to authenticate profits to people and institutions (Yasir and Majid, 2017; Yasir *et al.*, 2017). A meaningful connection based on trust is found between members in HEIs, which forms a constructive connection relationship that increases the concentration to which members support and grab information between members and consent to the allocation of relevant information (Chowdhury, 2005; Levin *et al.*, 2006). The outcome is consistent with the findings of previous studies (Yasir and Majid, 2017; Yasir *et al.*, 2017).

Organizational climate is employed as a new enabler in HEIs and focuses on elements that support the making of individual connections (Mehra *et al.*, 2001; Tseng and Jung, 2011). This can be mentioned that the climate in a firm exhibits a lively function in offering an environment for relational and social exchange (Jaw and Liu, 2003; Chen and Huang, 2007; Tseng and Jung, 2011). The suggestions verify the claims made in previous research, particularly by Jain *et al.* (2015) and Rodriguez *et al.* (2016). These relate to the significance of climate in a firm to communal message amongst people.

This study ascertains that KM processes considerably impact universities' performance in research. This presents that in universities conducting research, the real commitment of KM processes can direct towards improved HEIs performance. The outcome is in proportion to the KBV theory, which mentions that KM can help administrate competent knowledge benefits and help achieve higher organizational performance (Grant, 1996). As per Rowley (2000), knowledge management is important in HEIs, and the outcome of the present research also validates his results. Also, a positive connection between organizational performance and KM processes is found in other HEIs research (Ahmad *et al.*, 2017; Latilla *et al.*, 2018; Abubakar *et al.*, 2019; Iqbal *et al.*, 2019; Meher and Mishra, 2019). Thus, a condition is attached to the KBV in HEIs that managing information resources efficiently can influence HEIs' functioning. Also, this study validates with Iqbal *et al.* (2019) that an organization's strength depends on the efficient commitment of KM processes, and the organizational performance can establish a lively reason for the possible practical advantages and increased organizational performance.

The present study results convey an important realistic evaluation of the mediating position of KWS between KM processes and organizational performance. The results show that KM processes have a meaningful and positive impact on KWS, increasing academic functioning. These outcomes are equivalent to the previous studies in concentrated knowledge division (Mládková *et al.*, 2015; Shujahat *et al.*, 2018; Sahibzada *et al.*, 2020). The ending of the present research shows that KM processes help in KWS concerning enhanced knowledge methods and collective learning and improved organizational evaluation, practice, processes, talents, capability, and peoples' cognitive methods, thus confirming KBV theory (Grant, 1996; Gold *et al.*, 2001; Shih *et al.*, 2010). In previous research, sharing information, attainment, and use has been attached to organizational performance (Wang *et al.*, 2016a, 2016b; Iqbal *et al.*, 2019). This research further states that KM processes (knowledge acquisition, creation, storage, sharing, and utilization) accelerate KWS authenticating Herzberg and Drucker's theory (Herzberg, 1966, 1974; Drucker, 1999; Kianto *et al.*, 2019; Sahibzada *et al.*, 2020). The empirical results of this investigation support that sharing knowledge and knowledge acquisition, creation, storage, and utilization can maximize KWS contentment and excellence, ensuring higher levels of organizational performance and sustainable competitive advantage in advanced educational activities.

However, the multi-group analysis described an insignificant difference in the association between KM enablers, KM processes, and organizational performance through the interceding methods of KWS across Pakistan and China. A considerable impact of KM processes on KWS was established between Pakistan and China. These outcomes showed that KM and their effectual administration processes are almost the same across diverse societies and districts in the world's surroundings. Thus, the result proves the current

research and upcoming trends on the impact of KM processes on KWS (Sahibzada *et al.*, 2020a, d, e). To conclude, KM processes with minor dissimilarities are usually performed in Pakistan and China.

### *5.2 Conclusion*

The findings of this study suggest that KM processes can significantly increase employee satisfaction, which can help promote organizational performance in HEIs. The analysis also revealed that trust and organizational climate are essential factors in KM processes and that these factors contribute to organizational performance. According to the study's findings, university administrations and policymakers should cultivate alternative techniques for knowledge leaders to ensure effective management and involvement of KM processes among workers in the future (Rehman and Iqbal, 2020; Sahibzada *et al.*, 2020b, 2021f).

Limited studies have looked into the relationship between KM processes and organizational performance in HEIs (Ahmad *et al.*, 2017; Iqbal *et al.*, 2019; Sahibzada *et al.*, 2020e). The present work addressed this gap by testing the model in China and Pakistan and the mediating effect of KWS previously overlooked in the literature. The current study demonstrates that the KM processes contribute to knowledge workers' satisfaction regarding improved knowledge methods and collective learning, organizational evaluation, practice, processes, talents, capabilities, and people's cognitive procedures. Also, not only knowledge sharing but also knowledge acquisition, development, storage, and utilization can foster KWS, resulting in a greater operational performance for HEIs. Thus, contributing to the KM literature as the impact of the KM processes on organizational performance is not yet evident in the literature. To summarize, the current study adds to the existing literature by assessing trust, organizational climate, KM processes, KWS, and organizational performance in research institutions in China and Pakistan, and the findings substantiate the hypothesis that KM processes significantly increase KWS capacity toward improved organizational performance.

### *5.3 Implications*

The study attempted to fill a gap in the existing literature by using the KBV theory to explain the influence of KM processes on HEIs performance via the mediating role of KWS. Furthermore, this study expands on an extensive literature analysis on the KBV theory by exploring the links between trust, organizational climate, KM processes, knowledge workers' satisfaction, and OP from HEIs. The study's findings imply that administrators and educational officials should embrace and develop knowledge-oriented initiatives to improve organizational climate and build a trustworthy environment, ultimately improving KM processes among academicians (Whisnant and Khasawneh, 2014; Sahibzada *et al.*, 2020c).

Previous studies have investigated the relationship between KM processes and organizational performance in research-intensive universities (Sahibzada *et al.*, 2020c, d). However, in HEIs in cross-cultural settings, the intervening instrument of KWS was not found and tested. As a result, the findings of this study demonstrated that KM processes increase knowledge workers' satisfaction, which, in turn, influences the overall performance of HEIs. This study emphasized the importance of HEIs strengthening their KM processes. Managers in HEIs should prioritize knowledge creation, acquisition, storage, sharing, and utilization to improve employee performance. The combination of KM enablers like trust and creating an organizational climate encourages the administration to invest, which can help HEIs to achieve excellent performance in this difficult area.

### *5.4 Limitations and future research directions*

Future studies suggest increasing the sample size using other data collection approaches such as time-lag or longitudinal research designs. Future researchers may also consider

additional leadership styles as KM enablers, such as sustainable leadership, leader-member exchange, empowering leadership, or servant leadership. Additionally, KOL can be evaluated in terms of its dimensions using fsQCA (Sahibzada *et al.*, 2020a), as well as other mediating variables such as innovation quality (structural, radical, incremental and disruptive), knowledge hiding, knowledge coupling, and worker performance, which are still underdeveloped in HEIs. Moreover, the studied relationship can be examined at a multi-level to estimate team-level aggregated performance.

## References


Abubakar, A.M., Elrehail, H., Alatailat, M.A. and ElçI, A. (2019), "Knowledge management, decision-making style and organizational performance", *Journal of Innovation Knowledge*, Vol. 4 No. 2, pp. 104-114.

Ahmad, N., Lodhi, M.S., Zaman, K. and Naseem, I. (2017), "Knowledge management: a gateway for organizational performance", *Journal of the Knowledge Economy*, Vol. 8 No. 3, pp. 859-876.

Ahmed, Q., Sumbal, M.S., Akhtar, M.N. and Tariq, H. (2021), "Abusive supervision and the knowledge worker productivity: the mediating role of knowledge management processes", *Journal of Knowledge Management*, Vol. 25 No. 10, pp. 2506-2522.

Al-Abdullat, B.M. and Dababneh, A. (2018), "The mediating effect of job satisfaction on the relationship between organizational culture and knowledge management in Jordanian banking sector", *Benchmarking: An International Journal*, Vol. 25 No. 2, pp. 517-544.

Alavi, M. and Leidner, D.E. (2001), "Knowledge management and knowledge management systems: conceptual foundations and research issues", *MIS Quarterly*, Vol. 25 No. 1, pp. 107-136.

Appleyard, M.M. (1996), "How does knowledge flow? Interfirm patterns in the semiconductor industry", *Strategic Management Journal*, Vol. 17 No. S2, pp. 137-154.

Bano, S. and Taylor, J. (2015), "Universities and the knowledge-based economy: perceptions from a developing country", *Higher Education Research Development*, Vol. 34 No. 2, pp. 242-255.

Barley, W.C., Treem, J.W. and Kuhn, T. (2018), "Valuing multiple trajectories of knowledge: a critical review and agenda for knowledge management research", *Academy of Management Annals*, Vol. 12 No. 1, pp. 278-317.

Bennett, R. and Gabriel, H. (1999), "Organisational factors and knowledge management within large marketing departments: an empirical study", *Journal of Knowledge Management*, Vol. 3 No. 3, pp. 212-225.

Bock, G.-W., Zmud, R.W., Kim, Y.-G. and Lee, J.-N. (2005), "Behavioral intention formation in knowledge sharing: examining the roles of extrinsic motivators, social-psychological factors, and organizational climate", *MIS Quarterly*, Vol. 29 No. 1, pp. 87-111.

Bontis, N. and Serenko, A. (2009), "Longitudinal knowledge strategising in a long-term healthcare organisation", *International Journal of Technology Management*, Vol. 47 No. 1, pp. 276-297.

Bratianu, C. and Bejinaru, R. (2017), *Knowledge strategies for increasing IC of universities. European Conference on Intellectual Capital, Academic Conferences International Limited*, 2017, pp. 34-41.

Bryant, S.E. and Terborg, J.R. (2008), "Impact of peer mentor training on creating and sharing organizational knowledge", *Journal of Managerial Issues*, Vol. 20 No. 1, pp. 11-29.

Butt, M.A., Nawaz, F., Hussain, S., Sousa, M.J., Wang, M., Sumbal, M.S. and Shujahat, M. (2018), "Individual knowledge management engagement, knowledge-worker productivity, and innovation performance in knowledge-based organizations: the implications for knowledge processes and knowledge-based systems", *Computational Mathematical Organization Theory*, Vol. 25 No. 3, pp. 336-356.

Chatterji, N. and Kiran, R. (2017), "Role of human and relational capital of universities as underpinnings of a knowledge economy: a structural modelling perspective from north Indian universities", *International Journal of Educational Development*, Vol. 56 No. 3, pp. 52-61.

Chatzoudes, D., Chatzoglou, P. and Vraimaki, E. (2015), "The central role of knowledge management in business operations: developing a new conceptual framework", *Business Process Management Journal*, Vol. 21 No. 5, pp. 1117-1139.

Chen, C.-J. and Huang, J.-W. (2007), "How organizational climate and structure affect knowledge management—the social interaction perspective", *International Journal of Information Management*, Vol. 27 No. 2, pp. 104-118.

Chiu, C.-N. and Chen, H.-H. (2016), "The study of knowledge management capability and organizational effectiveness in Taiwanese public utility: the mediator role of organizational commitment", *SpringerPlus*, Vol. 5 No. 1, pp. 1520-1554.

Chou, T.-C., Chang, P.-L., Tsai, C.-T. and Cheng, Y.-P. (2005), "Internal learning climate, knowledge management process and perceived knowledge management satisfaction", *Journal of Information Science*, Vol. 31 No. 4, pp. 283-296.

Chowdhury, S. (2005), "The role of affect-and cognition-based trust in complex knowledge sharing", *Journal of Managerial Issues*, Vol. 17 No. 3, pp. 310-326.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*, Routledge, Hillsdale, NJ.

Davenport, T.H. and Prusak, L. (1998), *Working Knowledge: How Organizations Manage what They Know*, Harvard Business Press, Boston.

Delshab, V., Winand, M., Sadeghi Boroujerdi, S., Hoeber, L. and Mahmoudian, A. (2022), "The impact of knowledge management on performance in nonprofit sports clubs: the mediating role of attitude toward innovation, open innovation, and innovativeness", *European Sport Management Quarterly*, Vol. 22 No. 2, pp. 139-160.

Dirks, K.T. and Ferrin, D.L. (2002), "Trust in leadership: meta-analytic findings and implications for research and practice", *Journal of Applied Psychology*, Vol. 87 No. 4, p. 611.

Dixon, N.M. (2000), *Common Knowledge: How Companies Thrive by Sharing what They Know*, Harvard Business School Press.

Donate, M.J. and De Pablo, J.D.S. (2015), "The role of knowledge-oriented leadership in knowledge management practices and innovation", *Journal of Business Research*, Vol. 68 No. 2, pp. 360-370.

Donate, M.J. and Guadamillas, F. (2015), "An empirical study on the relationships between knowledge management, knowledge-oriented human resource practices and innovation", *Knowledge Management Research Practice*, Vol. 13 No. 2, pp. 134-148.

Drucker, P.F. (1998), "Management's new paradigms", *Forbes Magazine*, Vol. 10 No. 2, pp. 98-99.

Drucker, P.F. (1999), "Knowledge-worker productivity: the biggest challenge", *California Management Review*, Vol. 41 No. 2, pp. 79-94.

Edmondson, A. (1999), "Psychological safety and learning behavior in work teams", *Administrative Science Quarterly*, Vol. 44 No. 2, pp. 350-383.

Falk, R.F. and Miller, N.B. (1992), *A Primer for Soft Modeling*, University of Akron Press.

Feiz, D., Dehghani Soltani, M. and Farsizadeh, H. (2019), "The effect of knowledge sharing on the psychological empowerment in higher education mediated by organizational memory", *Studies in Higher Education*, Vol. 44 No. 1, pp. 3-19.

Fernandes, V. (2012), "(Re) discovering the PLS approach in management science", *M@ n@ gement*, Vol. 15 No. 1, pp. 101-123.

Floyd, S.W. and Lane, P.J. (2000), "Strategizing throughout the organization: managing role conflict in strategic renewal", *Academy of Management Review*, Vol. 25 No. 1, pp. 154-177.

Ford, D.P. (2004), "Trust and knowledge management: the seeds of success", *Handbook on Knowledge Management 1*, Springer.

Fornell, C. and Larcker, D.F. (1981), "Evaluating structural equation models with unobservable variables and measurement error", *Journal of Marketing Research*, Vol. 18 No. 1, pp. 39-50.

Fullwood, R. and Rowley, J. (2017), "An investigation of factors affecting knowledge sharing amongst UK academics", *Journal of Knowledge Management*, Vol. 21 No. 5, pp. 1254-1271.

Gold, A.H., Malhotra, A. and Segars, A.H. (2001), "Knowledge management: an organizational capabilities perspective", *Journal of Management Information Systems*, Vol. 18 No. 1, pp. 185-214.

Grant, R.M. (1996), "Toward a knowledge-based theory of the firm", *Strategic Management Journal*, Vol. 17 No. S2, pp. 109-122.

Hair, J.F. (2006), *Multivariate Data Analysis*, Pearson Education, Upper Saddle River, NJ.

Hair, J.F., Ringle, C.M. and Sarstedt, M. (2012), "Partial least squares: the better approach to structural equation modeling?", *Long Range Planning*, Vol. 45 Nos 5-6, pp. 312-319.

Hair, J.J.F., Sarstedt, M., Hopkins, L. and Kuppelwieser, V.G. (2014), "Partial least squares structural equation modeling (PLS-SEM) an emerging tool in business research", *European Business Review*, Vol. 26 No. 2, pp. 106-121.

Hair, J.F., Hult, G.T.M., Ringle, C. and Sarstedt, M. (2016), *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*, Sage Publications, Los Angeles.

Hansen, M.T. (1999), "The search-transfer problem: the role of weak ties in sharing knowledge across organization subunits", *Administrative Science Quarterly*, Vol. 44 No. 1, pp. 82-111.

Henseler, J., Ringle, C.M. and Sinkovics, R.R. (2009), "The use of partial least squares path modeling in international marketing", in *New Challenges to International Marketing*, Emerald Group Publishing, Bingley.

Herzberg, F. (1966), *Work and the Nature of Man*, Oxford.

Herzberg, F. (1974), *The Motivation-Hygiene Concept and Problems of Manpower Personnel Administration*, BYU Electronic Media Department, Provo, UT.

Hofstede, G. (2001), "Culture's recent consequences: using dimension scores in theory and research", *International Journal of Cross Cultural Management*, Vol. 1 No. 1, pp. 11-17.

Holste, J.S. and Fields, D. (2010), "Trust and tacit knowledge sharing and use", *Journal of Knowledge Management*, Vol. 14 No. 1, pp. 128-140.

Huang, J.-W. and Li, Y.-H. (2009), "The mediating effect of knowledge management on social interaction and innovation performance", *International Journal of Manpower*, Vol. 30 No. 3, pp. 285-301.

Huff, L. and Kelley, L. (2003), "Levels of organizational trust in individualist versus collectivist societies: a seven-nation study", *Organization Science*, Vol. 14 No. 1, pp. 81-90.

Iqbal, A., Latif, F., Marimon, F., Sahibzada, U.F. and Hussain, S. (2019), "From knowledge management to organizational performance: modelling the mediating role of innovation and intellectual capital in higher education", *Journal of Enterprise Information Management*, Vol. 32 No. 1, pp. 36-59.

Jain, K.K., Sandhu, M.S. and Goh, S.K. (2015), "Organizational climate, trust and knowledge sharing: insights from Malaysia", *Journal of Asia Business Studies*, Vol. 9 No. 1, pp. 54-77.

Janz, B.D. and Prasarnphanich, P. (2003), "Understanding the antecedents of effective knowledge management: the importance of a knowledge-centered culture", *Decision Sciences*, Vol. 34 No. 2, pp. 351-384.

Janz, B.D., Colquitt, J.A. and Noe, R.A. (1997), "Knowledge worker team effectiveness: the role of autonomy, interdependence, team development, and contextual support variables", *Personnel Psychology*, Vol. 50 No. 4, pp. 877-904.

Jaw, B.S. and Liu, W. (2003), "Promoting organizational learning and self-renewal in Taiwanese companies: the role of HRM", *Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan in alliance with the Society of Human Resources Management*, Vol. 42 No. 3, pp. 223-241.

Jones, M.C., Cline, M. and Ryan, S. (2006), "Exploring knowledge sharing in ERP implementation: an organizational culture framework", *Decision Support Systems*, Vol. 41 No. 2, pp. 411-434.

Kianto, A., Vanhala, M. and Heilmann, P. (2016), "The impact of knowledge management on job satisfaction", *Journal of Knowledge Management*, Vol. 20 No. 4, pp. 621-636.

Kianto, A., Shujahat, M., Hussain, S., Nawaz, F. and Ali, M. (2019), "The impact of knowledge management on knowledge worker productivity", *Baltic Journal of Management*, Vol. 14 No. 2, pp. 178-197.

Kline, R.B. (2011), *Principles and Practice of Structural Equation Modeling*, The Guilford Press, New York, NY.

Koohang, A., Paliszkiewicz, J. and Goluchowski, J. (2017), "The impact of leadership on trust, knowledge management, and organizational performance: a research model", *Industrial Management Data Systems*, Vol. 117 No. 3, pp. 521-537.

Kulkarni, U.R., Ravindran, S. and Freeze, R. (2006), "A knowledge management success model: theoretical development and empirical validation", *Journal of Management Information Systems*, Vol. 23 No. 3, pp. 309-347.

Laal, M. (2011), "Lifelong learning: what does it mean?", *Procedia-Social Behavioral Sciences*, Vol. 28 No. 2011, pp. 470-474.

Latilla, V.M., Frattini, F., Petruzzelli, A.M. and Berner, M. (2018), "Knowledge management, knowledge transfer and organizational performance in the arts and crafts industry: a literature review", *Journal of Knowledge Management*, Vol. 22 No. 6, pp. 1310-1331.

Lee, K.C., Lee, S. and Kang, I.W. (2005), "KMPI: measuring knowledge management performance", *Journal of Information Management*, Vol. 42 No. 3, pp. 469-482.

Lee, M.K., Cheung, C.M., Lim, K.H. and Ling Sia, C. (2006), "Understanding customer knowledge sharing in web-based discussion boards: an exploratory study", *Internet Research*, Vol. 16 No. 3, pp. 289-303.

Lee, Gillespie, N., Mann, L. and Wearing, A. (2010), "Leadership and trust: their effect on knowledge sharing and team performance", *Management Learning*, Vol. 41 No. 4, pp. 473-491.

Lei, H., Nguyen, T.T. and Le, P.B. (2019), "How knowledge sharing connects interpersonal trust and innovation capability: the moderating effect of leadership support", *Chinese Management Studies*, Vol. 13 No. 2, pp. 276-298.

Levin, D.Z., Whitener, E.M. and Cross, R. (2006), "Perceived trustworthiness of knowledge sources: the moderating impact of relationship length", *Journal of Applied Psychology*, Vol. 91 No. 5, pp. 1163-1171.

Lewis, R. (2010), *When Cultures Collide: Leading across Cultures*, Nicholas Brealey International, Boston and London.

Lim, K.K., Ahmed, P.K. and Zairi, M. (1999), "Managing for quality through knowledge management", *Total Quality Management*, Vol. 10 Nos 4-5, pp. 615-621.

Lin, H.-F. and Lee, G.-G. (2006), "Effects of socio-technical factors on organizational intention to encourage knowledge sharing", *Management Decision*, Vol. 44 No. 1, pp. 74-88.

Liu, C.-H., Wang, J.S. and Lin, C.-W. (2017), "The concepts of big data applied in personal knowledge management", *Journal of Knowledge Management*, Vol. 21 No. 1, pp. 213-230.

Lo, W.Y.W. (2016), "The concept of greater China in higher education: adoptions, dynamics and implications", *Comparative Education*, Vol. 52 No. 1, pp. 26-43.

Magner, N., Johnson, G.G. and Elfrink, J. (1994), "Evidence on the relationship between procedural and distributive justice in performance appraisal and accounting faculty attitudes and performance", *Journal of Accounting Education*, Vol. 12 No. 4, pp. 325-341.

Mahamed Ismail, N.A., Welch, C. and Xu, M. (2015), "Towards a sustainable quality of university research: knowledge sharing", *Knowledge Management Research Practice*, Vol. 13 No. 2, pp. 168-177.

Mahdavi Mazdeh, M. and Hesamamiri, R. (2014), "Knowledge management reliability and its impact on organizational performance: an empirical study", *Program*, Vol. 48 No. 2, pp. 102-126.

Martelo-Landroguez, S. and Cepeda-Carrión, G. (2016), "How knowledge management processes can create and capture value for firms?", *Knowledge Management Research Practice*, Vol. 14 No. 4, pp. 423-433.

Masa'deh, R.E., Obeidat, B.Y. and Tarhini, A. (2016), "A Jordanian empirical study of the associations among transformational leadership, transactional leadership, knowledge sharing, job performance, and firm performance: a structural equation modelling approach", *Journal of Management Development*, Vol. 35 No. 5, pp. 681-705.

Masa'deh, R.E., Shannak, R., Maqableh, M. and Tarhini, A. (2017), "The impact of knowledge management on job performance in higher education: the case of the University of Jordan", *Journal of Enterprise Information Management*, Vol. 30 No. 2, pp. 244-262.

Mayer, R.C., Davis, J.H. and Schoorman, F.D. (1995), "An integrative model of organizational trust", *Academy of Management Review*, Vol. 20 No. 3, pp. 709-734.

Mcneish, J. and Mann, I.J.S. (2010), "Knowledge sharing and trust in organizations", *IUP Journal of Knowledge Management*, Vol. 8 Nos 1/2, pp. 18-38.

Meher, J.R. and Mishra, R.K. (2019), "Assessing the influence of knowledge management practices on organizational performance", *VINE Journal of Information Knowledge Management Systems*, Vol. 49 No. 3, pp. 440-456.

Mehra, A., Kilduff, M. and Brass, D.J. (2001), "The social networks of high and low self-monitors: implications for workplace performance", *Administrative Science Quarterly*, Vol. 46 No. 1, pp. 121-146.

Mládková, L., Zouharová, J. and Nový, J. (2015), "Motivation and knowledge workers", *Procedia-Social Behavioral Sciences*, Vol. 207 No. 4, pp. 768-776.

Moorman, C., Zaltman, G. and Deshpande, R. (1992), "Relationships between providers and users of market research: the dynamics of trust within and between organizations", *Journal of Marketing Research*, Vol. 29 No. 3, pp. 314-328.

Munir, H., Jianfeng, C. and Ramzan, S. (2019), "Personality traits and theory of planned behavior comparison of entrepreneurial intentions between an emerging economy and a developing country", *International Journal of Entrepreneurial Behavior*, Vol. 25 No. 3, pp. 554-580.

Muqadas, F., Rehman, M., Aslam, U. and Ur-Rahman, U.-. (2017), "Exploring the challenges, trends and issues for knowledge sharing: a study on employees in public sector universities", *VINE Journal of Information Knowledge Management Systems*, Vol. 47 No. 1, pp. 2-15.

Ngah, R., Tai, T. and Bontis, N. (2016), "Knowledge management capabilities and organizational performance in roads and transport authority of Dubai: the mediating role of learning organization", *Knowledge Process Management*, Vol. 23 No. 3, pp. 184-193.

Nonaka, I. and Takeuchi, H. (1995), *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*, Oxford University Press.

Obeidat, B.Y., Al-Suradi, M.M., Masa'deh, R.E. and Tarhini, A. (2016), "The impact of knowledge management on innovation: an empirical study on Jordanian consultancy firms", *Management Research Review*, Vol. 39 No. 10, pp. 1214-1238.

Palvalin, M. (2017), "How to measure impacts of work environment changes on knowledge work productivity–validation and improvement of the SmartWoW tool", *Measuring Business Excellence*, Vol. 21 No. 2, pp. 175-190.

Palvalin, M., Van Der Voordt, T. and Jylhä, T. (2017), "The impact of workplaces and self-management practices on the productivity of knowledge workers", *Journal of Facilities Management*, Vol. 15 No. 4, pp. 423-438.

Pang, K. and Lu, C.-S. (2018), "Organizational motivation, employee job satisfaction and organizational performance: an empirical study of container shipping companies in Taiwan", *Maritime Business Review*, Vol. 3 No. 1, pp. 36-52.

Pérez, L.S., Manuel, M.P.J. and José, V.O.C. (2004), "Managing knowledge: the link between culture and organizational learning", *Journal of Knowledge Management*, Vol. 8 No. 6, pp. 93-104.

Peng, J., Moffett, S. and Mcadam, R. (2010), "Knowledge management in China: a review", *Journal of Technology Management in China*, Vol. 5 No. 2, pp. 158-175.

Porter, L.W., Steers, R.M., Mowday, R.T. and Boulian, P.V. (1974), "Organizational commitment, job satisfaction, and turnover among psychiatric technicians", *Journal of Applied Psychology*, Vol. 59 No. 5, pp. 603-609.

Ramjeawon, V.P. and Rowley, J. (2018), "Knowledge management in higher education institutions in Mauritius", *International Journal of Educational Management*, Vol. 32 No. 7, pp. 1319-1332.

Razmerita, L., Kirchner, K. and Nielsen, P. (2016), "What factors influence knowledge sharing in organizations? A social dilemma perspective of social media communication", *Journal of Knowledge Management*, Vol. 20 No. 6, pp. 1225-1246.

Rehman, U.U. and Iqbal, A. (2020), "Nexus of knowledge-oriented leadership, knowledge management, innovation and organizational performance in higher education", *Business Process Management Journal*, Vol. 26 No. 6, pp. 1731-1758.

Ringle, C.M., Wende, S. and Will, A. (2005), *SmartPLS 2.0 (M3) Beta*, Universityof Hamburg, Hamburg.

Ringle, C.M., Sarstedt, M., Mitchell, R. and Gudergan, S.P. (2018), "Partial least squares structural equation modeling in HRM research", *The International Journal of Human Resource Management*, Vol. 1 No. 1, pp. 1-27.

Rios-Ballesteros, N. and Fuerst, S. (2021), "Exploring the enablers and microfoundations of international knowledge transfer", *Journal of Knowledge Management*, Vol. 26 No. 7, pp. 1868-1898.

Rodriguez, I.A., Garcia, A., Morais, S.C., Muniz, J. and Munyon, T.P. (2016), "Workers' perspective about organizational climate in knowledge management: automotive assembly-line case", *IFIP International Conference on Advances in Production Management Systems*, Springer, pp. 723-730.

Rowley, J. (2000), "Is higher education ready for knowledge management?", *International Journal of Educational Management*, Vol. 14 No. 7, pp. 325-333.

Sahibzada, U.F., Jianfeng, C., Latif, F. and Shafait, Z. (2019), "Development and validation of a multidimensional instrument for measuring internal marketing in Chinese higher education", *Journal of Enterprise Information Management*, Vol. 32 No. 3, pp. 413-435.

Sahibzada, U.F., Cai, J., Latif, K.F. and Sahibzada, H.F. (2020), "Knowledge management processes, knowledge worker satisfaction, and organizational performance", *Aslib Journal of Information Management*, Vol. 72 No. 1, pp. 112-129.

Sahibzada, U.F., Cai, J., Latif, K.F. and Sahibzada, H.F. (2020a), "Knowledge management processes, knowledge worker satisfaction, and organizational performance: symmetric and asymmetrical analysis", *Aslib Journal of Information Management*, Vol. 72 No. 1, pp. 112-129.

Sahibzada, U.F., Cai, J., Latif, K.F., Shafait, Z. and Sahibzada, H.F. (2020b), "Interpreting the impact of knowledge management processes on organizational performance in Chinese higher education: mediating role of knowledge worker productivity", *Studies in Higher Education*, Vol. 47 No. 4 pp. 713-730.

Sahibzada, U.F., Cai, J., Latif, K.F., Shah, S.A. and Sahibzada, H.F. (2020c), "Refuelling knowledge management processes towards organisational performance: mediating role of creative organisational learning", *Knowledge Management Research Practice*.

Sahibzada, U.F., Jianfeng, C., Latif, K.F. and Sahibzada, H.F. (2020d), "Fueling knowledge management processes in Chinese higher education institutes (HEIs): the neglected mediating role of knowledge worker satisfaction", *Journal of Enterprise Information Management*, Vol. 33 No. 6, pp. 1395-1417.

Sahibzada, U.F., Latif, K.F., Xu, Y. and Khalid, R. (2020e), "Catalyzing knowledge management processes towards knowledge worker satisfaction: fuzzy-set qualitative comparative analysis", *Journal of Knowledge Management*, Vol. 24 No. 10, pp. 2373-2400.

Sahibzada, H.F., Jianfeng, C., Sahibzada, U.F., Khalid, R. and Afshan, G. (2021), "Unpacking knowledge management and organizational performance: a comparison between emerging and developing countries", *Aslib Journal of Information Management*, Vol. 73 No. 6, pp. 793-813.

Sahibzada, U.F., Latif, K.F. and Xu, Y. (2021f), "Symmetric and asymmetric modeling of knowledge management enablers to knowledge management processes and knowledge worker productivity in higher education institutes", *Journal of Enterprise Information Management*, Vol. 35 No. 3, pp. 729-756.

Sahibzada, U.F., Xu, Y., Afshan, G. and Khalid, R. (2021g), "Knowledge-oriented leadership towards organizational performance: symmetrical and asymmetrical approach", *Business Process Management Journal*, Vol. 27 No. 6, pp. 1720-1746.

Sarstedt, M., Ringle, C.M., Henseler, J. and Hair, J.F. (2014), "On the emancipation of PLS-SEM: a commentary on Rigdon (2012)", *Long Range Planning*, Vol. 47 No. 3, pp. 154-160.

Schmitz, C., Van Meijl, H., Kyle, P., Nelson, G.C., Fujimori, S., Gurgel, A., Havlik, P., Heyhoe, E., D'croz, D.M. and Popp, A. (2014), "Land-use change trajectories up to 2050: insights from a global agro-economic model comparison", *Journal of Agricultural Economics*, Vol. 45 No. 1, pp. 69-84.

Seleim, A. and Khalil, O. (2007), "Knowledge management and organizational performance in the Egyptian software firms", *International Journal of Knowledge Management*, Vol. 3 No. 4, pp. 37-66.

Shafait, Z., Khan, M.A., Sahibzada, U.F., Dacko-Pikiewicz, Z. and Popp, J. (2021), "An assessment of students' emotional intelligence, learning outcomes, and academic efficacy: a correlational study in higher education", *Plos One*, Vol. 16 No. 8, p. e0255428.

Shahzad, K., Bajwa, S.U., Siddiqi, A.F.I., Ahmid, F. and Raza Sultani, A. (2016), "Integrating knowledge management (KM) strategies and processes to enhance organizational creativity and performance: an empirical investigation", *Journal of Modelling in Management*, Vol. 11 No. 1, pp. 154-179.

Shahzadi, A., Li, S., Sahibzada, U.F., Malik, M., Khalid, R. and Afshan, G. (2021), "The dynamic relationship of knowledge management processes and project success: modeling the mediating role of knowledge worker satisfaction", *Business Process Management Journal*, Vol. 27 No. 6, pp. 1657-1676.

Sharma, N. (2008), "The origin of the 'data information knowledge Wisdom' (DIKW) Hierarchy", February.

Shih, K.-H., Chang, C.-J. and Lin, B. (2010), "Assessing knowledge creation and intellectual capital in banking industry", *Journal of Intellectual Capital*, Vol. 11 No. 1, pp. 74-89.

Shujahat, M., Ali, B., Nawaz, F., Durst, S. and Kianto, A. (2018), "Translating the impact of knowledge management into knowledge-based innovation: the neglected and mediating role of knowledge-worker satisfaction", *Human Factors Ergonomics in Manufacturing Service Industries*, Vol. 28 No. 4, pp. 200-212.

Singh, A.K. and Sharma, V. (2011), "Knowledge management antecedents and its impact on employee satisfaction: a study on Indian telecommunication industries", *The Learning Organization*, Vol. 18 No. 2, pp. 115-130.

Smith, F.J. (1976), *The Index of Organizational Reactions (IOR)*, American Psychological Association, Washington.

Sumbal, M.S., Tsui, E., Irfan, I., Shujahat, M., Mosconi, E. and Ali, M. (2019), "Value creation through big data application process management: the case of the oil and gas industry", *Journal of Knowledge Management*, Vol. 23 No. 8, pp. 1566-1585.

Sveiby, K.-E. and Simons, R. (2002), "Collaborative climate and effectiveness of knowledge work–an empirical study", *Journal of Knowledge Management*, Vol. 6 No. 5, pp. 420-433.

Szulanski, G. (1996), "Exploring internal stickiness: Impediments to the transfer of best practice within the firm", *Strategic Management Journal*, Vol. 17 No. S2, pp. 27-43.

Thani, F.N. and Mirkamali, S.M. (2018), "Factors that enable knowledge creation in higher education: a structural model", *Data Technologies Applications*, Vol. 52 No. 3, pp. 424-444.

Thomas, A. and Gupta, V. (2021), "The role of motivation theories in knowledge sharing: an integrative theoretical reviews and future research agenda", *Kybernetes*, Vol. 51 No. 1, pp. 116-140.

Thomas, A. and Paul, J. (2019), "Knowledge transfer and innovation through university-industry partnership: an integrated theoretical view", *Knowledge Management Research Practice*, Vol. 17 No. 4, pp. 436-448.

Trivella, L. and Dimitrios, N.K. (2015), "Knowledge management strategy within the higher education", *The Case of GreeceProcedia-Social Behavioral Sciences*, Vol. 175 No. 2, pp. 488-495.

Tseng, S.-M. (2010), "The correlation between organizational culture and knowledge conversion on corporate performance", *Journal of Knowledge Management*, Vol. 14 No. 2, pp. 269-284.

Tseng, F.-C. and Jung, F.Y. (2011), "Exploring the influence of organizational ethical climate on knowledge management", *Journal of Business Ethics*, Vol. 101 No. 2, pp. 325-342.

Turner, Y. and Acker, A. (2017), *Education in the New China: Shaping Ideas at Work*, Routledge, New York.

Turriago-Hoyos, A., Thoene, U. and Arjoon, S. (2016), "Knowledge workers and virtues in Peter Drucker's management theory", *SAGE Open*, Vol. 6 No. 1, pp. 1-9.

Veer-Ramjeawon, P. and Rowley, J. (2019), "Embedding knowledge management in higher education institutions (HEIs): a comparison between two countries", *Studies in Higher Education*, Vol. 1 No. 2, pp. 1-17.

Viñas-Bardolet, C., Torrent-Sellens, J. and Guillen-Royo, M. (2018), "Knowledge workers and job satisfaction: evidence from Europe", *Journal of the Knowledge Economy*, Vol. 1 No. 2, pp. 1-25.

Wang, Z., Sharma, P.N. and Cao, J. (2016a), "From knowledge sharing to firm performance: a predictive model comparison", *Journal of Business Research*, Vol. 69 No. 10, pp. 4650-4658.

Wang, Z., Wang, N., Cao, J. and Ye, X. (2016b), "The impact of intellectual capital–knowledge management strategy fit on firm performance", *Management Decision*, Vol. 54 No. 8, pp. 1861-1885.

Weiss, D.J., Dawis, R.V. and England, G.W. (1967), "Manual for the Minnesota satisfaction questionnaire", *Minnesota Studies in Vocational Rehabilitation*, Vol. 22 No. 1, pp. 1-120.

Whisnant, B. and Khasawneh, O. (2014), "The influence of leadership and trust on the sharing of tacit knowledge: exploring a path model", *Journal of Business Studies Quarterly*, Vol. 6 No. 2, pp. 1-15.

Wiig, K.M. (1999), *Introducing Knowledge Management into the Enterprise*, CRC Press, Boca Raton, FL.

Wong, K.K.-K. (2013), "Partial least squares structural equation modeling (PLS-SEM) techniques using SmartPLS", *Marketing Bulletin*, Vol. 24 No. 1, pp. 1-32.

Yasir, M. and Majid, A. (2017), "Impact of knowledge management enablers on knowledge sharing: is trust a missing link in SMEs of emerging economies?", *World Journal of Entrepreneurship, Management Sustainable Development*, Vol. 13 No. 1, pp. 16-33.

Yasir, M., Majid, A. and Yasir, M. (2017), "Nexus of knowledge-management enablers, trust and knowledge-sharing in research universities", *Journal of Applied Research in Higher Education*, Vol. 9 No. 3, pp. 424-438.

Zack, M., Mckeen, J. and Singh, S. (2009), "Knowledge management and organizational performance: an exploratory analysis", *Journal of Knowledge Management*, Vol. 13 No. 6, pp. 392-409.

Zand, D.E. (1972), "Trust and managerial problem solving", *Administrative Science Quarterly*, Vol. 17 No. 2, pp. 229-239.

Zander, U. and Kogut, B. (1995), "Knowledge and the speed of the transfer and imitation of organizational capabilities: an empirical test", *Organization Science*, Vol. 6 No. 1, pp. 76-92.

Zutshi, A., Creed, A., Bhattacharya, A., Bavik, A., Sohal, A. and Bavik, Y.L. (2021), "Demystifying knowledge hiding in academic roles in higher education", *Journal of Business Research*, Vol. 137 No. 2021, pp. 206-221.

## Appendix

| |
|---|
| *Trust* |
| There is a very high level of trust throughout this university |
| In this university, subordinates have a great deal of trust for seniors |
| If someone in this university makes a promise, others within the university will almost always trust that the person will do his or her best to keep the promise |
| Seniors in this company trust their subordinates to make good decisions |
| *Organizational Climate* |
| Members in my department keep close ties with each other |
| Members in my department consider another members' standpoint highly |
| Members in my department have a strong feeling of 'one team' |
| Members in my department cooperate well with each other |
| My department encourages suggesting ideas for new opportunities |
| My department puts much value on taking risks even if that turns out to be a failure |
| My department encourages finding new methods to perform a task |
| I can trust my leader evaluation to be good |
| Objectives which are given to me are reasonable |
| My leader does not show favoritism to any one |
| *Knowledge Management Processes* |
| *Creation* |
| My university workers constantly generate new ideas |
| My university workers adapt their work to meet customer requirements |
| Members of my team actively talk with each other and share knowledge |
| My department transforms individual knowledge to shared knowledge |
| Members of my department regularly share knowledge with other teams |
| My department regularly creates innovative processes |
| *Acquisition* |
| Knowledge is obtained from students |
| Knowledge is obtained from employee |
| Knowledge is obtained from partners/other stakeholders (Media, Education, Communication, Agencies) |
| *Storage* |
| The department I work for uses the databases, repositories and information technology applications to store the knowledge for easy access by all lecturers |
| The department I work for uses various written devices such as newsletter, manuals to store the knowledge which capture from the lecturers |
| The department I work for has several publications to display the capture knowledge |
| The department I work for has several mechanisms to store the knowledge for patent and copyright |
| *Sharing* |
| My university makes constantly updated information available to me |
| My university has systems in place that efficiently capture workers' knowledge |
| My university is highly committed to research and development |
| My university does all it can to launch new products and services |
| *Utilization* |
| There exist incentive and benefit policies for new idea suggestions in utilizing existing knowledge |
| Our university effectively manages different source and type of knowledge |
| Workflow diagrams are required and used in performing tasks |
| Our university utilizes available knowledge in improving service provided to its students |
| Our university applied available knowledge to improve its performance |
| *Knowledge Worker Satisfaction* |
| *Nature of Work* |
| I have chance to try my own methods of doing the job |
| I am satisfied with the nature of work given to me |
| I have freedom to use my own judgment |

**Table A1.**
Questionnaire

(*continued*)

| |
|---|
| *Satisfaction with Supervisor* |
| I am satisfied with my current leader |
| My leader is a role model for me |
| Whenever need arises my leader provides me counseling |
| My leader works with me as a mentor |
| *Justice* |
| The rules used to evaluate my performance are fair |
| The procedures used to evaluate my performance are fair |
| The policies used to evaluate my performance are fair |
| I find that my values and the organization's (university) values are very similar |
| *Organizational Commitment* |
| I am proud to tell others that I am part of this organization (university) |
| I really care about the fate of this organization (university) |
| I am Satisfied with the performance appraisal system being implemented in my organization (university) |
| *Organizational Performance* |
| There has been an integral improvement in the operations, administrations, services and overall university performance |
| There has been an integral improvement in the relationship between the university and its students (e.g. no of students, student retention rates, student satisfaction, etc.) |
| There has been an integral improvement in university effectiveness and efficiency (e.g. timing of launching new programs, quality control or management procedures, etc.) |
| There has been an integral improvement in resource development (e.g. student skills and personnel development, etc.) |
| There has been an integral improvement in preparing for the future of the university (e.g. quality/depth of planning, indicators of partnerships and alliances, anticipating and preparing for changes in the environment, etc.) |

**Table A1.**

**Corresponding author**
Umar Farooq Sahibzada can be contacted at: umar_sahibzada@hotmail.com

For instructions on how to order reprints of this article, please visit our website:
**www.emeraldgrouppublishing.com/licensing/reprints.htm**
Or contact us for further details: **permissions@emeraldinsight.com**

Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Nonlinear Interaction of 3D Kinetic Alfv'en Wave with Null Point and Turbulence Generation in the Solar Corona.

**Garima Patel** ( esz208161@iitd.ac.in )
Indian Institute of Technology Delhi
**Neha Pathak**
Delhi Technological University
**R. Uma**
Indian Institute of Technology Delhi
**R.P. Sharma**
Indian Institute of Technology Delhi

**Research Article**

**Keywords:** Solar corona heating–Sun: corona– turbulence, Wave, Null points

**Posted Date:** September 6th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-2014435/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Solar Physics
DOI: 10.1007/•••••-•••-•••-••••-•

# Nonlinear Interaction of 3D Kinetic Alfvén Wave with Null Point and Turbulence Generation in the Solar Corona *Solar Physics*

**Garima Patel [1] · Neha Pathak [2] · R. Uma [1] · R. P. Sharma [1]**

© Springer ••••

**Abstract** In the present paper, we have studied the nonlinear Kinetic Alfvén Waves (KAW) in the vicinity of the null point. We have considered the nonlinearity due to ponderomotive effects associated with KAW in the solar corona. A 3D model equation representing the dynamics of KAW is developed in this Null point scenario. Using numerical methods, we have solved the model equation for solar corona parameters. The pseudospectral method and the finite difference method have been applied to tackle spatial integration and temporal evaluation, respectively. The outcome of the simulation demonstrates the formation of localized structures. With the evolution of time, these localized structures become more chaotic. Chaotic structures (turbulent) can efficiently transfer energy. The power spectrum of these turbulent structures shows the Kolmogorov spectral index of nearly $-\frac{5}{3}$ in the inertial range followed by a steeper spectrum of nearly $-3.3$ (in the range of $-2$ to $-4$). These structures also lead to the generation of the current sheet. To understand the physics of our model, we have also done a semi-analytical study for our model equation. Semi-analytical calculations reveal that the current sheet structures have scale sizes of the order of ion gyro-radius. The relevance of this investigation to the current observations by Parker Solar Probe has also been discussed.

✉ Garima Patel
esz208161@iitd.ac.in

✉ Neha Pathak
npsneha90@gmail.com

✉ R. Uma
ruma@ces.iitd.ac.in

✉ R. P. Sharma
rpsharma@ces.iitd.ac.in

[1] Department of Energy Science and Engineering, Indian Institute of Technology Delhi, New Delhi 110016, India

[2] Department of Applied Physics, Delhi Technological University, Shahbad Daulatpur, 5 Bawana Road, New Delhi-110042, India

**Keywords:** Solar corona heating–Sun: corona– turbulence, Wave, Null points

## 1. Introduction

A big scientific mystery of solar physics is how the solar corona is hotter than the photosphere. The explanations of this have been proposed by various mechanisms; however, a concise theory has yet to be developed to explain the qualitative and quantitative description in detail. Energy transport and dissipation of waves coming from the disturbances of photospheric and lower surfaces with the role of shuffling magnetic field lines are now undoubtedly accepted to be the main resources of this much energy rise in the solar corona region. Models proposed by Biermann, 1946; Schwarzschild, 1948 in which the convection zone generates acoustic waves and the outward transport of mechanical energy passing through the photosphere meets the need for the non-radiative energy required for heating the corona (Stein and Leibacher, 1974). Turbulence-based coronal heating models (Van Ballegooijen *et al.*, 2011; Dahlburg *et al.*, 2012, 2016; Einaudi *et al.*, 1996; Strauss, 1976) predict that transverse magnetic fluctuations propagating in strong axial magnetic field directions merge with nonlinear effects to produce turbulent energy cascading with Kolmogorov $-\frac{5}{3}$ scale.

The ubiquitous nature of Alfvenic Waves has ensured that they remain the focus of increasingly detailed laboratory and space investigations (Gekelman, 1999). Dissipation of magnetic waves plays a crucial role in coronal heating due to efficient transmission of the Alfvén waves to the solar atmosphere (Hollweg, 1978) and causes the corona to heat (Wentzel, 1974)). Alfvén Waves are supposed to be involved in coronal heating but due to propagating through a dispersive medium like solar corona, it changes their angular trajectory and appeared as a KAW. The existence of a strong magnetic field background is observed to cascade perpendicular directional energy and cluster the energy fluctuations (Matthaeus and Lamkin, 1986; Matthaeus, Goldstein, and Roberts, 1990). This perpendicular directionality preferred by the KAW on stage plays a crucial role. For example, (Carbone, Veltri, and Mangeney, 1990; Oughton, Priest, and Matthaeus, 1994; Shebalin, Matthaeus, and Montgomery, 1983).

Kinetic Alfvén Waves are of great interest because of their essential role in various space and astrophysical plasma phenomena (Wu and Chao, 2004). Mechanisms such as resonance absorption (Ionson, 1978) and phase mixing (Heyvaerts and Priest, 1983) enhance viscous or resistive energy dissipation, but they take too long in the corona. However, the KAW has an electric field component in the same direction as the background magnetic field. This electric field component has been shown to interact with plasma particles and stochastically energized them up to the order of 100 eV (equivalent to a temperature rise of $\sim 10^6$ K) within a very small time in accordance with the amplitude of the turbulence present. The power per unit area, which is absorbed by the plasma particles from this electric field is found to be proportional to the third power of the velocity amplitude and this power is of the order of the energy associated with the turbulent cascade (Malara *et al.*, 2019). This shows the KAW to be at the

centre of coronal heating theory, but the generation of turbulence by the KAW needs more study.

The critical step in coronal heating is to recognise a mechanism for converting the magnetic energy to plasma heating. Since classical dissipation coefficients are negligible in the corona because the coronal plasma is low collisional, plasma heating requires the generation of steeper gradients in tiny spatial scales (Malara *et al.*, 2019). Some numerical simulations have explained how the generation of current sheets occurs, but these models and simulations still need more improvements.

Due to complex magnetic field structures, the topological inhomogeneity of magnetic fields is being sprayed over the turbulent plasma of the solar corona. Since measuring the coronal magnetic field is not an easy task, so the magnetic field emerging from the photosphere is extrapolated to observe the coronal magnetic field structures (Van Doorsselaere, Nakariakov, and Verwichte, 2008). Extrapolation of magnetic potential field (Brown and Priest, 2001) provides the existence of locations where the vanishing of the magnetic field and Alfvén speed are found (Dungey, 1958; Sweet, 1958). These specific topological structures originated due to randomness of the field lines, turning up from the lower solar atmospheres are known as the Null point (McLaughlin and Hood, 2005). The study of MHD Waves around the null points is a fundamental coronal process. Null points are observed to be significantly contributing to drive some high energetic events like solar flares (Shibata and Magara, 2011). MHD Wave perturbations and reconnection are omnipresent in the solar corona. These null points are frequently seen around the magnetic reconnection sites i.e. the null points are common at the reconnection sites. The accumulation of currents near null points has been proposed by (McLaughlin and Hood, 2004), which increases the importance of the null point in dissipative events by which wave energy can be translated into heat.

Therefore, we seek the contribution of null points in coronal heating. The presence of null points in the solar corona being proportional to the complexity of the magnetic flux distribution is inevitable (Close, Parnell, and Priest, 2004; Longcope and Parnell, 2009). Thus, the encounter of MHD waves and magnetic topology at some point will undoubtedly occur; i.e., MHD wave propagation will interact with coronal null points. The earlier study of MHD wave behaviour in the vicinity of magnetic null points often provides critical insights into other areas of plasma physics, including mode conversion, reconnection, and intensification of current (Galsgaard and Nordlund, 1997; Pontin and Galsgaard, 2007; Rickard and Titov, 1996). Different types of plasma waves and turbulence have been seen in the vicinity of null points in the coronal reconnection sites (Thurgood and McLaughlin, 2013). KAW with high amplitudes that cause pondermotive non-linearity have also been investigated, but the effect of the presence of null points has not been studied yet. The Kinetic Alfvén Waves (KAW) and the magnetic null point interact to redistribute the energy and cascade it with a steeper power spectrum gradient.

In this paper, the developed equations and numerical simulations have been studied to investigate the generation of turbulence from the magnetic reconnection sites of the corona region. The study of the non-linear interaction of

Kinetic Alfvén Wave in the vicinity of a Null point profile has been performed by introducing pondermotive density perturbation.

This paper is organised as follows- Section 2 gives the basic development of the model equation. Section 3 describes the numerical simulation technique used. Further, section 4 applies a semi-analytical method to describe the physics behind localization, and the scale size of these structures is calculated. Section 5 discusses the summary and conclusions of the present paper.

## 2. The Model

Consider a 3D Kinetic Alfvén Wave which have wave vector $\boldsymbol{k}_0 = \boldsymbol{k}_{0x}\hat{x} + \boldsymbol{k}_{0y}\hat{y} + \boldsymbol{k}_{0z}\hat{z}$. This wave is propagating in a medium having the background magnetic field along the z-axis, i.e. $B_0\hat{z}$. We assume that there is an additional magnetic field $\delta\boldsymbol{B}$ due to presence of a null point as $\frac{B_0}{L}(x, -y, 0)$. Here, L is the length scale for the variation of the magnetic field. Therefore, total magnetic field is given as $\boldsymbol{B} = B_0\hat{z} + \frac{B_0}{L}(x\hat{x} - y\hat{y})$. The electric and magnetic fields can be expressed as $\boldsymbol{E} = -\nabla\phi - \frac{1}{c}\frac{\partial\tilde{A}_z}{\partial t}$ and $\boldsymbol{B}_\perp = \hat{z} \times \nabla\tilde{A}_z$ , here $\phi$ is the scalar potential of the field and $\tilde{A}_z$ is the parallel component of the vector potential, respectively. The speed of light is denoted as c.

The ion equation of motion can be written as

$$\frac{\partial \boldsymbol{v}_i}{\partial t} = -\frac{e}{m_i}\left(\nabla\phi + \frac{1}{c}\frac{\partial\tilde{A}_z}{\partial t}\right) - \frac{e}{m_i c}(\boldsymbol{v}_i \times \boldsymbol{B}_0 + \boldsymbol{v}_i \times \delta\boldsymbol{B}) - \frac{k_B T_i}{n_0 m_i}\nabla n \qquad (1)$$

Here $\boldsymbol{v}_i$ is the ion velocity, $k_B$ is the Boltzmann constant, $T_i$ is the ion temperature, $n_0$ is the background density, and $n$ is the density modification in the background density. The absolute value of charge for electron and ion are denoted as e and mass of ion as $m_i$. For perpendicular component, taking cross product of equation (1) with $\hat{z}$.

$$\begin{aligned}\hat{z} \times \frac{\partial \boldsymbol{v}_i}{\partial t} = \quad & -\frac{e}{m_i}\left(\hat{z} \times \nabla\phi + \frac{1}{c}\hat{z} \times \frac{\partial\tilde{A}_z}{\partial t}\right) - \frac{e}{m_i c}\hat{z} \times (\boldsymbol{v}_i \times \boldsymbol{B}_0) - \\ & \frac{e}{m_i c}\hat{z} \times (\boldsymbol{v}_i \times \delta\boldsymbol{B}) - \frac{k_B T_i}{n_0 m_i}(\hat{z} \times \nabla n)\end{aligned}$$

Taking again the cross product with $\hat{z}$ which will give us the perpendicular components with reverse of sign, and applying the time derivative on the equation will give,

$$\begin{aligned}-\frac{\partial^2 \boldsymbol{v}_{i\perp}}{\partial t^2} = \quad & \frac{e}{m_i}\frac{\partial}{\partial t}(\nabla_\perp\phi) - \omega_{ci}\left(\hat{z} \times \frac{\partial \boldsymbol{v}_{i\perp}}{\partial t}\right) - \frac{e v_{iz}}{m_i c}(\hat{z} \times \delta\boldsymbol{B}) + \\ & \frac{k_B T_i}{n_0 m_i}\frac{\partial}{\partial t}(\nabla_\perp n)\end{aligned}$$

We have used the ion cyclotron frequency $\omega_{ci} = \frac{eB}{m_i c}$,

for the second term of RHS, using equation (1) in the above equation-

$$\left(\frac{\partial^2}{\partial t^2}+\omega_{ci}^2\right)\boldsymbol{v}_{i\perp}=-\frac{c}{B}\omega_{ci}\frac{\partial}{\partial t}(\nabla_\perp\phi)-\frac{c}{B}\omega_{ci}^2(\hat{z}\times\nabla_\perp\phi)-$$
$$\omega_{ci}\frac{k_BT_i}{n_0m_i}(\hat{z}\times\nabla_\perp n)-\frac{k_BT_i}{n_0m_i}\frac{\partial}{\partial t}(\nabla_\perp n)+\frac{e\omega_{ci}}{m_ic}v_{iz}\delta\boldsymbol{B}+\frac{ev_{iz}}{m_ic}(\hat{z}\times\delta\boldsymbol{B}) \quad (2)$$

Taking a dot product with $\nabla_\perp$ in equation (2) and taking ion continuity equation into consideration :

$$\left(\frac{\partial^2}{\partial t^2}+\omega_{ci}^2\right)\frac{n_i}{n_0}=-\frac{c}{B}\omega_{ci}(\nabla_\perp^2\phi)-\frac{k_BT_i}{n_0m_i}(\nabla_\perp^2 n)-\frac{2e\delta B}{m_ic}\omega_{ci}(\nabla_\perp v_z) \quad (3)$$

So one can get the following from modifying further the equation (3):

$$\left(\frac{\partial^2}{\partial t^2}+\omega_{ci}^2+\omega_{ci}\frac{2e\delta B}{m_ic}\right)\left(\frac{\partial^2}{\partial t^2}-c_s^2\frac{\partial^2}{\partial z^2}\right)n+\frac{en_0}{m_i}\frac{\partial^2}{\partial t^2}(\nabla_\perp^2\phi)+$$
$$\frac{k_BT_i}{m_i}(\nabla_\perp^2 n)=0 \quad (4)$$

Now for electrons, the expression for the equation of motion can be written as-

$$\frac{\partial\boldsymbol{v}_e}{\partial t}=-\frac{e}{m_e}\boldsymbol{E}-\frac{e}{m_ec}(\boldsymbol{v}_e\times\boldsymbol{B})-(\boldsymbol{v}_e\times\delta\boldsymbol{B})-\frac{k_BT_e}{n_0m_e}\nabla n$$

After taking the z component of electrons, the equation of motion can be written as

$$\frac{\partial\boldsymbol{v}_{ez}}{\partial t}=-\frac{e}{m_e}\boldsymbol{E}_z-\frac{e}{m_ec}(\boldsymbol{v}_e\times\boldsymbol{B})_z-(\boldsymbol{v}_e\times\delta\boldsymbol{B})-\frac{k_BT_e}{n_0m_e}\frac{\partial n}{\partial z} \quad (5)$$

Since $\boldsymbol{B}$ is in the z-direction so cross product term will vanish and the equation (5) in terms of potential fields is-

$$-\frac{m_e}{e}\frac{\partial(-n_0ev_{ez})}{\partial t}=n_0e\left(\frac{\partial\phi}{\partial z}+\frac{1}{c}\frac{\partial\tilde{A}_z}{\partial t}\right)-k_BT_e\frac{\partial n}{\partial z}$$

Using the value of current density $j_z=-n_0\,ev_{ez}$ in the above equation. one can get equation in the z direction as-

$$\frac{\partial\phi}{\partial z}+\frac{1}{c}\frac{\partial\tilde{A}_z}{\partial t}=-\frac{m_e}{n_0e^2}\frac{\partial(j_z)}{\partial t}+\frac{k_BT_e}{n_0e}\frac{\partial n}{\partial z} \quad (6)$$

Again taking the derivative of (6) in the z-direction.

$$\frac{\partial^2\phi}{\partial z^2}+\frac{1}{c}\frac{\partial^2\tilde{A}_z}{\partial z\partial t}=-\frac{m_e}{n_0e^2}\frac{\partial\nabla_z(j_z)}{\partial t}+\frac{k_BT_e}{n_0e}\frac{\partial^2 n}{\partial z^2} \quad (7)$$

From the electron continuity equation and $\frac{\partial n}{\partial t} + \nabla.(n\boldsymbol{v}_e) = 0$ using Ampere's law, $j_z = -\frac{c}{4\pi}\nabla_\perp^2 A_z$ author will get:

$$\frac{\partial n}{\partial t} + \frac{c}{4\pi e}\nabla_\parallel(\nabla_\perp^2.\tilde{A}_\parallel) = 0$$
$$\frac{\partial^2 n}{\partial t^2} + \frac{c}{4\pi e}\nabla_\parallel\frac{\partial(\nabla_\perp^2.\tilde{A}_\parallel)}{\partial t} = 0 \tag{8}$$

Using equations (7),(8) and (4), the dynamical equation of the KAW in terms of vector potential is written as

$$(1-\lambda_e^2\nabla_\perp^2)\frac{\partial^2\tilde{A}_z}{\partial t^2} = v_A^2\left[1-\rho_s^2\nabla_\perp^2 - 2\frac{\delta B}{B_0}\right]\frac{\partial^2\tilde{A}_z}{\partial z^2} \tag{9}$$

Here $\rho_s$ is the radius of the ion gyro defined as $\rho_s \approx \frac{c_s}{\omega_{ci}}$, $\omega_{ci} = \frac{eB_0}{m_i c}$ is the ion gyro frequency, the ion sound speed is given as $c_s = \left(\frac{k_B(T_e+T_i)}{m_i}\right)^{\frac{1}{2}}$, and $\lambda_e = \left(\frac{m_e c^2}{4\pi n_0 e^2}\right)^{\frac{1}{2}}$ is the collision-less electron skin depth (inertial length). In the above eq. $v_A = (B_0^2/4\pi n_0 e^2)^{1/2}$ is the Alfvén speed. Due to density modification, the Alfvén speed takes the following form:

$$v_A^2 = \frac{B_0^2}{4\pi(n_0+\delta n)e^2} = \frac{B_0^2}{4\pi n_0 e^2(1+\frac{\delta n}{n_0})} \approx v_A^2\left(1-\frac{\delta n}{n_0}\right).$$

For the present field $\boldsymbol{B} = B_{0x}\hat{x} + B_{0y}\hat{y} + B_{0z}\hat{z}$ where $B_{0x} = \frac{B_0}{L}x$, $B_{0y} = -\frac{B_0}{L}y$, $B_{0z} = B_0$. we use the modification.

$$\frac{\partial}{\partial z} = \frac{(B_0\hat{z})}{B_0}.\nabla = \frac{\boldsymbol{B}.\nabla}{B_0} = \frac{1}{B_0}\left(B_{0x}\hat{x} + B_{0y}\hat{y} + B_{0z}\hat{z}\right).\nabla$$
$$= \frac{B_{0x}}{B_0}\frac{\partial}{\partial x} + \frac{B_{0y}}{B_0}\frac{\partial}{\partial y} + \frac{\partial}{\partial z}$$

Furthermore, using the above modification in equation (9)

$$\left(\frac{\partial}{\partial z}\right)\frac{\partial}{\partial z} = \left(\frac{B_{0x}}{B_0}\frac{\partial}{\partial x} + \frac{B_{0y}}{B_0}\frac{\partial}{\partial y} + \frac{\partial}{\partial z}\right)\left(\frac{B_{0x}}{B_0}\frac{\partial}{\partial x} + \frac{B_{0y}}{B_0}\frac{\partial}{\partial y} + \frac{\partial}{\partial z}\right)$$
$$\left(\frac{\partial}{\partial z}\right)\frac{\partial}{\partial z} = \left(\frac{B_{0x}}{B_0}\right)^2\frac{\partial^2}{\partial x^2} + 2\left(\frac{B_{0x}}{B_0}\frac{B_{0y}}{B_0}\right)\frac{\partial^2}{\partial x\partial y} + 2\left(\frac{B_{0x}}{B_0}\right)\frac{\partial^2}{\partial x\partial z}$$
$$+\left(\frac{B_{0y}}{B_0}\right)^2\frac{\partial^2}{\partial y} + 2\left(\frac{B_{0y}}{B_0}\right)\frac{\partial^2}{\partial z\partial y} + \frac{\partial^2}{\partial z^2}$$

Rewrite equation (9) using the above modification and neglecting the higher-order terms:

$$(1-\lambda_e^2\nabla_\perp^2)\frac{\partial^2\tilde{A}_z}{\partial t^2} - v_A^2\left(1-\rho_s^2\nabla_\perp^2 - \frac{\delta n}{n_0}\right)\left[\left(\frac{B_{0x}}{B_0}\right)^2\frac{\partial^2}{\partial x^2} +\right.$$

$$2\left(\frac{B_{0x}}{B_0}\frac{B_{0y}}{B_0}\right)\frac{\partial^2}{\partial x\partial y}+2\left(\frac{B_{0x}}{B_0}\right)\frac{\partial^2}{\partial x\partial z}+\left(\frac{B_{0y}}{B_0}\right)^2\frac{\partial^2}{\partial y^2}+$$
$$2\left(\frac{B_{0y}}{B_0}\right)\frac{\partial^2}{\partial z\partial y}+\frac{\partial^2}{\partial z^2}\Bigg]\tilde{A}_z=0 \tag{10}$$

Next, we assume the envelope solution for vector potential in the above equation (10) in the following form $\tilde{A}_z = A_z(x,y,z,t)e^{\iota(k_{0x}x+k_{0y}y+k_{0z}z-\omega_0 t)}$. If the effect of field modification due to the null point is not present, then equation (10) with envelope solution gives the usual dispersion relation of KAW as:

$$\omega_0^2=\frac{k_{0z}^2v_A^2(1+\rho_s^2k_{0\perp}^2))}{(1+\lambda_e^2k_{0\perp}^2)}$$

Now, substituting the envelope solution into our fully fledged nonlinear equation (10), we get the following expression.

$$-2\iota\omega_0(1+\lambda_e^2k_{0\perp}^2)\frac{\partial A_z}{\partial t}+2\iota k_{0x}(\lambda_e^2\omega_0^2-v_A^2\rho_s^2k_{0z}^2)\frac{\partial A_z}{\partial x}+2\iota k_{0x}(\lambda_e^2\omega_0^2-$$
$$v_A^2\rho_s^2k_{0z}^2)\frac{\partial A_z}{\partial y}-2\iota v_A^2k_{0z}(1+\rho_s^2k_{0\perp}^2)\frac{\partial A_z}{\partial z}+(\lambda_e^2\omega_0^2-v_A^2\rho_s^2k_{0z}^2)\frac{\partial^2A_z}{\partial x^2}+(\lambda_e^2\omega_0^2$$
$$-v_A^2\rho_s^2k_{0z}^2)\frac{\partial^2A_z}{\partial y^2}-4\omega_0k_{0\perp}\lambda_e^2\frac{\partial A_z}{\partial t\partial z}-4v_A^2\rho_s^2k_{0z}^2\frac{\partial A_z}{\partial z\partial x}-4v_A^2\rho_s^2k_{0z}^2\frac{\partial A_z}{\partial z\partial y}+$$
$$\iota2\omega_0\lambda_e^2\frac{\partial^2A_z}{\partial t\partial^2x}+\iota2\omega_0\lambda_e^2\frac{\partial^2A_z}{\partial t\partial^2y}+\iota2v_A^2\rho_s^2k_{0z}\frac{\partial^3A_z}{\partial^2x\partial z}+\iota2v_A^2\rho_s^2k_{0z}\frac{\partial^3A_z}{\partial^2y\partial z}+$$
$$\iota2v_A^2\rho_s^2k_{0\perp}\frac{\partial^3A_z}{\partial x\partial^2z}+\iota2v_A^2\rho_s^2k_{0\perp}\frac{\partial^3A_z}{\partial y\partial^2z}+v_A^2\rho_s^2\frac{\partial^4A_z}{\partial^2x\partial^2z}+v_A^2\rho_s^2\frac{\partial^4A_z}{\partial^2y\partial^2z}-$$
$$v_A^2\rho_s^2k_{0\perp}^2k_{0z}^2(1+\rho_s^2k_{0\perp}^2-\frac{\delta n}{n_0})\Bigg[\left(\frac{B_{0x}}{B_0}\right)^2k_{0x}^2+\left(\frac{B_{0y}}{B_0}\right)^2k_{0y}^2+$$
$$2\left(\frac{B_{0x}}{B_0}\right)k_{0x}k_{0z}+2\left(\frac{B_{0y}}{B_0}\right)k_{0y}k_{0z}+\left(\frac{B_{0x}}{B_0}\right)\left(\frac{B_{0y}}{B_0}\right)k_{0y}k_{0x}\Bigg]A_z=0 \tag{11}$$

The following normalising parameters are used to make equation (11) in dimensionless form as

$$x_n=y_n=\frac{2(\lambda_e^2\omega_0^2-v_A^2\rho_s^2k_{0z}^2)}{v_A^2\rho_s^2k_{0z}^2k_{0\perp}},\quad tn=\frac{2\omega_0(1+\lambda_e^2k_{0\perp}^2)}{v_A^2\rho_s^2k_{0z}^2k_{0\perp}^2},\quad z_n=\frac{2(1+\rho_s^2k_{0\perp}^2)}{\rho_s^2k_{0\perp}^2k_{0z}}$$

The fluctuation in density is given using the fluid model Yadav and Sharma (2014) $\frac{\delta n}{n_0}=|A_z|^2$

Therefore we obtain the dimensionless nonlinear dynamical equation as follows:

$$-\iota\frac{\partial A_z}{\partial t}+\iota\frac{\partial A_z}{\partial x}+\iota\frac{\partial A_z}{\partial y}-\iota\frac{\partial A_z}{\partial z}+C_1\frac{\partial^2A_z}{\partial x^2}+C_1\frac{\partial^2A_z}{\partial y^2}-C_2\frac{\partial^2A_z}{\partial z^2}-C_3\frac{\partial^2A_z}{\partial t\partial z}$$
$$-C_4\frac{\partial^2A_z}{\partial x\partial z}-C_4\frac{\partial^2A_z}{\partial y\partial z}+\iota C_5\frac{\partial^3A_z}{\partial t\partial^2x}+\iota C_5\frac{\partial^3A_z}{\partial t\partial^2y}+\iota C_6\frac{\partial^3A_z}{\partial^2x\partial z}+\iota C_6\frac{\partial^3A_z}{\partial^2y\partial z}+$$

$$\iota C_7\frac{\partial^3 A_z}{\partial x\partial^2 z}+\iota C_7\frac{\partial^3 A_z}{\partial y\partial^2 z}+C_8\frac{\partial^4 A_z}{\partial^2 z\partial^2 x}+C_8\frac{\partial^4 A_z}{\partial^2 z\partial^2 y}-|A_z|^2-\frac{(1+\rho_s^2k_\perp^2)}{k_{0z}^2\rho_s^2k_\perp^2}$$
$$\left[\left(\frac{B_{0x}}{B_0}\right)^2k_{0x}^2+\left(\frac{B_{0y}}{B_0}\right)^2k_{0y}^2+2\left(\frac{B_{0x}}{B_0}\right)k_{0x}k_{0z}+2\left(\frac{B_{0y}}{B_0}\right)k_{0y}k_{0z}+\right.$$
$$\left.\left(\frac{B_{0x}}{B_0}\right)\left(\frac{B_{0y}}{B_0}\right)k_{0y}k_{0x}\right]A_z=0 \tag{12}$$

Where

$$C_1 = \frac{\lambda_e^2\omega_0^2-v_A^2\rho_s^2k_{0z}^2}{(v_A^2\rho_s^2k_{0z}^2k_{0\perp}^2x_n^2)},\ C_2=\frac{(1+\rho_s^2k_{0\perp}^2)}{\rho_s^2k_{0z}^2k_{0\perp}^2z_n^2},\ C_3=\frac{4\omega_0\lambda_e^2}{v_A^2\rho_s^2k_{0z}^2k_\perp x_nt_n},$$
$$C_4 = \frac{4}{k_{0z}k_{0\perp}x_nz_n},C_5=\frac{2\omega_0\lambda_e^2}{v_A^2\rho_s^2k_{0z}^2k_{0\perp}^2x_n^2t_n},C_6=\frac{2}{k_{0z}k_{0\perp}^2x_n^2z_n},$$
$$C_7 = \frac{2}{k_{0z}^2k_{0\perp}x_nz_n^2},\ C_8=\frac{1}{k_{0z}^2k_{0\perp}^2x_n^2z_n^2}$$

Equation (12) is used to model the evolution of nonlinear (coherent) structure and turbulence generation by the interaction of nonlinear KAW in the neighbourhood of the null point. We have taken a simple profile of the magnetic field near the null point. Substituting the profile of magnetic null point $\boldsymbol{B}=\frac{B_0}{L}(x,-y,0)$ in equation (12), we have the final equation as

$$-\iota\frac{\partial A_z}{\partial t}+\iota\frac{\partial A_z}{\partial x}+\iota\frac{\partial A_z}{\partial y}-\iota\frac{\partial A_z}{\partial z}+C_1\frac{\partial^2 A_z}{\partial x^2}+C_1\frac{\partial^2 A_z}{\partial y^2}-C_2\frac{\partial^2 A_z}{\partial z^2}-C_3\frac{\partial^2 A_z}{\partial t\partial z}$$
$$-C_4\frac{\partial^2 A_z}{\partial x\partial z}-C_4\frac{\partial^2 A_z}{\partial y\partial z}+\iota C_5\frac{\partial^3 A_z}{\partial t\partial^2 x}+\iota C_5\frac{\partial^3 A_z}{\partial t\partial^2 y}+\iota C_6\frac{\partial^3 A_z}{\partial x\partial^2 z}+\iota C_6\frac{\partial^3 A_z}{\partial y\partial^2 z}+$$
$$C_8\frac{\partial^4 A_z}{\partial^2 z\partial^2 x}+C_8\frac{\partial^4 A_z}{\partial^2 z\partial^2 y}-|A_z|^2-\frac{(1+\rho_s^2k_{0\perp}^2)}{k_{0z}^2\rho_s^2k_{0\perp}^2}\left[\left(\frac{x}{L}\right)^2k_{0x}^2+\left(\frac{-y}{L}\right)^2k_{0y}^2-\right.$$
$$\left.\left(\frac{x}{L}\right)\left(\frac{y}{L}\right)k_{0y}k_{0x}\right]A_z=0 \tag{13}$$

## 3. Numerical Simulation and Result

Here we solve the equation (12) in a periodic spatial domain $\left(\frac{2\pi}{\alpha_x}\right)\times\left(\frac{2\pi}{\alpha_y}\right)\times\left(\frac{2\pi}{\alpha_z}\right)$ with grid points $(128\times128\times128)$. we have applied the pseudo-spectral method for space integration and finite difference method with the predictor-corrector scheme to study the temporal evolution. In the simulation, it is taken as $\alpha_x=\alpha_y=\alpha_z=0.2$. These are the wave numbers of the perturbation. Our present study considers the initial condition (Biskamp and Welter, 1989) for the simulation as

$$A_z(x,y)=\cos(2x+2.3)+\cos(y+4.1) \tag{14}$$

We have used the typical solar corona plasma parameters as :

$B_0 \approx 32\ G$, $n_0 \approx 10^8\ cm^{-3}$, $T_i \approx 235\ eV$ $T_e \approx 37\ eV$ (Champeaux, Passot, and Sulem (1997)).

We have further calculated the followings parameters:

$v_A \approx 6.98 \times 10^8\ \frac{cm}{sec}$, $\omega_{ci} \approx 3 \times 10^5\ \frac{rad}{sec}$, $\omega_{pi} \approx 1.32 \times 10^7\ \frac{rad}{sec}$, $c_s \approx 7 \times 10^6\ \frac{cm}{sec}$, $\rho_s \approx 22.96\ cm$, $\omega_{ce} = 5.6 \times 10^8\ \frac{rad}{sec}$, $\omega_0 = 1.5 \times 10^5\ \frac{rad}{sec}$, $k_{0z} = 3.9 \times 10^{-4}\ cm^{-1}$, $k_{0\perp} = 9.5 \times 10^{-3}\ cm^{-1}$ and $k_{0\perp}\lambda_e = 0.2$.

For these parameters, one can find $x_n = y_n = 66cm \approx \lambda_e$ , $z_n = 1.6 \times 10^4 cm \approx \frac{1}{k_{0z}}$ and $t_n = 39\ \mu sec \approx \frac{1}{\omega_0}$.

Before proceeding further to solve the present equation, we cross-checked the accuracy of the code. Accuracy has been confirmed by setting up the modified nonlinear Schrodinger equation (MNLS) and using the conservation of the plasmon number, i.e. $\Sigma_k |A_z|^2$. It remained constant up to an accuracy of $10^{-5}$ during the computation. After validating the reliability of the code, we updated it to solve the equation of the dynamical model. Now, the model equation is solved using numerical techniques.

At different times, the spatial evolution of the vector potential $A_z$ of KAW in the x-y plane is shown in Figure 2 at fixed z. As time evolves the profile is modified. Initially, the coherent structures are formed and with time evolution we can see the evolving chaotic structures and with further increasing time from $t = 30$ to $t = 60$ these structures become more random and irregular showing the turbulence generation and evolution in the reconnection site. The evolution of these structures confirms an indication of turbulence.

In Figure 3, the contour plot of the magnetic flux and the current density are shown. The spatial evolution of current density $j = -\nabla^2 A_z$ has been studied with the time evolution. As time evolves, we see that the proper X-O structures get distorted, and we get a fully chaotic pattern with increasing amplitude as the temporal evolution occurs. The presence of these chaotic structures indicates the generation and evolution of the turbulence.

Furthermore, in Figure 4, the spectrum plot of the normalised field intensity associated with KAW in the solar corona is plotted. The dark red and green lines are not fit but are used for reference purposes. We find the Kolmogorov $-\frac{5}{3}$ features in the inertial range and after the breakpoint, it shows the steepening of the spectra plot going up to $-3.3$ (in the range of $-2$ to $-4$). The observation from the Parker Solar Probe (PSP) mission is giving a vital contribution to our research area whereas till now only observations are being reported in the range of 0.17 AU to 0.25AU. A study of the data obtained by PSP shows that the power spectra are steepening in the range of $-3.2$ to $-5.8$,as reported by Huang *et al.* (2022). As this PSP continues to follow its path, it will get closer to the solar corona and continue to provide us with more data. We can further proceed to generalise our model to different parameters which can be tested and will improve our understanding according to the further observations.

The interactions between the localised structures and the coronal particles are random and frequent in nature. The stochastic interaction, also known as Fermi acceleration of the second kind, has been associated with particle acceleration and heating phenomena (Vlahos and Isliker, 2018). To further elucidate the proper mechanism of the thermal tail that improves the velocity distribution function as a result of turbulence, Sharma and Kumar (2010) presents the

Fokker–Planck diffusive formalism. This formalism describes the evolution of the velocity distribution function due to the interaction of KAW and the coronal particles. The recurrent interaction between plasma particles and the localized intense fields can be studied by using the quasi-ilinear diffusion equation, which is given as

$$\frac{\partial f}{\partial t} = \frac{\partial}{\partial v}\left(D(v)\frac{\partial f}{\partial v}\right) \tag{15}$$

Here, diffusion coefficient is $D(v)$ and $f(t, v)$ is the distribution function of velocity.

The distribution function can be treated as time independent if the observation time $t_{obs}$ is greater than the characteristic time taken by ponder-motive nonlinearity in settling up. In this case the distribution function is proposed as $f(v) \propto v^{2-\eta}$. The spectral index $[\eta]$ greatly affects the behaviour of the velocity distribution function. Assuming a quasi-steady state, the distribution function for the localised structures, as per Sharma and Kumar (2010) takes the form $f(v) \propto v^{-1.3}$ where $\eta = 3.3$ from the simulation result. This enhances the thermal tail segment of the distribution function and pave the way for thermal heating of the particles.

For intense phase of localized field structures, the quasi-linear theory is inadequate and the fractional diffusion approach should be taken into consideration. In this approach, the fractional diffusion equation describes the interaction of coronal particles with the given intense localized fields (Bian and Browning, 2008). The distribution function $[g(v, t)]$, in this scenario is given as $g(v) \sim v^{-(1+\mu)}$. The simulation's output provides the spectral index value as $(\mu = 3.3)$. Consequently, the distribution function is provided by $g(v) \sim v^{-(4.4)}$. As a result, the thermal tail of the energetic electrons becomes strengthened, which causes the corona to heat.

So, we anticipate the strengthening of the thermal tail of the energetic electrons and subsequent coronal heating and particle acceleration due to this interaction between the localised structures and coronal particles.

## 4. semi-analytical Model

To understand the physical mechanism of the coherent structures and current sheet generation, we have solved the model equation by the semi-analytical method. Since the numerical method solution covers a wide domain of space and time. Therefore, we have developed a steady-state model for simplification. In this situation, from equation (11) we obtain the following form,

$$\left(\omega_0^2\lambda_e^2 - k_{0z}^2 v_A^2\rho_s^2\right)\frac{\partial^2 A_z}{\partial x^2} + \left(\omega_0^2\lambda_e^2 - k_{0z}^2 v_A^2\rho_s^2\right)\frac{\partial^2 A_z}{\partial y^2} - 2\iota k_{0z} v_A^2\left(1 + \rho_s^2 k_{0\perp}^2\right)$$
$$\frac{\partial A_z}{\partial z} - v_A^2\rho_s^2 k_{0\perp}^2 k_{0z}^2\frac{\delta n}{n_0}A_z - v_A^2(1 + k_{0\perp}^2\rho_s^2)\left[\left(\frac{B_{0x}}{B}\right)^2 k_{0x}^2 + \left(\frac{B_{0y}}{B}\right)^2 k_{0y}^2 +\right.$$

$$2\left(\frac{B_{0x}}{B}\right)\left(\frac{B_{0y}}{B}\right)k_{0x}k_{0y}\Bigg]A_z = 0 \tag{16}$$

Here we have assumed the variation of the vector potential $A_z$ in terms of eikonal (Akhmanov, Sukhorukov, and Khokhlov, 1968) as

$$A_z = A_{00}(x,y,z)e^{\iota k_{0z}S(x,y,z)}$$

Substituting this assumed solution in equation (16) and comparing real and imaginary parts, we get

$$\begin{aligned}
&\Upsilon\left(\frac{\partial^2 A_{00}}{\partial x^2}+\frac{\partial^2 A_{00}}{\partial y^2}\right)-\Upsilon k_{0z}^2 A_{00}\left(\frac{\partial^2 S}{\partial x^2}+\frac{\partial^2 S}{\partial y^2}\right)+2\left(\frac{1+k_{0\perp}^2\rho_s^2}{k_{0\perp}^2\rho_s^2}\right)\\
&A_{00}\frac{\partial S}{\partial z}-\frac{\delta n}{n_0}A_{00}-\left(\frac{1+k_{0\perp}^2\rho_s^2}{k_{0\perp}^2\rho_s^2}\right)\Bigg[\left(\frac{B_{0x}}{B_0}\right)^2 k_{0x}^2+\left(\frac{B_{0y}}{B_0}\right)^2 k_{0y}^2+\\
&2\left(\frac{B_{0x}}{B_0}\right)\left(\frac{B_{0y}}{B_0}\right)k_{0x}k_{0y}\Bigg]A_{00}=0
\end{aligned} \tag{17}$$

$$2\Upsilon\frac{\partial S}{\partial x}\frac{\partial A_{00}}{\partial x}+2\Upsilon\frac{\partial S}{\partial y}\frac{\partial A_{00}}{\partial y}-2\left(\frac{1+k_{0\perp}^2\rho_s^2}{k_{0\perp}^2\rho_s^2}\right)\frac{\partial A_{00}}{\partial z}=0 \tag{18}$$

Here,

$$\Upsilon=\frac{\left(\omega_0^2\lambda_e^2-k_{0z}^2v_A^2\rho_s^2\right)}{k_{0\perp}^2\rho_s^2k_{0z}^2v_A^2}$$

Further, we have assumed the solution of equations (17) and (18) in the form of a Gaussian profile as follows

$$\begin{aligned}
A_{00}&=\frac{A_{00}'^2}{f_1(z)f_2(z)}e^{\left(-\frac{x^2}{r_{01}^2f_1}-\frac{y^2}{r_{02}^2f_2}\right)}\\
S&=\beta_1(z)\frac{x^2}{2}+\beta_2(z)\frac{y^2}{2}
\end{aligned}$$

Substituting above assumed solution in equation (18), we obtained

$$\beta_1=\frac{v_A^2(1+k_{0\perp}^2\rho_s^2)}{(\omega_0^2\lambda_e^2-k_{0z}^2v_A^2\rho_s^2)}\frac{1}{f_1}\frac{\partial f_1}{\partial z} \tag{19}$$

$$\beta_2=\frac{v_A^2(1+k_{0\perp}^2\rho_s^2)}{(\omega_0^2\lambda_e^2-k_{0z}^2v_A^2\rho_s^2)}\frac{1}{f_2}\frac{\partial f_2}{\partial z} \tag{20}$$

Here, $f_1, f_2$ are the beam width parameters of the wave,$r_{01}$ and $r_{02}$ are the scale size, and $\beta_1$ and $\beta_2$ are slowly varying functions of z. Using equation (19)–(20) in equation (17) along with the paraxial approximation i.e.,$(x \ll r_{01}f_1, y \ll r_{02}f_2)$

and equating the coefficient of $x^2$ and $y^2$ on both sides, we obtain the dimensionless differential equations for beam width parameter as

$$\frac{\partial^2 f_1}{\partial Z^2} = R_d^2 \frac{(\omega_0^2\lambda_e^2 - k_{0z}^2 v_A^2 \rho_s^2)}{(1+\rho_s^2 k_{0\perp}^2)} \left( \frac{4(\omega_0^2\lambda_e^2 - v_A^2\rho_s^2 k_{0z}^2)}{v_A^4 k_{0z}^2 (1+\rho_s^2 k_{0\perp}^2)} \frac{1}{r_{01}^4 f_1^3} - \frac{k_{0x}^2}{v_A^2 k_{0z}^2 L^2} f_1 - \frac{\rho_s^2 k_{0\perp}^2}{(1+\rho_s^2 k_{0\perp}^2)} \frac{(\alpha A_{00}'^2)}{r_{01}^2 f_1^3 f_2^2} \right) \tag{21}$$

$$\frac{\partial^2 f_2}{\partial Z^2} = R_d^2 \frac{(\omega_0^2\lambda_e^2 - k_{0z}^2 v_A^2 \rho_s^2)}{(1+\rho_s^2 k_{0\perp}^2)} \left( \frac{4(\omega_0^2\lambda_e^2 - v_A^2\rho_s^2 k_{0z}^2)}{v_A^4 k_{0z}^2 (1+\rho_s^2 k_{0\perp}^2)} \frac{1}{r_{02}^4 f_2^3} - \frac{k_{0y}^2}{v_A^2 k_{0z}^2 L^2} f_2 - \frac{\rho_s^2 k_{0\perp}^2}{(1+\rho_s^2 k_{0\perp}^2)} \frac{(\alpha A_{00}'^2)}{r_{02}^2 f_2^3 f_1^2} \right) \tag{22}$$

Here $Z = \frac{z}{R_d}$ and $R_d = k_{0z} r_{01}^2$.

In the above equations (21) and (22) the first term is due to the finite transverse size of the KAW structures and leads to the spreading effect, the second term is due to the null point, which is responsible for convergence of the structures, and the last term is due to the density modification in the background, which is responsible for convergence of the structures (both in x and y). The critical value of the scale size can be found by solving equations (21) and (22) for a given power value. For a initially plane wavefront, we assume the following initial conditions $f_1 = f_2 = 1$ at $z = 0$ and $\frac{\partial f_1}{\partial z} = \frac{\partial f_1}{\partial z} = 0$ at $z = 0$ .

$$1 - \frac{v_A^2(1+\rho_s^2 k_{0\perp}^2) k_{0x}^2 r_{01}^4}{4(\omega_0^2\lambda_e^2 - v_A^2\rho_s^2 k_{0z}^2) L^2} - \frac{\rho^2 k_{0\perp}^2 v_A^2 k_{0z}^2 (\alpha A_{00}'^2) r_{01}^2}{4(\omega_0^2\lambda_e^2 - v_A^2\rho_s^2 k_{0z}^2)} = 0 \tag{23}$$

$$1 - \frac{v_A^2(1+\rho_s^2 k_{0\perp}^2) k_{0x}^2 r_{02}^4}{4(\omega_0^2\lambda_e^2 - v_A^2\rho_s^2 k_{0z}^2) L^2} - \frac{\rho^2 k_{0\perp}^2 v_A^2 k_{0z}^2 (\alpha A_{00}'^2) r_{02}^2}{4(\omega_0^2\lambda_e^2 - v_A^2\rho_s^2 k_{0z}^2)} = 0 \tag{24}$$

As it is clear from equations (23) and (24) that in the absence of KAW, i.e. when the power of KAW ($\alpha A_{00}'^2$) is zero, the scale size of the localized KAW structures is quadratic in $r_{01}^2$ and $r_{02}^2$, respectively, with a valid root of the equation as

$$r_{01}^2 = \left( \frac{4(\omega_0^2\lambda_e^2 - v_A^2\rho_s^2 k_{0z}^2) L^2}{(1+\rho_s^2 k_{0\perp}^2) k_{0x}^2 v_A^2} \right)^{\frac{1}{2}}, \qquad r_{02}^2 = \left( \frac{4(\omega_0^2\lambda_e^2 - v_A^2\rho_s^2 k_{0z}^2) L^2}{(1+\rho_s^2 k_{0\perp}^2) k_{0y}^2 v_A^2} \right)^{\frac{1}{2}} \tag{25}$$

Magnetic fluxes are related to current density by Ampere's law. So, we have studied the evolution of magnetic flux and current density. To calculate the size of the scale of the current sheets along the x and y directions, if the relation can be used as $j = -\nabla^2 A_{00}$, where $A_{00} = \frac{A_{00}'^2}{f_1 f_2} exp\left(-\frac{x^2}{r_{01}^2 f_1} - \frac{y^2}{r_{02}^2 f_2}\right)$. The expression further simplified to give the scale size of current sheet along x direction as the order of $\frac{\rho_s}{\sqrt{3}}$ Similar expression can be found for y component as $\frac{\rho_s}{\sqrt{3}}$. Further, if

we take into account the power of KAW ($\alpha A'^2_{00}$), then for any finite power the solution of (21) and (22) with assumed initial value of $f_1$ and $f_2$ will give some different value of current density. So, the KAW with different finite power will give a varying scale size. Consequently, it demonstrates that the scale size of coherent structures and current sheet dimensions depends on the KAW power as well.

## 5. Conclusion

The present work demonstrates the generation of localised structures when KAW is powerful enough and propagates in an environment where null points are present (as considered here). The model equation to represent a 3-D Kinetic Alfvén Waves interacting with the null point using the nonlinear density perturbation has been developed and numerically solved. The formation of the localized structure of the magnetic field provides a pathway for energy transfer from a larger scale to a smaller spatial scale. Small-scale localized structure formation shows the coupling of Alfvén waves with null points. This localization of the wave and the resulting turbulence leads to energy transfer between large and small scales. The generated turbulence spectrum shows that for the energy spectrum of the power in the inertial region for $k_{0\perp}\rho_s < 1$, we have Kolmogorov scaling nearly $k^{-\frac{5}{3}}$, but after the inertial zone for $k_{0\perp}\rho_s > 1$ the slope started to change and gets steeper, having power spectra in the range $k^{-2}$ to $k^{-4}$. The steepening assures us that the energy transfer rate from one scale to the other smaller scales got increased in the temporally evolved turbulence. Therefore, the signature of turbulence is apparent in the present power spectrum at the reconnection site. This emergence of turbulence ensures the enhancement of the energy-transfer rate. We have also presented the energetic particle distribution function by using this turbulence with the help of the Fokker-Planck equation and velocity space diffusion coefficient. The semi-analytical study has also been done which gives us an estimate of the scale size of coherent structures formed. The order of the current sheet size turns out to be of the order of the ion gyro radius but it gets affected by the power of KAW significantly. Therefore, we find that the localized magnetic structure and resulting turbulence are the outcomes of this KAW interaction,near the null point and responsible for coronal heating.

**Acknowledgement** This research is supported by the Council for Scientific & Industrial Research (CSIR), New Delhi, India.

## 6. Figures

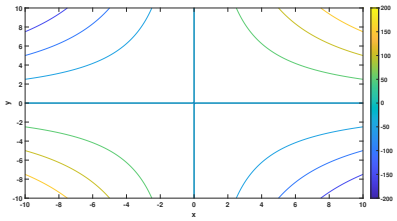


**Figure 1.** *Schematic diagram of equilibrium vector potential $A_z$ structure near Null point.*

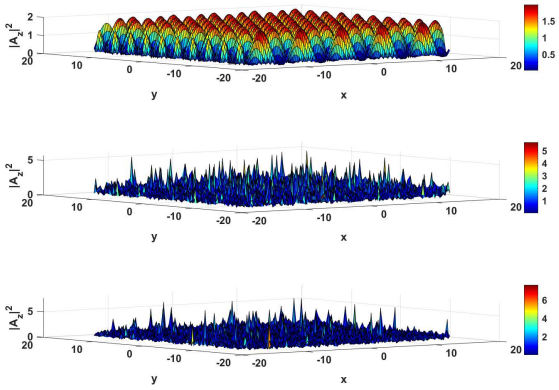


**Figure 2.** *Normalised vector potential showing the coherent structures and localization of the KAW at (a) $t = 0$, (b) $t = 30$ and (c) $t = 60$.*

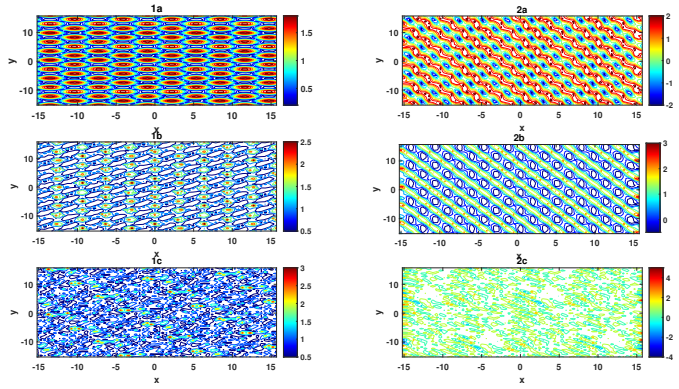



**Figure 3.** *Evolution of (1) vector potential $A_z$ and (2) current sheet in x-y plane at different time (a) $t = 0$, (b) $t = 4$ and (c) $t = 10$.*

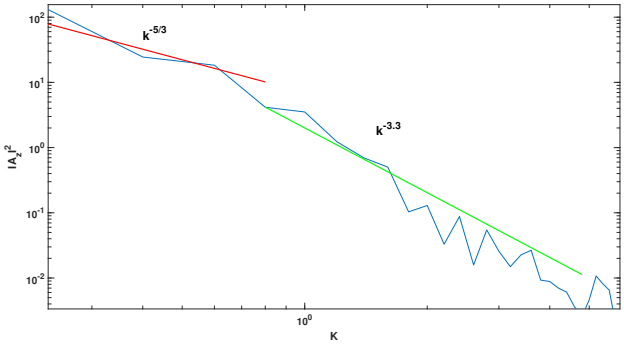



**Figure 4.** *The magnetic power spectra associated with Kinetic Alfvén Wave.*

## References


Akhmanov, S.A., Sukhorukov, A.P., Khokhlov, R.V.: 1968, Self-focusing and diffraction of light in a nonlinear medium. *Soviet physics USPEKHI* **10**(5), 609. [akhmanov1968self]

Bian, N., Browning, P.: 2008, Particle acceleration in a model of a turbulent reconnecting plasma: a fractional diffusion approach. *The Astrophysical Journal* **687**(2), L111. [bian2008particle]

Biermann, L.: 1946, The meaning of chromospheric turbulence and the uv excess of the sun. *Naturwischenschaften* **33**, 118. [biermann1946]

Biskamp, D., Welter, H.: 1989, Dynamics of decaying two-dimensional magnetohydrodynamic turbulence. *Physics of Fluids B* **1**(10), 1964. DOI. [Biskamp1989]

Brown, D.S., Priest, E.R.: 2001, The topological behaviour of 3D null points in the Sun's corona. *Astronomy and Astrophysics* **367**(1), 339. DOI. [Brown2001]

Carbone, V., Veltri, P., Mangeney, A.: 1990, Coherent structure formation and magnetic field line reconnection in magnetohydrodynamic turbulence. *Physics of Fluids A: Fluid Dynamics* **2**(8), 1487. [carbone1990]

Champeaux, S., Passot, T., Sulem, P.L.: 1997, Alfven wave filamentation. *Journal of Plasma Physics* **58**(4), 665. DOI. [CHAMPEAUX1997]

Close, R.M., Parnell, C.E., Priest, E.R.: 2004, Separators in 3D Quiet-Sun Magnetic Fields. *Solar Physics* **225**(1), 21. DOI. http://link.springer.com/10.1007/s11207-004-3259-0. [Close2004]

Dahlburg, R., Einaudi, G., Rappazzo, A., Velli, M.: 2012, Turbulent coronal heating mechanisms: coupling of dynamics and thermodynamics. *Astronomy and Astrophysics* **544**, L20. [dahlburg2012]

Dahlburg, R., Einaudi, G., Taylor, B., Ugarte-Urra, I., Warren, H., Rappazzo, A., Velli, M.: 2016, Observational signatures of coronal loop heating and cooling driven by footpoint shuffling. *The Astrophysical Journal* **817**(1), 47. [dahlburg2016]

Dungey, J.W.: 1958, 15. The neutral point discharge theory of solar flares. A reply to Cowling's criticism. *Symposium - International Astronomical Union* **6**, 135. DOI. [Dungey1958]

Einaudi, G., Velli, M., Politano, H., Pouquet, A.: 1996, Energy release in a turbulent corona. *The Astrophysical Journal* **457**(2), L113. [einaudi1996]

Galsgaard, K., Nordlund, Å.: 1997, Heating and activity of the solar corona: 3. Dynamics of a low beta plasma with three-dimensional null points. *Journal of Geophysical Research: Space Physics* **102**(A1), 231. DOI. http://doi.wiley.com/10.1029/96JA02680. [Galsgaard1997]

Gekelman, W.: 1999, Review of laboratory experiments on alfvén waves and their relationship to space observations. *Journal of Geophysical Research: Space Physics* **104**(A7), 14417. [gekelman1999]

Heyvaerts, J., Priest, E.R.: 1983, Coronal heating by phase-mixed shear alfvén waves. *Astronomy and Astrophysics* **117**, 220. [heyvaerts1983]

Hollweg, J.V.: 1978, Alfvén waves in the solar atmosphere. *Solar Physics* **56**(2), 305. [hollweg1978]

Huang, S.Y., Xu, S.B., Zhang, J., Sahraoui, F., Andrés, N., He, J.S., Yuan, Z.G., Deng, X.H., Jiang, K., Wei, Y.Y., Xiong, Q.Y., Wang, Z., Yu, L., Lin, R.T.: 2022, Anisotropy of magnetic field spectra at kinetic scales of solar wind turbulence as revealed by the parker solar probe in the inner heliosphere. *The Astrophysical Journal Letters* **929**(1), L6. DOI. https://doi.org/10.3847/2041-8213/ac5f02. [Huang_2022]

Ionson, J.A.: 1978, Resonant absorption of alfvénic surface waves and the heating of solar coronal loops. *The Astrophysical Journal* **226**, 650. [ionson1978]

Longcope, D.W., Parnell, C.E.: 2009, The Number of Magnetic Null Points in the Quiet Sun Corona. *Solar Physics* **254**(1), 51. DOI. http://link.springer.com/10.1007/s11207-008-9281-x. [Longcope2009]

Malara, F., Nigro, G., Valentini, F., Sorriso-Valvo, L.: 2019, Electron heating by kinetic alfvén waves in coronal loop turbulence. *The Astrophysical Journal* **871**(1), 66. [malara2019electron]

Matthaeus, W., Lamkin, S.L.: 1986, Turbulent magnetic reconnection. *The Physics of fluids* **29**(8), 2513. [matthaeus1986]

Matthaeus, W.H., Goldstein, M.L., Roberts, D.A.: 1990, Evidence for the presence of quasi-two-dimensional nearly incompressible fluctuations in the solar wind. *Journal of Geophysical Research Space Physics* **95**(A12), 20673. [matthaeus1990]

McLaughlin, J.A., Hood, A.W.: 2004, MHD wave propagation in the neighbourhood of a two-dimensional null point. *Astronomy and Astrophysics* **420**(3), 1129. DOI. http://www.aanda.org/10.1051/0004-6361:20035900. [McLaughlin2004]

McLaughlin, J.A., Hood, A.W.: 2005, MHD wave propagation in the neighbourhood of two null points. *Astronomy and Astrophysics* **435**(1), 313. DOI. [McLaughlin2005]

Oughton, S., Priest, E.R., Matthaeus, W.H.: 1994, The influence of a mean magnetic field on three-dimensional magnetohydrodynamic turbulence. *Journal of Fluid Mechanics* **280**, 95. [oughton1994]

Pontin, D.I., Galsgaard, K.: 2007, Current amplification and magnetic reconnection at a three-dimensional null point: Physical characteristics. *Journal of Geophysical Research: Space Physics* **112**(A3), n/a. DOI. http://doi.wiley.com/10.1029/2006JA011848. [Pontin2007]

Rickard, G., Titov, V.: 1996, Current accumulation at a three-dimensional magnetic null. *The Astrophysical Journal* **472**(2), 840. [rickard1996current]

Schwarzschild, M.: 1948, On noise arising from the solar granulation. *The Astrophysical Journal* **107**, 1. [schwarzschild1948]

Sharma, R., Kumar, S.: 2010, Landau damped kinetic alfvén waves and coronal heating. *Journal of plasma physics* **76**(2), 239. [sharma2010landau]

Shebalin, J.V., Matthaeus, W.H., Montgomery, D.: 1983, Anisotropy in mhd turbulence due to a mean magnetic field. *Journal of Plasma Physics* **29**(3), 525. [shebalin1983]

Shibata, K., Magara, T.: 2011, Solar Flares: Magnetohydrodynamic Processes. *Living Reviews in Solar Physics* **8**. DOI. http://link.springer.com/10.12942/lrsp-2011-6. [Shibata2011]

Stein, R.F., Leibacher, J.: 1974, Waves in the Solar Atmosphere. *Annual Review of Astronomy and Astrophysics* **12**(1), 407. DOI. https://www.annualreviews.org/doi/10.1146/annurev.aa.12.090174.002203. [Stein1974]

Strauss, H.R.: 1976, Nonlinear, three-dimensional magnetohydrodynamics of noncircular tokamaks. *The Physics of Fluids* **19**(1), 134. [strauss1976]

Sweet, P.A.: 1958, 14. The neutral point theory of solar flares. *Symposium - International Astronomical Union* **6**, 123. DOI. [Sweet1958]

Thurgood, J.O., McLaughlin, J.A.: 2013, Nonlinear Alfvén wave dynamics at a 2D magnetic null point: ponderomotive force. *Astronomy and Astrophysics* **555**, A86. DOI. http://www.aanda.org/10.1051/0004-6361/201321338. [Thurgood2013]

Van Ballegooijen, A., Asgari-Targhi, M., Cranmer, S., DeLuca, E.: 2011, Heating of the solar chromosphere and corona by alfvén wave turbulence. *The Astrophysical Journal* **736**(1), 3. [van2011]

Van Doorsselaere, T., Nakariakov, V.M., Verwichte, E.: 2008, Detection of Waves in the Solar Corona: Kink or Alfvén? *The Astrophysical Journal* **676**(1), L73. DOI. https://iopscience.iop.org/article/10.1086/587029. [VanDoorsselaere2008]

Vlahos, L., Isliker, H.: 2018, Particle acceleration and heating in a turbulent solar corona. *Plasma Physics and Controlled Fusion* **61**(1), 014020. [vlahos2018particle]

Wentzel, D.G.: 1974, Coronal heating by Alfv n waves. *Solar Physics* **39**(1), 129. DOI. http://link.springer.com/10.1007/BF00154975. [Wentzel1974]

Wu, D.J., Chao, J.K.: 2004, Recent progress in nonlinear kinetic alfvén waves. *Nonlinear Processes in Geophysics* **11**(5/6), 631. DOI. https://npg.copernicus.org/articles/11/631/2004/. [DJwu]

Yadav, N., Sharma, R.P.: 2014, Nonlinear Interaction of 3D Kinetic Alfvén Waves and Ion Acoustic Waves in Solar Wind Plasmas. *Solar Physics* **289**(5), 1803. DOI. [Yadav2014]

Geocarto International

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgei20

# Novel Ensemble of Deep Learning Neural Network and Support Vector Machine for Landslide Susceptibility Mapping in Tehri Region, Garhwal Himalaya

Sunil Saha, Anik Saha, Tusar Kanti Hembram, Barnali Kundu & Raju Sarkar

**To cite this article:** Sunil Saha, Anik Saha, Tusar Kanti Hembram, Barnali Kundu & Raju Sarkar (2022): Novel Ensemble of Deep Learning Neural Network and Support Vector Machine for Landslide Susceptibility Mapping in Tehri Region, Garhwal Himalaya, Geocarto International, DOI: 10.1080/10106049.2022.2120638

**To link to this article:** https://doi.org/10.1080/10106049.2022.2120638

Accepted author version posted online: 02 Sep 2022.

Submit your article to this journal

Article views: 11

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tgei20

Check for updates

# Novel Ensemble of Deep Learning Neural Network and Support Vector Machine for Landslide Susceptibility Mapping in Tehri Region, Garhwal Himalaya

**Sunil Saha[1], Anik Saha[2], Tusar Kanti Hembram[3], Barnali Kundu[4], Raju Sarkar[5*]**

[1]Department of Geography, University of Gour Banga, Malda-732103, West Bengal, India
E-mail: sunilgeo.88@gmail.com

[2]Department of Geography, University of Gour Banga, Malda-732103, West Bengal, India
E-mail: aniksahamalda@gmail.com

[3]Department of Geography, Nistarini Women's College, Purulia, West Bengal, India
E-mail:tusarpurulia1991@gmail.com

[4]Department of Geography, University of Gour Banga, Malda-732103, West Bengal, India
E-mail:barnalimalda29@gmail.com

[5]Department of Civil Engineering, Delhi Technological University, Delhi - 110 042, India
E-mail: rajusarkar@dce.ac.in

*Corresponding: rajusarkar@dce.ac.in (Raju Sarkar)

Accepted Manuscript

## Abstract

Over the years, landslide has become one of the most destructive events that can happen in hilly areas. Tehri, a region in the Himalayas is no different. Current research aids in the construction of ensemble models of DLNN and SVM, which are then compared with various SVM kernels. Landslide susceptibility mapping in the Tehri region of the Himalayas has been worked upon using a deep learning (DLNN), four machine learnings (SVM-RBF, SVM-Linear, SVM-Polynomial, SVM-Sigmoid), and their novel ensembles i.e. DLNN with SVM-RBF, DLNN with SVM-Linear, DLNN with SVM-Polynomial and DLNN with SVM-Sigmoid. 16 environmental, geographical, hydrological, soil condition and topographical landslide conditioning variables (LCVs) have been considered. These models were trained using 70% of inventory landslides and tested using 30% of the same. The results revealed the superiority of DLNN, DLNN-SVM (RBF), DLNN-SVM (Linear) models which quantified 28.32%, 26.96%, and 22.41% of the area highly susceptible for landslide, respectively.

**Keywords:** Landslide susceptibility map; deep learning algorithms; Tehri region; ensemble model; Kernels

## 1. Introduction

Landslides, mainly in the hilly regions, are the significant environmental disasters that cause the demolition of natural and human habitats, roads and other transport routes, upland agriculture,

deforestation, and several other economic and cultural structures. In the field of Engineering Geology, landslides are defined as the downhill movement of mass material on a slope (Mandal et al. 2021). Globally, hilly regions are intensely influenced by landslide events due to high instability of landmasses over the slope (Conforti & Ietto, 2019). In India, states and union territories (UTs) around the Himalayan mountainous belts like Jammu & Kashmir, Himachal Pradesh, Uttarakhand, Northern part of West Bengal, Sikkim and North-Eastern states are rigorously affected by the landslides. Mainly the landslides are the product of a large multiplicity of geo-environmental processes which entail meteorological, geological and anthropogenic factors. The key factors which affect landslides were talked about by Kornejady et al (2015) and Nepal et al. (2019). Mainly the arrangement of substratum geology (lithological structure, formation, amount of weathering), geomorphology (inclination of slope, relative relief & aspect), soil type (deepness, arrangement, porosity and permeability), hydrological and anthropogenic factors (lands use, road) are the significant intrinsic factors of the landslide (Haoet al. 2020). Landslides are activated by numerous external causal factors like rain, seismic activity, explosion and drilling, heavy downpour, and flash floods. In numerous studies, it has been found that the reservoir has brought negative effects on the adjacent socio-physical environment (Nilsson& Berggren, 2000). Modification of slopes in high altitude regions due to excavation and reservoir construction, causes slope instability. Gradual and increasing sedimentation into the reservoir causes excessive pressure on the underlying rock beds. Under such condition of hydrology and lithology, structural disruption is the major cause for the instability of the slope in the reservoir edge region.

Therefore, spatial prediction of landslide-prone regions might be an imperative and initially standard tool for decision-making in the field of disaster management (Moharrami et al. 2020). Recognition of the zones which are probable for the landslide is based on the supposition that in future the landslides will happen under those circumstances which enhances the instability in the past and present (Saponaro et al. 2015; Chen et al. 2015). In landslide studies, different scholars described zonation as the dissection of the land into regions and ranking those regions according to the amount of definite hazard of landslides on the slopes (Kanungo et al. 2006; Bora 2016). However, based on the absolute field inspection in the inaccessible areas, prediction is not possible and it is very much time and cost consuming as well as requires a venturesome and a hard working person (Kanungo et al. 2006; Daneshvar, 2014). To resolve these issues to some extent, remote sensing data and geo-informatics offers successful means to map landslides with advanced techniques. An aerial coverage of a landslide-prone area earlier to rainfall and after the rainfall to can be determined based on temporal resolution sensing data (Pradhan, 2010, Sameen et al., 2020).

In the context of GIS, a variety of qualitative and quantitative methodologies have been utilised for mapping the susceptibility to landslides. Qualitative methods assume subjective knowledge of experts for weighting the factors which affect landslides (Vahidnia et al 2010; He et al

Accepted Manuscript

2019). The use of semi-quantitative procedures such as the analytical hierarchy process, combined landslide frequency ratio, fuzzy logic approach, and weighted linear combination depends on the weighting and rating of the components (WLC) (Van et al., 2016; Psomiadis et al. 2020). The subjective approach based on GIS has been used by Meena et al. (2019), Pham et al. (2018) in various portions of the Himalayan region. As like Chauhan et al (2010), Das et al. (2010) applied multivariate logistic regression technique for the identification of landslide hazard zones in Indian Himalayas. In the Himalayan region, the Fuzzy logic approach has been carried out by Anbalagan et al. (2015). Li et al. (2020) used Geographically Weighted Regression for mapping the landslide susceptibility in the Qingchuan Area of China and got good result. However, with the use of recent advanced computational intelligence i.e. machine learning classifiers and deep learning approaches, the accuracy of the mapping landslide susceptibility has been improved (Van Dao et. Al., 2020; Tang et al., 2020; Ngo et al., 2021; Akinci, 2022). These have some advantages over the aforementioned methods i.e. can deal with the massive dataset, have intelligence of advanced classification, wider framework of analysis and ease of handling inputs (Ngo et al., 2021, Saha and Saha (2022a).

The performance of ML models in landslide susceptibility mapping (LSM) has approved relatively better than those of the traditional methods (Mandal et al., 2021; Akinci et al,. 2021a). Among the most popular and success ML models of LSM are Random Forest (Achour et al., 2020; Akinci et al., 2020), Boosted Regression Tree (Pourghasemi et al. 2021; Saha et al, 2021), Naive Bayes (He et al. 2019; Youssef et al., 2021), ANN (Saha et al.,2020a; Youssef et al., 2021), SVM (Roy et al., 2019; Akinci et al., 2021a) etc. The detailed review of the ML models in respect of LSM have articulated by Merghadi et al. 2020; Huang et al. 2020. Deep learning (DL) models have also performed better and in some cases showed more efficiency (Wang et al., 2020). Convolutional Neural Network (CNN), Deep Learning Convolutional Neural Network (DCNN) and Deep Learning Neural Network (DLNN) are effectively and widely used in modelling hazards. The comparison between ML and DL models in landslide mapping has shown by Mohan et al. (2020); Bui et al. (2020). These literatures have shown although some satisfactory results but also, have recommended advancement in this regard. Ensemble of two or more efficient can be complementary and can provide a better outcome in this regard. The studies on ensemble models have shown more accurate LSM mapping such as Kadavi et al. (2018); Frang et al. (2021).

Now the main research question is if we make an ensemble of DL and ML algorithms, the accuracy of the landslide map will increase or not. In the current work, landslide susceptibility assessment has been performed by applying DL and ML models. The ensemble approach is a type of learning in which a number of models, such as classifiers, are created and merged in a systematic manner in order to address a particular computational intelligence issue. The ensemble approach is mostly used to improve a model's performance (classification, estimation, etc.) or reduce the likelihood of an unforeseen poor model being chosen. However, the application of ensemble model of

Accepted Manuscript

DLNN and SVM integrating with RS-GIS is new in India. In this regard, recent research helps to create ensemble models of DLNN and SVM and compares them with various SVM kernels for the prediction of landslide susceptibility. Therefore, this study's objectives are to: (1) assess the effectiveness of the DLNN, SVM-RBF, SVM-Linear, SVM-Polynomial, and SVM-Sigmoid in assessing landslide susceptibility; (2) rank the LCFS; and (3) provide spatial information and direction in decision-making processes regarding landslide management in the study area.

## 2. Study area

The Tehri reservoir rim area is located in the considerably undulating topography of the Lesser Himalayan section (Fig.1). In between the Bhagirathi and Bhilangana Rivers, the Tehri reservoir was constructed and extended in length up to 70 km. Many settlements, farmlands and secondary forests has been grown in this moderate to steep sloped territory. This region is characterized by ridges, cliffs, valleys, spurs, mature terraces and abrupt steep slopes that are deeply dissected. Thick and spare forests cover the spurs and ridges, whereas settlements, farms and agricultural land are typically distributed on the moderately slanted slopes. There are several developmental activities on the boundary slopes of the Tehri reservoir. This reservoir is filled with water in the monsoon season and it dries out during the summer season. Besides, a certain amount of recession occurs throughout the reservoir operation. The reservoir drawdown happens in the summer and rainy seasons between the dead storage level (740 m) and the full reservoir level (MRL 830 m) and this absorbs bounding slopes as the reservoir hits the MRL. The saturated slopes collapse in various positions at the moment of sudden drawdown.

Instability in the slopes of the reservoir boundary region is dependent on the following factors: (i) the materials which formed the slope, (ii) the geometry of the slope, (iii) the vegetation cover on the slopes, and (iv) the activity induced by humans all along the reservoir. The alteration of the reservoir's water leads to the unsteadiness of cliffs, leading to landslides, subsidence, sinkholes, etc. Typical reservoir-induced landslides have been found to persist on the way to the upper reaches of slopes and intruding villages, which emerged at the reservoir level. The reservoir region comes under the division of the great Himalaya which is called the Lesser Himalaya. In the Lesser Himalaya, the rocks of the Tehri reservoir area are characterized by a number of lithological units. The major rock units covering the area are Chandpur, Rautgara, Nagthat, Krol, Jaunsar, Blaini, Damtha, Tejam, Deoban and Mussoorie groups. The central portion of the reservoir edge region is characterised by moderate to low-grade phyllite and extremely worn quartzite. Structural discontinuities like joints and plane of foliation makes quartzite and phyllite susceptible to landslides. Because of the frequent appearances of shearing and weathered tops, rocks of this region are vulnerable to landslides at places.

**Figure 1.** SOMEWHERE HERE

## 3. Material and methods

Accepted Manuscript

### 3.1 Data types and sources

All the required data and their types, sources and other details have been given in Table 1.

The detailed methodology has been briefed in the following heads (Fig. 2):

a) First of all, 16 conditioning factors have been considered in 5 groups such as topographical, hydrological, soil quality, geological and environmental factors by applying the Relief-F factor selection technique.

b) Landslide inventory map has been prepared and also the training and testing dataset has been prepared for modelling and validation purpose.

c) Modelling process has been performed for each of the selected individual and ensemble models i.e. SVM (RBF), SVM (Linear), SVM (Polynomial), SVM (Sigmoid), DLNN, DLNN-SVM (RBF), DLNN-SVM (Linear), DLNN-SVM (Polynomial), DLNN-SVM (Sigmoid).

d) Success and prediction performance of each model is evaluated by applying multi-statistical measures such as sensitivity, specificity, accuracy, F1 measure, Jaccard coefficient, Matthews correlation coefficient(MCC), root mean square error(RMSE) and area under the receiving curve (AUC)

e) The contribution of landslide conditioning factors in producing LSM has been evaluated using frequency ratio.

**Figure 2.** SOMEWHERE HERE

**Table 1.** SOMEWHERE HERE

### 3.2 Landslide inventory map (LIM)

An inventory map of landslides is a significant input of susceptibility analysis of landslide employing various methods (Kalantar et al. 2018; Kavzoglu et al. 2014). It represents the positions of past landslides incidences that have left strong impacts in a region but they don't inculcate on sequential development and differences. Landslide inventory maps (LIM) serve as foundation for further additional maps. A LIM has been prepared by means of collected data during a field survey using handheld GPS and also from the past recorded database (Fig. 3). The field survey was conducted in April 2019. A total of 165 landslide locations have been considered of which 70% is considered as training set and the rest of 30% as testing set. Splitting of the training and testing data set has been done by using the random subset technique. At the time of field survey area, width and depth of landslide was measured. The largest landslide was 512 $m^2$ and the smallest landslide was 106 $m^2$. The average width of the landslide was 27m and average length was 135m.

**Figure 3. SOMEWHERE HERE**

### 3.3 Landslide conditioning factors (LCFs)

Accepted Manuscript

For modelling the landslide susceptibility using scientific method consideration of appropriate LCFs is essential (Achour et al., 2020; Akinci et al., 2020; Pourghasemi et al. 2021; Saha et al, 2021; He et al. 2019; Youssef et al., 2021). There are no specific rules for selecting the LCFs. In the present first literatures on landslide have been reviewed (Akinci et al., 2020; Pourghasemi et al. 2021; Saha et al, 2021; He et al. 2019; Youssef et al., 2021). Then the assortment of the landslide conditioning factors (LCFs) has been done mainly based on the accessibility and the relevancy of data in respect of a landslide event. Selected 16 LCFs for the present study have been detailed below (Fig. 4):

***Elevation:*** Variation of elevation in each area is one of the most influential factors in case of soil erosion and movement of slope mass (Bunn et al. 2019). Commonly, elevation determines various natural and biological factors, like rainfall, temperature, vegetation cover, and the human activity. As a result, these physical situations have determined the stability of slope and generate the failure of slope. Elevation of this area has been divided into 5 classes (Fig. 4A) and the maximum and minimum height has been found as 2692m and 591m respectively.

***Slope angle and slope aspect:*** The probability of landslide primarily depends on the region's slope angle (Psomiadis et al. 2020). The high degree of slope raises the chances of landslides and in the gentle slope, there is less possibility of landslides (Abedin et al. 2020). The slope angle of the study area was divided into five classes (Fig. 4B). Landslide susceptibility is also caused by the topographical aspect, i.e. slope orientation (Ma et al. 2019). On any slope, the concentration of sunlight which is related to the temperature and with other climatic condition is largely governed by the slope aspect.

***Curvature:*** Curvature is also a significant factor that affects landslides (Chen & Li 2020). With decreasing curvature value, the possibility of occurrence of landslide decreases (Chen & Li 2020). Three groups, such as concave, flat and convex, were classified in the curvature map (Fig. 4D) of the study area.

***Rainfall:*** High amount of rainfalls influence the soil particles by adding extra amount of saturation to the soil profiles that enhances the regular incidences of landslides (Fan et al. 2016). It is expressed that the soil and rock are essentially separated by intensive precipitation, leading to a high probability of landslide occurrence. The derived rainfall map showed that in different parts of the study area, the amount of rainfall varies from 950mm to 1450mm.

***Geology:*** Various geological aspects like underlying rock structure, substances of the terrain, rock history of a particular area largely determine the stability of a slope (Zhao & Chen, 2020). The total study area has formed under 7 different geologic formations, viz. Mandholi formation, Chandrpur formation, Barinag formation, Deoban formation, Nautgara formation, Blaini formation, and Nagthat formation (Fig. 4F). The major rock units covering the area are Chandpur, Rautgara, Nagthat, Krol, Jaunsar, Blaini, Damtha, Tejam, Deoban and Mussoorie groups. The central portion of the reservoir edge region is characterised by moderate to low-grade phyllite and extremely worn quartzite.

Accepted Manuscript

Structural discontinuities like joints and plane of foliation makes quartzite and phyllite susceptible to landslides.

***Lineament density:*** The incidence of landslides is highly influenced by various geological structures such as fault, fold, fracture, etc. (Bunn et al. 2019). Higher lineament density raises the risk of a landslide as it reduces the cohesiveness of rock materials (Bera et al. 2019). The study area has been divided into 5 lineament density zones in the following diagram (Fig. 4G).

***Land use/ landcover (LULC):*** In the assessment of landslide threat, the LULC configuration of a region plays a critical role (Singh & Kumar 2017). LULC pattern of an area plays a significant role in the study of landslide hazard (Singh & Kumar 2017). LULC map has been derived from Landsat 8 satellite imagery performing supervised image classification technique. The categories classified are settlement, agriculture, dense forest, open forest, scrub forest, and barren land (Fig. 4K).

***Distance from roads:*** A huge number of landslides occurred on the unprocessed cut-slopes, which take place through the road construction (Al-Najjar et al. 2019). Cut-slopes usually dip in the direction of the road at steep angles. The incidence of landslides on these cut-slopes arises during the rainy season.

***NDVI:*** Landslide is a combined process that is influenced by various attributes of vegetation (Naidu et al. 2018). High vegetation cover protects the slope from landslide events. Bare ground, on the other hand, is more vulnerable to landslides. The NDVI value when closer to 1 implies a high degree of greenness, while closer to -1 denotes minimal or absence of vegetation cover over an area. The mathematical expression of NDVI computation is as f Eq. 1.

$$NDVI = \frac{NIR - R}{NIR + R} \quad \text{(Eq. 1)}$$

***Soil type and soil depth:*** Soil types and depth play a significant role in determining susceptibility of landslides (Wang et al. 2020). For its low cohesiveness, the coarse grain soil particles are more likely to fragment. Four types of soil have been identified mainly in this study area viz. forest sandy loam, alluvial sandy loam, forest black clay, and hill granular sandy soil (Fig. 4H). The risk of landslide increases with the increasing soil depth since the cohesion of the soil particles is lower as compared to the rock particles (Almeida et al. 2017).

***Topographic Position Index (TPI):*** TPI is a responsible factor in the process of flow of debris and also defines the appearance of the settings of geomorphology (valley, ridge and slope etc) (Riley et al. 2017). This index is also considered as an influencing parameter of landslide since the incidence of landslide habitually occurs on the ridges. Based on the TPI values, 6 different types of landforms, such as valley, lower slope, flat slope, middle slope, high slope, and ridge have been identified for this region (Fig. 4J).

Accepted Manuscript

***Topographic Wetness Index (TWI):*** TWI is generally used as a tool to predict the quantity of soil moisture of an area (Raduła et al. 2018; Kavzoglu et al. 2014) The areas with high wetness are more susceptible to landslide (Moayedi et al. 2019). ***Topographic Roughness Index (TRI):*** TRI is a morphometric measure that represents the non-homogenous state of a surface. TRI displays the local variation of curvatures or surface gradients (Różycka et al.2017). The measurement of the TRI was intended to quantify the topographical heterogeneity by determining the unevenness of the slope or elevation within a certain location.

***Sediment Transport Index (STI):*** STI is strongly associated with the release of sediments from land into the channel (Meena et al. 2019a). This sediment supply uncertainty varies with the terrain, maximum slope angle, rock formation, length of the slope etc. With the increasing potentiality of sediment transport, the probability of landslide also increases and vice versa.

**Table 2.** SOMEWHERE HERE

**Figure 4.** SOMEWHERE HERE

### 3.4 Relief-F and multi-collinearity analysis

First, the Relief-F technique was developed by Kira and Rendell (1992) and then it was further enhanced by Challita et al (2018). It is very common among the different factor selection methods because of the efficiency and clarity of the Relief-F. (Elsayyad et al. 2017; Fang et al. 2020). To measure the average merit (AM) of different LCFs, the Relief-F technique was used in this research. To know the selected factors are independent or not multi-collinearity test has been carried out using variance influence factor (VIF) and tolerance method. Regarding the multi-collinearity test can be found in the literature of Gayen et al. (2017).

### 3.5 Deep learning neural network (DLNN)

DLNN is a well-admired machine learning (ML) model which has three layers viz. a layer of input, numerous hidden layers and an output layer (Adil et al. 2020). Fig. 5 depicts the structural layout of the DLNN model used in the current study's mapping of landslide susceptibility. The layer of output has two feasible labels such as a negative label and a positive label. These categorization outputs are attained from the final layer which is hidden and revealed in the layer of output (Bello et al. 2020). DLNN has some specific recompense over the algorithm of traditional machine learning, and therefore, in the ground of forecast analysis, much more importance has been given to it (Bello et al. 2020). The DLNN model differs from other ML models in that it makes the best use of amorphous data, correctly understands the training dataset, is robust enough to recognise novel data, and can construct further learning models by adding more layers to the neural network system. The use of the activation function of Rectified Linear Unit (ReLU) ensures the fee-forward neural network guidance

Accepted Manuscript

process of the DLNN model. The gradient of receding can be reduced with this method. In the literature of Bui et al. (2020) more details about the DLNN can be found.

**Figure 5.** SOMEWHERE HERE

### 3.6 Support Vector Machine (SVM)

The SVM is a well-known algorithm of supervised machine learning which is depending on the theory of statistical learning (Uddin et al. 2019; Ozer et al. 2020; Akinci et al., 2021a). SVM is one of the widely used ML models in the field of predicting various natural hazards like gully erosion (Arabameri et al. 2020), landslide forecasting (Xie et al. 2017), mapping of flood susceptibility (Choubin et al. 2019) etc. SVM is used to resolve the analysis of regression and problems of multi-faceted classifier. Based on the principle of statistical learning and minimization of structural risk the SVM is formulated. SVMs have the capability to reduce a higher occurrence of the superior error. This technique divides the training dataset from formation of hyper-plane. Dividing hyper-plane (HyPl) is geared up in the actual space of n coordinates (xi parameters in vector x) among the points of two different classes (Robotham et al. 2011; Bui et al., 2012). Highest limit of division among the classes is revealed by SVM and consequently, it constructs a categorization HyPl in the central of the utmost limit. The high dimensional element from the space of input separated classes from the space, the divider is a constructive linear (Robotham et al. 2011, Saha et al., 2020b). In the present study four kernels i.e. linear, polynomial, radial basis function and sigmoid have been used for preparing the LSMs. Details regarding the SVM can be found in the literatures of Tehrany et al. (2014) and Roy et al. (2019). Afterwards, the ensemble models of DLNN and SVM with different kernels, were prepared to compare the results of individual models and ensemble models.

### 3.7 Frequency Ratio (FR)

The frequency ratio is defined as the ratio of landslide occurrence possibilities to non-landslide possibilities in a specific region (Shirani et al. 2018). This model assumes the spatial relationship between landslide incident locations and various factors affecting the frequency of landslides, improves the accuracy of categorization and the connection between landslide locations and various factors in the study area (Dou et al. 2019). FR is used to evaluate the relative significance of LCFs. For implementing the FR model, all LCFs have been converted into different classes (Table 2) and then statistics of the cells of landslides and non-landslides have been extracted to compute using the formula Eq. 2.

$$Fr_i = \frac{b}{a} = \frac{\frac{landslide_cell_i}{landslide_cell_{tot}}}{\frac{no_landslide_cell_i}{no_landslide_cell_{tot}}} \quad \text{(Eq. 2)}$$

Accepted Manuscript

The FR value represents the degree of association of the LCFs with landslides. FR value >1 and <1 represents a high and low correlation with the incidence of landslide respectively (Akinci et al., 2021b).

### 3.8 Validation techniques

For measuring the accuracy of the used models using different techniques True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) have been used. The appropriately identified pixels for landslide occurrence and absence, respectively, are represented by TP and TN.TP and TN are the appropriately classified pixels for landslide existence and absence, respectively. FP stands for the number of pixels that were mistakenly assigned to the landslide presence class whereas FN stands for the number of pixels that were incorrectly assigned to the landslide absence class. ***MAE:*** By dividing the total variance between the observed and predicted values by the overall number of observations, the mean absolute error (MAE) was calculated (Von Rudorff & Von Lilienfeld, 2020).The mathematical expression is as Eq. 3:

$$MAE = \frac{1}{N}\sum_{i=1}^{n}\left|V_{predict} - V_{observ}\right| \quad \text{(Eq. 3)}$$

***RMSE:*** The square root of the ratio of the difference between anticipated and observed values is the RMSE (Root Mean Square Error) (Juneja et al. 2019).It articulates the mathematical concept as Eq. 4:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}[V_{predict} - V_{observ}]^2}{N}} \quad \text{(Eq. 4)}$$

Where N is size of the sample, $V_{predict}$ is predicted and $V_{observ}$ is the observed values of the dependent variable.

## 4. Results

### 4.1 Relief-F and multi-collinearity test

The Relief-F average merit (AM) results for each of the LCFs are shown in Table 3. Highest AM has been found for elevation followed by geology, soil type and soil depth i.e. the elevation has the highest contribution in the incidence of landslide susceptibility. Lowest AM has been found for the TRI and STI followed by TWI, TPI, aspect and NDVI. TRI and STI have a very insignificant effect on the incidence of landslide in the present study area. To determine if there is multi-collinearity issue, the threshold values of VIF (Variance inflation factor) and TOL (Tolerance) are used: VIF <5 and tolerance >0.2 (Lin et al., 2017; Roy et al., 2019; Ajin et al., 2022; Saha et al., 2022b). Among the selected landslide conditioning factors no one has exceeded the threshold values of VIF and TOL.

Accepted Manuscript

Therefore, multi-collinearity problem does not exist among the selected LCFs. So, all the selected LCFS are independent and can be used for landslide modelling (Table 4).

**Table 3.** SOMEWHERE HERE

**Table 4.** SOMEWHERE HERE

### 4.2 Landslide Susceptibility Maps (LSM)

Produced LSMs of SVM (RBF), SVM (Linear), SVM (Polynomial), SVM (Sigmoid), DLNN, DLNN-SVM(RBF), DLNN-SVM(Linear) DLNN-SVM (Polynomial), DLNN-SVM (Sigmoid) models have been shown in Fig. 6, 7 and 8. Table 5 shows the areal distribution of the landslide vulnerability classes of different models i.e. very low, low, medium, high and very high. In the LSMs of each model, a very high landslide vulnerability zone was observed predominantly along both the banks of the Bhagirathi River (Fig. 6, 7 and 8). The areas of low to very low susceptibility were mainly located around the edge of the basin. These two groups cover an average of over half of the basin area for all models. The propensity to landslide is very strong predominantly towards both the riverbank, but it steadily declines towards the edge of the basin.

**Table 5.** SOMEWHERE HERE

**Figure 6.** SOMEWHERE HERE

**Figure 7.** SOMEWHERE HERE

**Figure 8.** SOMEWHERE HERE

### 4.3 Result of factors importance analysis by FR method

The FR based analysis of the LCFs depicts that every class of LCFs is regarded as a definite landslide occurrence (Table 6). It has been found that in the case of elevation, in the very low elevation class, the FR value is the highest (0.0527) among all classes followed by the class low (0.0137). In case of slope factor, the FR value is the highest in the class of very high (0.0171) followed by the class of high (0.0153) i.e. the landslide susceptibility is higher in very high and high slope regions (Table 6). The northeast-facing slope segments have recorded more landslides followed by the other categories of slope aspect and FR is lowest for flat slope areas. The concave slope class has high susceptibility to landslide in the case of curvature among three classes. In the very low precipitation class (0.0202) and very high precipitation class (0.0166), a high FR value was observed (Table 6). The highest and lowest FR value has been found for the Mondholi geological formation and Chadpur geological formation respectively (Table 6). Similarly, the FR value is highest under forest black loam soil followed by alluvial sandy loam soil class. The table reveals that very highly deep soil has the maximum FR value (0.0178) followed by highly deep soil (0.0122) among all the soil depth groups for the soil depth factor. Past event of landslides are found more prominent in upper slope (0.0207)

Accepted Manuscript

and valley (0.0075) areas as revealed by the factor TPI. It has been observed for the various types of LULC that the FR value is highest for barren land, followed by agricultural land and scrubland. The FR values for Barren Land, Agricultural Land and Scrub are 0.0201, 0.0157 and 0.0150 respectively (Table 6). For the road distance factor, the largest and smallest FR value was observed for the class of 100-200 (0.0186) followed by the class of 50-100 (0.0116) and 200-400 (0.0108)

In case of NDVI, high landslide susceptibility has been found for the class of 0.175 to 0.291(0.0186) followed by the class of 0.087 to 0.175 (0.0116). For TRI, TWI and STI highest FR value have been found for the class of 5.614 to 8.876 (0.0389), 3 to 11.24 (0.0851) and 4.73 to 9.25 (0.0129), respectively (Table 6).

**Table 6.** SOMEWHERE HERE

**4.4. Validation and comparison**

The AUC values using the training dataset for SVM (RBF), SVM (Linear), SVM (Polynomial), SVM (Sigmoid), DLNN, DLNN-SVM(RBF), DLNN-SVM (Linear), DLNN-SVM (Polynomial), DLNN-SVM (Sigmoid) are 84.26%, 83.82%, 82.71%, 80.98%, 91.23%, 87.56%, 86.37%, 85.66%, 84.51% respectively (Table 7). AUC values considering the testing dataset for the mentioned models are 86.17%, 85.82%, 85.11%, 82.56%, 91.94%, 90.72%, 90.08%, 87.19% and 87.08% respectively (Fig. 9). Accuracy value has been found highest in case of DLNN and lowest in case of SVM (Sigmoid). As per the measures of sensitivity, the index values considering training data are 0.951, 0.943, 0.866, 0.865, 0.976, 0.956, 0.94, 0.926 and 0.889 for SVM (RBF), SVM (Linear), SVM (Polynomial), SVM (Sigmoid), DLNN, DLNN-SVM(RBF), DLNN-SVM(Linear), DLNN-SVM (Polynomial), DLNN-SVM (Sigmoid) respectively and for the testing dataset, the index values have been found 0.956, 0.951, 0.906, 0.874, 0.984, 0.965, 0.882, 0.893 and 0.876, respectively (Table 7). Highest sensitivity index value has been found in case of DLNN and lowest value has been found in case of SVM (Sigmoid). In case of specificity index, the values are found 0.827, 0.877, 0.798, 0.798, 0.91, 0.798, 0.881, 0.785, 0.744 for SVM (RBF), SVM (Linear), SVM (Polynomial), SVM (Sigmoid), DLNN, DLNN-SVM (RBF), DLNN-SVM (Linear), DLNN-SVM (Polynomial), DLNN-SVM (Sigmoid) model respectively using the training data and for the testing dataset the values have been found 0.896, 0.865, 0.715, 0.684, 0.872, 0.807, 0.901, 0.784, 0.712 (Table 7). Highest and lowest Specificity values have been found in case of DLNN and SVM (Sigmoid) respectively.

In case of SVM (RBF), using train dataset, the F1 measure, Jaccard, MCC values have been found as 0.901, 0.898, 0.038 and using testing dataset the values are 0.912, 0.906, 0.912, respectively (Table 7). For SVM (Linear) model, the F1 measure, Jaccard, MCC values have been found as 0.878, 0.901, 0.042 and 0.884, 0.908, 0.055 for training and testing dataset, respectively (Table 7). For SVM (Polynomial) model, these values are 0.892, 0.887, 0.872 and 0.898, 0.835, 0.868 for training and testing dataset, respectively (Table 7). In case of SVM (Sigmoid), for training and testing dataset, the F1 measure, Jaccard, MCC values have been found as 0.882, 0.881, 0.865 and 0.875, 0.829, 0.845,

Accepted Manuscript

respectively. For DLNN, for training and testing dataset, these index values are 0.920, 0.950, 0.914 and 0.954, 0.922, 0.939, respectively. Similarly, 0.911, 0.926, 0.911 and 0.954, 0.889, 0.884 values have been found using training and testing datasets for DLNN-SVM(RBF) model. In case of DLNN-SVM(Linear) model, the F1 measure, Jaccard, MCC values have been found as 0.938, 0.911, 0.937 and 0.956, 0.939, 0.884 considering train and test data respectively and for the DLNN-SVM (Polynomial) model these values are 0.938, 0.909, 0.809 and 0.935, 0.875, 0.884 derived using train and test landslides accordingly. In case of DLNN-SVM (Sigmoid) values have been found as 0.938, 0.909, 0.847 and 0.935, 0.875, 0.884 for training and testing datasets respectively for F1 measure, Jaccard, MCC.

Highest and lowest F1 measure values have been found for DLNN-SVM(Linear) and SVM (Sigmoid), respectively (Table 7). Highest Jaccard and MCC values have been found for DLNN and lowest Jaccard and MCC values have been found for SVM (Sigmoid). The MAE values for the training and testing dataset for SVM (RBF), SVM (Linear), SVM (Polynomial), SVM (Sigmoid), DLNN, DLNN-SVM(RBF), DLNN-SVM(Linear), DLNN-SVM (Polynomial), DLNN-SVM (Sigmoid) are 0.038, 0.042, 0.139, 0.126, 0.115, 0.064, 0.049, 0.172, 0.175 and 0.088, 0.055, 0.144, 0.124, 0.037, 0.036, 0.112, 0.163, 0.191 respectively (Table 7). Highest and lowest values for MAE and RMSE values have been found for DLNN-SVM (Sigmoid) and SVM (RBF) respectively.

**Figure 9.** SOMEWHERE HERE

**Table 7.** SOMEWHERE HERE

## 5. Discussion

As per the stated methodology, the feature score or average merit (AM) of landslide-related factors was checked in order to exclude LCFs that seemed to have no predictive potentiality for improving the models' prediction accuracy for landslide mapping (Van Dao et al. 2020). Relief-F assigns a feature score to each feature, which can then be used to rank and select the features with the highest scores. The results of Relief-F shown in the Table 3 revealed that the first eight LCFs that reflected higher AM values are elevation, geology, soil type, soil depth, distance from road, lineament density, slope angle and rainfall respectively. In the highly elevated area the degree of slope is more that allows frequent landslides. If an area has weak geology, loose soil and high soil depth the probability of landslide in that area will be very high. Thus, all the LCFs were examined further and their AM values are greater than "0," a condition that would add "noise" and signify null predictive power, associated with decreased predictive output. Elevation was also highlighted as a major LCF by Bui et al. (2016) in landslide analysis. The priority of other LCFs in terms of feature score or AM values was also found higher in the several pieces of literatures (Kutlag et al., 2017; Van Dao et al., 2020; Dang et al., 2020)

Accepted Manuscript

For the mapping of landslide susceptibility, in the present study, an individual model-based approach using four kernels of SVM such as RBF, Linear, Polynomial and Sigmoid and DLNN model and simultaneously, ensemble of these four kernels with DLNN was adopted for the Tehri region of the Tehri Garhwal Himalaya. DLNN is one of the most efficient deep learning approaches especially in landslide susceptibility modelling as revealed in the literatures of Bui et al. (2020); Costache et al. (2021).

SVM is also a widely accepted machine learning model mostly applied in classification problems and became the model of preference, outperforming rivals around (Marjanović et al. 2011; Hong et al. 2017). The performances of individual models evaluated by AUC-ROC showed that the success rate is highest for the SVM-RBF kernel followed by SVM-Linear, SVM-Polynomial and SVM-Sigmoid. Using the test data, the AUC value is highest for SVM-RBF followed by SVM-Linear, SVM-Polynomial and SVM-Sigmoid. SVM-RBF outperformed all models which were also confirmed by the work of Hong et al. (2017) in landslide susceptibility mapping at Suichuan area. However, the performance of these four kernel functions has differed from the present study in terms of performance hierarchy in landslide studies like Hong et al. (2017), Chen et al. (2018). But these studies have alike results in terms of AUC values as Hong et al. (2017) achieved AUC values 0.833 for RBF kernel, 0.716 for linear kernel, 0.810 for Polynomial and 0.680 for Sigmoid kernel. But the DLNN individually have outperformed as it has achieved success rate value of 91.23% and prediction rate value 91.94%. The ensemble approach of SVM four kernels with DLNN exhibited that DLNN-SVM (RBF) accomplished the best performance followed by DLNN-SVM (Linear), DLNN-SVM (Polynomial) and DLNN-SVM (Sigmoid). The performances of the ensemble methods reflected in the results as AUC of SVM was increased from 84.26% in SVM (RBF) to 87.56% in DLNN-SVM (RBF), 83.82% in SVM (Linear) to 86.37% in DLNN-SVM (Linear), 82.71% in SVM (Polynomial) to 85.66% in DLNN-SVM (Polynomial) and from 80.98% in SVM (Sigmoid) to DLNN-SVM (Sigmoid) using the training datasets which represent the success ability. The improvement of performance also reflected in prediction capability such as AUC was increased by 4.53% in DLNN-SVM (RBF), 4.26% in DLNN-SVM (Linear), 2.08% in DLNN-SVM (Polynomial) and 4.52% in DLNN-SVM (Sigmoid) ensemble model. These results have shown that the proper ensemble technique can improve the prediction output to some extent. This ensemble learning models not only decreases the variance of forecasts but it can also contribute to much more enhanced forecasts than any other single model (Laradji et al., 2015). However, in the present analysis, DLNN individually has outperformed the results of the ensemble models. The other evaluation measures also revealed the similar outcome as of ROC-AUC for all the models employed as shown in Table 7. The excellent predictive performances of DLNN can also be found in studies like Bui et al. (2020a); Huang et al. (2020a).

Accepted Manuscript

The result of Frequency Ratio analysis has revealed that both banks of Bhagirathi River have the high susceptibility of landslide which have varied slope gradient from $10^{o}$ to $69^{o}$, elevation ranges between 591m and 1020m, alluvial sandy loam and forest black clay soil cover and 130 to 140m depth of soil. There are several studies regarding landslide susceptibility that have also confirmed the mentioned nature of surfaces that are more vulnerable to landslides (Ayalew et al., 2005; Dou et al., 2019a).

## 6. Conclusion

In the Tehri reservoir region, the event of the landslide is frequent. The fluctuation of rainfall during monsoon and dry season and the constructional activities like civil structures, roads etc. have provided additional stress in the study region. In the present a novel ensemble of SVM and DLNN models have been applied to comprehend the landslide susceptible area in Tehri Garhwal Himalaya. According the validation techniques all the models have good capability to map the landslide susceptibility in the study region. The results showed that the ensemble models performed better than individual SVM models. The LSMs ascertained that around 41% of the study region along with the bank of Bhagirathi River has fallen under the zone of high to very highly susceptible. There are great role of elevation, geology, soil depth and rainfall in making these areas more susceptible to landslide. Although DLNN has approved as one of the efficient models for LSM in various existing research works with the AUC of 91.94%, the accuracy could be improved and to be verified by assembling other ML and deep learning approaches especially on the scenario of the present study area. Furthermore, the present study could be very indispensable for the planners as well as developers, as the reservoir of Tehri Dam and surrounding ecology are an enormous asset of this country. Hence, proper planning of the Tehri Garhwal Himalayan region is necessary for a long-term perspective.

**Funding**

No funding was received for this work.

**Acknowledgments:**

Authors would like to thanks the inhabitants of basin because they have helped a lot during our field visit. Also authors thankful to Survey of India (SOI), Geological Survey of India (GSI) and USGS to make available all the maps and data required for the study.

**Data availability statement:**

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

**References**

Accepted Manuscript

Abedin, J., Rabby, Y. W., Hasan, I., Akter, H. 2020. An investigation of the characteristics, causes, and consequences of June 13, 2017, landslides in Rangamati District Bangladesh. *Geoenvirn. Dis.*, *7*(1), 1-19.

Achour, Y., Pourghasemi, H. R. 2020. How do machine learning techniques help in increasing accuracy of landslide susceptibility maps? *Geosci. Front*, *11*(3), 871-883.

Adil, M., Ullah, R., Noor, S., Gohar, N. 2020. Effect of number of neurons and layers in an artificial neural network for generalized concrete mix design. *Neural Comp. and Appl.*, 1-9.

Ajin, R.S., Saha, S., Saha, A. Biju, A., Costache, R., Kuriakose, S.L, Enhancing the Accuracy of the REPTree by Integrating the Hybrid Ensemble Meta-Classifiers for Modelling the Landslide Susceptibility of Idukki District, South western India. *J Indian Soc Remote Sens* (2022). https://doi.org/10.1007/s12524-022-01599-4

Akinci, H., 2022. Assessment of rainfall-induced landslide susceptibility in Artvin, Turkey using machine learning techniques. Journal of African Earth Sciences, 191, p.104535.

Akinci, H., Kilicoglu, C., Dogan, S. 2020. Random forest-based landslide susceptibility mapping in coastal regions of Artvin, Turkey. ISPRS International Journal of Geo-Information, 9(9), 553

Akinci, H., Yavuz Ozalp, A. 2021b. Landslide susceptibility mapping and hazard assessment in Artvin (Turkey) using frequency ratio and modified information value model. *Acta Geophysica*, *69*(3), 725-745.

Akinci, H., Zeybek, M. 2021a. Comparing classical statistic and machine learning models in landslide susceptibility mapping in Ardanuc (Artvin), Turkey. *Natural Hazards*, *108*(2), 1515-1543.

Almeida, S., Holcombe, E. A., Pianosi, F., Wagener, T. 2017. Dealing with deep uncertainties in landslide modelling for disaster risk reduction under climate change. *Nat. Haz. and Earth Sys. Sci.*, *17*(2), 225-241.

Al-Najjar, H. A. H., Kalantar, B., Pradhan, B., Saeidi, V. 2019. Conditioning factor determination for mapping and prediction of landslide susceptibility using machine learning algorithms. In *Earth Res. and Enviro. Rem. Sens./GIS Appl. X* (Vol. 11156, p. 111560K). International Society for Optics and Photonics.

Anbalagan, R., Kumar, R., Lakshmanan, K., Parida, S., Neethu, S. 2015. Landslide hazard zonation mapping using frequency ratio and fuzzy logic approach, a case study of Lachung Valley, Sikkim. *Geoenv. Dis.*, *2*(1), 6.

Arabameri, A., AsadiNalivan, O., Saha, S., Roy, J., Pradhan, B., Tiefenbacher, J. P., & Thi Ngo, P. T. 2020. Novel Ensemble Approaches of Machine Learning Techniques in Modeling the Gully Erosion Susceptibility. *Rem. Sens.*, *12*(11), 1890.

Ayalew, L., Yamagishi, H., Marui, H., Kanno, T. 2005. Landslides in Sado Island of Japan: Part II. GIS-based susceptibility mapping with comparisons of results from two methods and verifications. *Eng. geology*, 81(4), 432-445.

Bello, M., Nápoles, G., Sánchez, R., Bello, R., Vanhoof, K. 2020. Deep neural network to extract high-level features and labels in multi-label classification problems. *Neurocomputing*, *413*, 259-270.

Bera, A., Mukhopadhyay, B. P., Das, D. 2019. Landslide hazard zonation mapping using multi-criteria analysis with the help of GIS techniques: a case study from Eastern Himalayas, Namchi, South Sikkim. *Nat. Haz.*, *96*(2), 935-959.

Bora, R. S. 2016. Landslide Hazards Zonation & its Management using Remote Sensing and GIS. *Rem. Sen. for Nat. Res. Manag. & Monitoring*, 286.

Bui, D. T., Hoang, N. D., Martínez-Álvarez, F., Ngo, P. T. T., Hoa, P. V., Pham, T. D., Costache, R. 2020. A novel deep learning neural network approach for predicting flash flood susceptibility: A case study at a high frequency tropical storm area. Sci. Total Environ., 701, 134413.

Accepted Manuscript

Bui, D. T., Tsangaratos, P., Nguyen, V. T., Van Liem, N., & Trinh, P. T. (2020). Comparing the prediction performance of a Deep Learning Neural Network model with conventional machine learning models in landslide susceptibility assessment. *Catena*, *188*, 104426.

Bui, D. T., Tuan, T. A., Klempe, H., Pradhan, B., Revhaug, I. 2016. Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. Landslides, 13(2), 361-378.

Bunn, M. D., Leshchinsky, B. A., Olsen, M. J., Booth, A. 2019. A simplified, object-based framework for efficient landslide inventorying using LIDAR digital elevation model derivatives. *Rem. Sens.*, *11*(3), 303.

Challita, N., Khalil, M. A., Beauseroy, P. 2018. Efficiency and stability of EN-ReliefF, a new method for feature selection. *Int. J. of Com. Aided Eng. and Tech.*, *10*(3), 320-339.

Chauhan, S., Sharma, M., Arora, M. K. 2010. Landslide susceptibility zonation of the Chamoli region, Garhwal Himalayas, using logistic regression model. *Landslides*, *7*(4), 411-423.

Chen, W., Li, Y. 2020. GIS-based evaluation of landslide susceptibility using hybrid computational intelligence models. *Catena*, *195*, 104777.

Chen, W., Panahi, M., Tsangaratos, P., Shahabi, H., Ilia, I., Panahi, S., Ahmad, B. B. 2019. Applying population-based evolutionary algorithms and a neuro-fuzzy system for modeling landslide susceptibility. *Catena*, *172*, 212-231.

Chen, W., Pourghasemi, H. R., Naghibi, S. A. 2018. A comparative study of landslide susceptibility maps produced using support vector machine with different kernel functions and entropy data mining models in China. *Bull. of Eng. Geology and the Envir.*, 77(2), 647-664.

Choubin, B., Moradi, E., Golshan, M., Adamowski, J., Sajedi-Hosseini, F., Mosavi, A. 2019. An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. *Sci. Total Environ.*, *651*, 2087-2096.

Conforti, M., Ietto, F. 2019. An integrated approach to investigate slope instability affecting infrastructures. *Bull. of Eng. Geology and the Envir.*,*78*(4), 2355-2375.

Costache, R., Arabameri, A., Blaschke, T., Pham, Q. B., Pham, B. T., Pandey, M., Costache, I. 2021. Flash-Flood Potential Mapping Using Deep Learning, Alternating Decision Trees and Data Provided by Remote Sensing Sensors. *Sensors*, 21(1), 280.

Daneshvar, M. R. M. 2014. Landslide susceptibility zonation using analytical hierarchy process and GIS for the Bojnurd region, northeast of Iran. *Landslides*, *11*(6), 1079-1091.

Dang, V. H., Hoang, N. D., Nguyen, L. M. D., Bui, D. T., Samui, P. 2020. A novel GIS-based random forest machine algorithm for the spatial prediction of shallow landslide susceptibility. Forests, 11(1), 118.

Dhibi, K., Fezai, R., Mansouri, M., Trabelsi, M., Kouadri, A., Bouzara, K., Nounou, M. 2020. Reduced kernel random forest technique for fault detection and classification in grid-tied PV systems. *IEEE J.of Photo.*, *10*(6), 1864-1871.

Dou, J., Yunus, A. P., Bui, D. T., Merghadi, A., Sahana, M., Zhu, Z., Pham, B. T. 2019. Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Sci. total environ.*, *662*, 332-346.

Dou, J., Yunus, A. P., Tien Bui, D., Sahana, M., Chen, C. W., Zhu, Z., . Thai Pham, B. 2019a. Evaluating GIS-based multiple statistical models and data mining for earthquake and rainfall-induced landslide susceptibility using the LiDAR DEM. *Rem. Sens.*, 11(6), 638.

Elsayyad, A., Al-Dhaifallah, M., Nassef, A. M. 2017. Features selection for arrhythmia diagnosis using Relief-F algorithm and support vector machine. In *2017 14th Int. Multi-Conference on Sys., Signals & Devices (SSD)* (pp. 461-468). IEEE.

Accepted Manuscript

Fan, L., Lehmann, P., Or, D. 2016. Effects of soil spatial variability at the hillslope and catchment scales on characteristics of rainfall-induced landslides. *Water Res. Research*, *52*(3), 1781-1799.

Fang, Z., Wang, Y., Peng, L., Hong, H. 2020. A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping. *Int. J. Geographical Infor. Sci.*, 1-27.

Gayen, A., Saha, S., 2017. Application of weights-of-evidence (WoE) and evidential belief function (EBF) models for the delineation of soil erosion vulnerable zones: a study on Pathro river basin, Jharkhand, India. Model. Earth Syst. Environ. 3 (3), 1123–1139. https://doi.org/10.1007/s40808-017-0362-4.

Ghosh, S., Dasgupta, A., Swetapadma, A. 2019. A study on support vector machine based linear and non-linear pattern classification. In *2019 Inter. Con. Intelligent Sus.Sys. (ICISS)* (pp. 24-28). IEEE.

Hao, L., Rajaneesh, A., van Westen, C., Sajinkumar, K. S., Martha, T. R., Jaiswal, P., McAdoo, B. G. 2020. Constructing a complete landslide inventory dataset for the 2018 monsoon disaster in Kerala, India, for land use change analysis. *Earth Sys. Sci. Data*, *12*(4), 2899-2918.

He, K., Li, Y., Ma, G., Hu, X., Liu, B., Ma, Z., Xu, Z. 2020. Failure mode analysis of post-seismic rockfall in shattered mountains exemplified by detailed investigation and numerical modelling. *Landslides*, 1-22.

He, Q., Shahabi, H., Shirzadi, A., Li, S., Chen, W., Wang, N., Chai, H., Bian, H., Ma, J., Chen, Y. and Wang, X., 2019. Landslide spatial modelling using novel bivariate statistical based Naïve Bayes, RBF Classifier, and RBF Network machine learning algorithms. *Science of the total environment*, *663*, pp.1-15.

He, Q., Shahabi, H., Shirzadi, A., Li, S., Chen, W., Wang, N., Wang, X. 2019. Landslide spatial modelling using novel bivariate statistical based Naïve Bayes, RBF Classifier, and RBF Network machine learning algorithms. *Sci. total environ.*, *663*, 1-15.

Hong, H., Pradhan, B., Bui, D. T., Xu, C., Youssef, A. M., Chen, W. 2017. Comparison of four kernel functions used in support vector machines for landslide susceptibility mapping: a case study at Suichuan area (China). *Geomatics, Nat.Haz. and Risk*, 8(2), 544-569.

Huang, F., Cao, Z., Jiang, S. H., Zhou, C., Huang, J., Guo, Z. 2020. Landslide susceptibility prediction based on a semi-supervised multiple-layer perceptron model. *Landslides*, *17*(12), 2919-2930.

Huang, F., Zhang, J., Zhou, C., Wang, Y., Huang, J., Zhu, L. 2020a. A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction. *Landslides*, 17(1), 217-229.

Human activities like expansion of agricultural land, settlement, construction of road etc. in the mountain area accelerate the rate of landslide.

Juneja, V. K., Golden, C. E., Mishra, A., Harrison, M. A., Mohr, T., Silverman, M. 2019. Predictive model for growth of Bacillus cereus during cooling of cooked rice. *Int. j. food microbiology*, *290*, 49-58.

Kalantar, B., Pradhan, B., Naghibi, S. A., Motevalli, A., Mansor, S. 2018. Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN). *Geomatics, Nat.Haz. and Risk*, *9*(1), 49-69.

Kanungo, D. P., Arora, M. K., Sarkar, S., Gupta, R. P. 2006. A comparative study of conventional, ANN black box, fuzzy and combined neural and fuzzy weighting procedures for landslide susceptibility zonation in Darjeeling Himalayas. *Eng. Geology*, *85*(3-4), 347-366.

Kavzoglu, T., Sahin, E. K., Colkesen, I. 2014. Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides*, *11*(3), 425-439.

Kim, P. Matlab deep learning. Mach. Learn. Neural Netw. Artif. Intell. 2017, 130, 21.

Accepted Manuscript

Kornejady, A., Heidari, K., Nakhavali, M. 2015. Assessment of landslide susceptibility, semi-quantitative risk and management in the Ilam dam basin, Ilam, Iran. *Environ. Res. Research*, *3*(1), 85-109.

Kumar, A., Sharma, R. K., Bansal, V. K. 2018. Landslide hazard zonation using analytical hierarchy process along National Highway-3 in mid Himalayas of Himachal Pradesh, India. *Environ. Earth Sci.*, *77*(20), 719.

Laradji, I. H., Alshayeb, M., Ghouti, L. 2015. Software defect prediction using ensemble learning on selected features. *Info. and Software Tech.*, 58, 388-402.

Lin, G. F., Chang, M. J., Huang, Y. C., & Ho, J. Y. (2017). Assessment of susceptibility to rainfall-induced landslides using improved self-organizing linear output map, support vector machine, and logistic regression. *Engineering Geology*, *224*, 62-74.

Ma, Z., Qin, S., Cao, C., Lv, J., Li, G., Qiao, S., Hu, X. 2019. The influence of different knowledge-driven methods on landslide susceptibility mapping: A case study in the Changbai Mountain Area, Northeast China. *Entropy*, *21*(4), 372.

Mandal, K., Saha, S. and Mandal, S., 2021. Applying deep learning and benchmark machine learning algorithms for landslide susceptibility modelling in Rorachu river basin of Sikkim Himalaya, India. *Geoscience Frontiers*, *12*(5), p.101203.

Marjanović, M., Kovačević, M., Bajat, B., Voženílek, V. 2011. Landslide susceptibility assessment using SVM machine learning algorithm. *Eng. Geology*, 123(3), 225-234.

Meena, S. R., GudiyangadaNachappa, T. 2019. Impact of Spatial Resolution of Digital Elevation Model on Landslide Susceptibility Mapping: A case Study in Kullu Valley, Himalayas. *Geosciences*, *9*(8), 360.

Meena, S. R., Mishra, B. K., TavakkoliPiralilou, S. 2019a. A hybrid spatial multi-criteria evaluation method for mapping landslide susceptible areas in kullu valley, himalayas. *Geosciences*, *9*(4), 156.

Moayedi, H., Osouli, A., Tien Bui, D., Foong, L. K. 2019. Spatial landslide susceptibility assessment based on novel neural-metaheuristic geographic information system based ensembles. *Sensors*, *19*(21), 4698.

Mohan, A., Singh, A. K., Kumar, B., Dwivedi, R. 2020. Review on remote sensing methods for landslide detection using machine and deep learning. *Transactions on Emerging Telecom. Tech.*, e3998.

Moharrami, M., Naboureh, A., GudiyangadaNachappa, T., Ghorbanzadeh, O., Guan, X., Blaschke, T. 2020. National-scale landslide susceptibility mapping in Austria using fuzzy best-worst multi-criteria decision-making. *ISPRS Inter. J. Geo-Infor.*, *9*(6), 393.

Naidu, S., Sajinkumar, K. S., Oommen, T., Anuja, V. J., Samuel, R. A., Muraleedharan, C. 2018. Early warning system for shallow landslides using rainfall threshold and slope stability analysis. *Geosci. Front.*, *9*(6), 1871-1882.

Nepal, N., Chen, J., Chen, H., Sharma, T. P. P. 2019. Assessment of landslide susceptibility along the Araniko Highway in Poiqu/BhoteKoshi/Sun Koshi Watershed, Nepal Himalaya. *Progress in Dis.Sci.*, *3*, 100037.

Ngailo, E., von Rosen, D., Singull, M. 2020. Approximation of misclassification probabilities in linear discriminant analysis with repeated measurements.

Ngo, P. T. T., Panahi, M., Khosravi, K., Ghorbanzadeh, O., Kariminejad, N., Cerda, A., Lee, S. 2021. Evaluation of deep learning algorithms for national scale landslide susceptibility mapping of Iran. *Geosci. Front.*, *12*(2), 505-519.

Nilsson, C., Berggren, K. 2000. Alterations of riparian ecosystems caused by river regulation: Dam operations have caused global-scale ecological changes in riparian ecosystems. How to protect river environments and human needs of rivers remains one of the most important questions of our time. *BioScience*, *50*(9), 783-792.

Accepted Manuscript

Ozer, M. E., Sarica, P. O., Arga, K. Y. 2020. New machine learning applications to accelerate personalized medicine in breast cancer: Rise of the support vector machines. *OMICS: A J. of Integra. Bio.*, *24*(5), 241-246.

Pham, B. T., Shirzadi, A., Bui, D. T., Prakash, I., Dholakia, M. B. 2018. A hybrid machine learning ensemble approach based on a radial basis function neural network and rotation forest for landslide susceptibility modeling: A case study in the Himalayan area, India. *Int. J. Sed Res.*, *33*(2), 157-170.

Pourghasemi, H. R., Sadhasivam, N., Amiri, M., Eskandari, S., Santosh, M. 2021. Landslide susceptibility assessment and mapping using state-of-the art machine learning techniques. Natural Hazards, 108(1), 1291-1316

Pradhan, B. 2010. Remote sensing and GIS-based landslide hazard analysis and cross-validation using multivariate logistic regression model on three test areas in Malaysia. *Advan. in space res.*, *45*(10), 1244-1256.

Psomiadis, E., Charizopoulos, N., Efthimiou, N., Soulis, K. X., Charalampopoulos, I. 2020. Earth Observation and GIS-Based Analysis for Landslide Susceptibility and Risk Assessment. *ISPRS Int.J. Geo-Infor.*, *9*(9), 552.

Raduła, M. W., Szymura, T. H., Szymura, M. 2018. Topographic wetness index explains soil moisture better than bioindication with Ellenberg's indicator values. *Eco.Indica.*, *85*, 172-179.

Riley, J. W., Calhoun, D. L., Barichivich, W. J., Walls, S. C. 2017. Identifying small depressional wetlands and using a topographic position index to infer hydroperiod regimes for pond-breeding amphibians. *Wetlands*, *37*(2), 325-338.

Robotham, H., Castillo, J., Bosch, P., Perez-Kallens, J. 2011. A comparison of multi-class supportsvector machine and classification tree methods for hydroacoustic classification of fish-schools in Chile. *Fisheries res.*, *111*(3), 170-176.

Roy, J., Saha, S., Arabameri, A., Blaschke, T., & Bui, D. T. (2019). A novel ensemble approach for landslide susceptibility mapping (LSM) in Darjeeling and Kalimpong districts, West Bengal, India. *Remote Sensing*, *11*(23), 2866.

Różycka, M., Migoń, P., Michniewicz, A. 2017. Topographic Wetness Index and Terrain Ruggedness Index in geomorphic characterisation of landslide terrains, on examples from the Sudetes, SW Poland. *Zeitschriftfürgeomorphologie, Supplementary issues*, *61*(2), 61-80.

Saha, A., & Saha, S. (2022a). Integrating the artificial intelligence and hybrid machine learning algorithms for improving the accuracy of spatial prediction of landslide hazards in Kurseong Himalayan Region. *Artificial Intelligence in Geosciences*, *3*, 14-27.

Saha, A., Saha, S. 2020b. Comparing the efficiency of weight of evidence, support vector machine and their ensemble approaches in landslide susceptibility modelling: A study on Kurseong region of Darjeeling Himalaya, India. *Rem.Sen. Appl.: Society and Environ.*, 100323.

Saha, S., Arabameri, A., Saha, A., Blaschke, T., Ngo, P. T. T., Nhu, V. H., Band, S. S. 2021. Prediction of landslide susceptibility in Rudraprayag, India using novel ensemble of conditional probability and boosted regression tree-based on cross-validation method. *Sci. total environ.*, *764*, 142928.

Saha, S., Saha, A., Hembram, T. K., Mandal, K., Sarkar, R., & Bhardwaj, D. (2022b). Prediction of spatial landslide susceptibility applying the novel ensembles of CNN, GLM and random forest in the Indian Himalayan region. *Stochastic Environmental Research and Risk Assessment*, 1-20.

Saha, S., Saha, A., Hembram, T. K., Pradhan, B., Alamri, A. M. 2020a. Evaluating the performance of individual and novel ensemble of machine learning and statistical models for landslide susceptibility assessment at Rudraprayag District of Garhwal Himalaya. *Appl.Sci.*, *10*(11), 3772.

Accepted Manuscript

Sameen, M. I., Pradhan, B., & Lee, S. (2020). Application of convolutional neural networks featuring Bayesian optimization for landslide susceptibility assessment. *Catena*, *186*, 104249.

Saponaro, A., Pilz, M., Wieland, M., Bindi, D., Moldobekov, B., Parolai, S. 2015. Landslide susceptibility analysis in data-scarce regions: the case of Kyrgyzstan. *Bull. of Eng. Geology and the Envir.*, *74*(4), 1117-1136.

Shirani, K., Pasandi, M., Arabameri, A. 2018. Landslide susceptibility assessment by dempster–shafer and index of entropy models, Sarkhoun basin, southwestern Iran. *Natural Hazards*, *93*(3), 1379-1418.

Singh, K., Kumar, V. 2017. Landslide hazard mapping along national highway-154A in Himachal Pradesh, India using information value and frequency ratio. *Arabian J.of Geosci.*, *10*(24), 539.

Tang, Y., Feng, F., Guo, Z., Feng, W., Li, Z., Wang, J., Sun, Q., Ma, H. and Li, Y., 2020. Integrating principal component analysis with statistically-based models for analysis of causal factors and landslide susceptibility mapping: A comparative study from the loess plateau area in Shanxi (China). Journal of Cleaner Production, 277, p.124159.

Tehrany, M. S., Pradhan, B., & Jebur, M. N. (2014). Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. *Journal of hydrology*, *512*, 332-343.

Tien Bui, D., Pradhan, B., Lofman, O.; Revhaug, I.2012. Landslide susceptibility assessment in Vietnam using support vector machines, decision tree, and Naïve Bayes Models. Math. Probl. Eng., 1–26.

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1), 1-16.

Vahidnia, M. H., Alesheikh, A. A., Alimohammadi, A., Hosseinali, F. 2010. A GIS-based neuro-fuzzy procedure for integrating knowledge and data in landslide susceptibility mapping. *Comp. & Geoscis*, *36*(9), 1101-1114.

Van Dao, D., Jaafari, A., Bayat, M., Mafi-Gholami, D., Qi, C., Moayedi, H., Pham, B. T. 2020. A spatially explicit deep learning neural network model for the prediction of landslide susceptibility. *Catena*, *188*, 104451.

Van, N. T. H., Van Son, P., Khanh, N. H., Binh, L. T. 2016. Landslide susceptibility mapping by combining the analytical hierarchy process and weighted linear combination methods: a case study in the upper Lo River catchment (Vietnam). *Landslides*, *13*(5), 1285-1301.

von Rudorff, G. F., von Lilienfeld, O. A. 2020. Rapid and accurate molecular deprotonation energies from quantum alchemy. *Physical Chemistry Chemical Physics*, *22*(19), 10519-10525.

Wang, G., Lei, X., Chen, W., Shahabi, H., Shirzadi, A. 2020. Hybrid computational intelligence methods for landslide susceptibility mapping. *Symmetry*, *12*(3), 325.

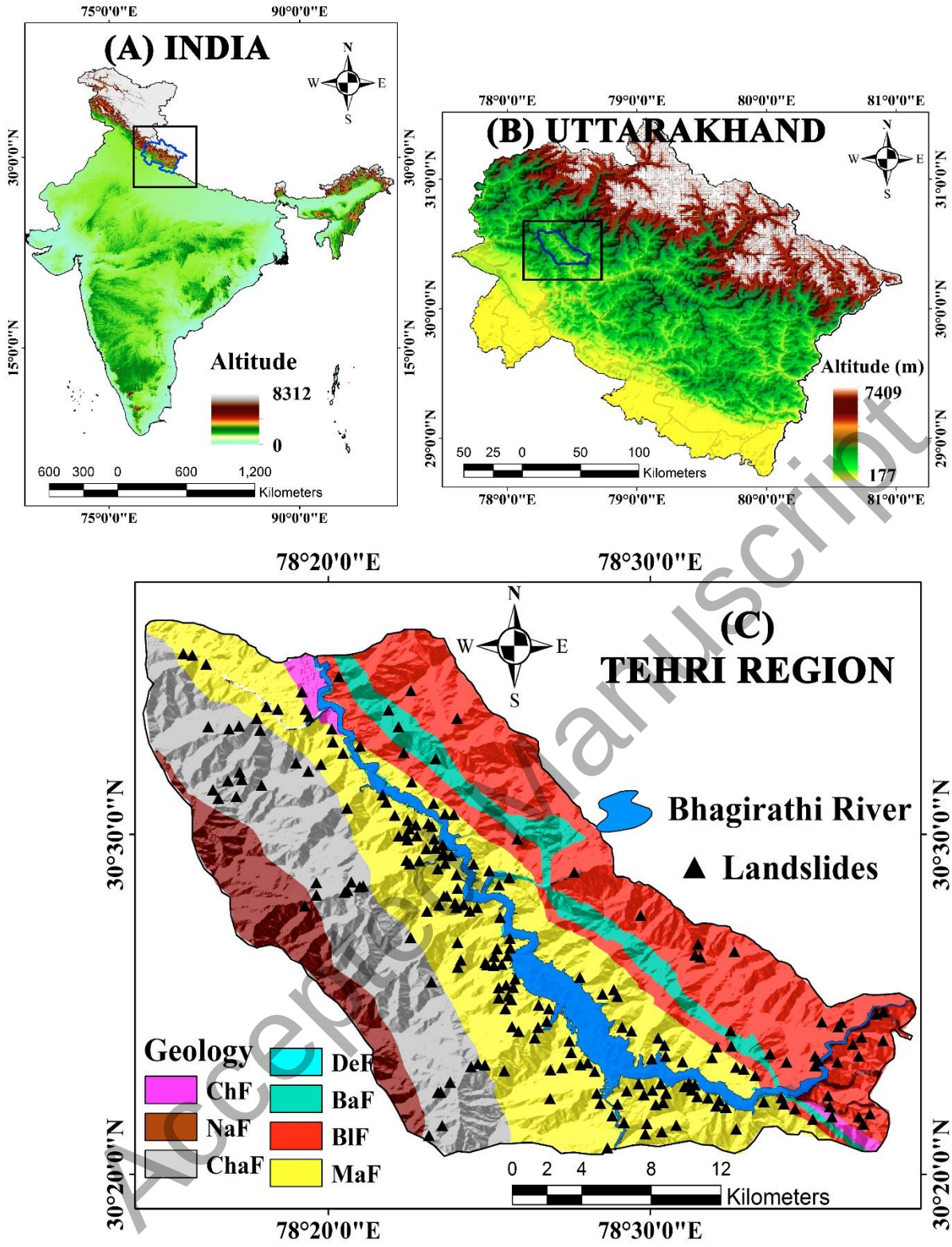
Xie, S., Liang, Y., Zheng, Z., Liu, H. 2017. Combined forecasting method of landslide deformation based on MEEMD, approximate entropy, and WLS-SVM. *ISPRS Int. J. Geo-Infor.*, *6*(1), 5.

Youssef, A. M., Pourghasemi, H. R. 2021. Landslide susceptibility mapping using machine learningalgorithms and comparison of their performance at Abha Basin, Asir Region, Saudi Arabia. *Geosci. Front.*, *12*(2), 639-655.

Zhao, X., Chen, W. 2020. Optimization of computational intelligence models for landslide susceptibility evaluation. *Rem.Sens.*, *12*(14), 2180.

Accepted Manuscript

## Figures




Accepted Manuscript

**Fig. 1.** Location map of the study area: A. India, B. Uttarakhand and C. Tehri region

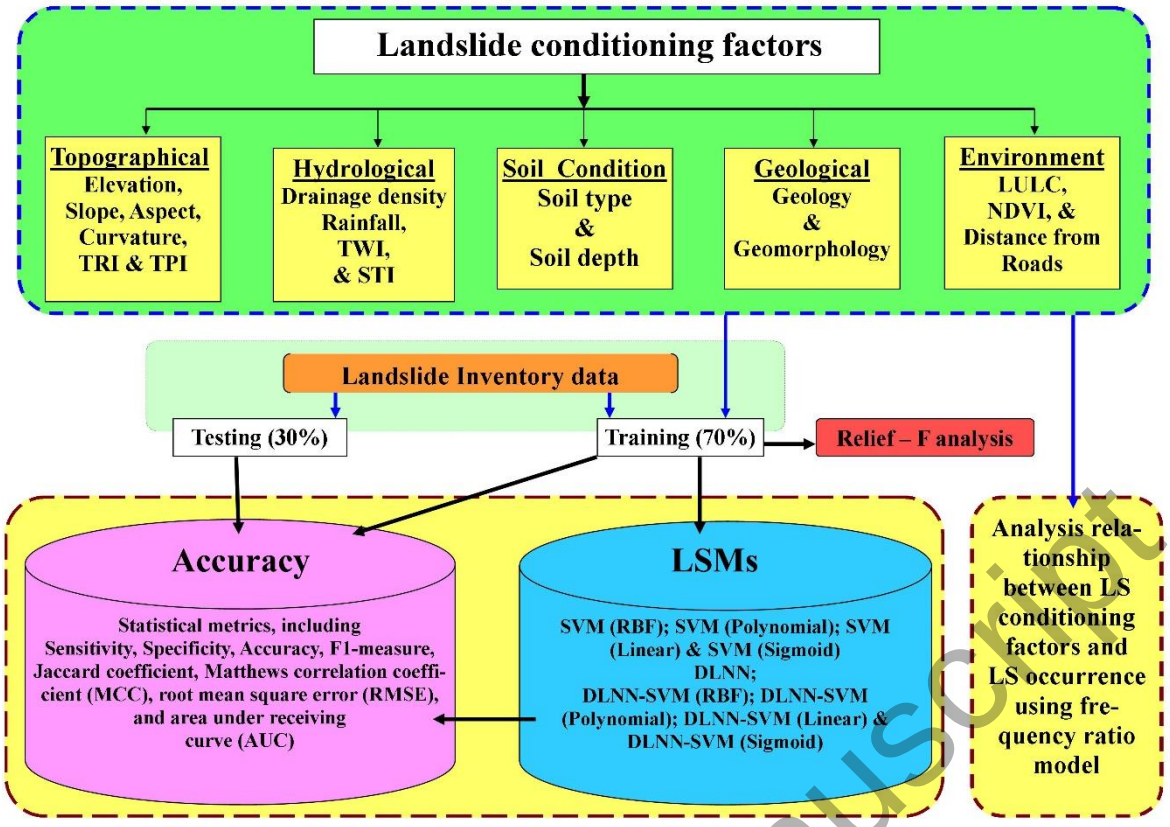



**Fig. 2.** Methodological flow chart of the present work

Accepted Manuscript

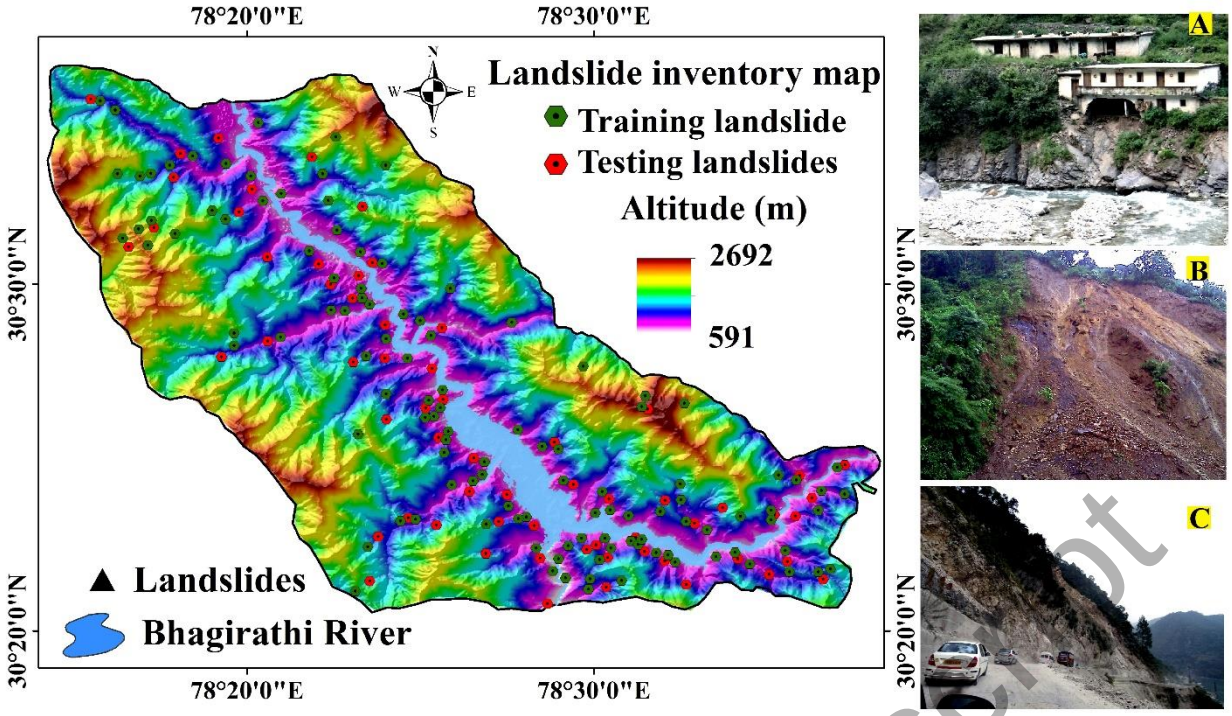
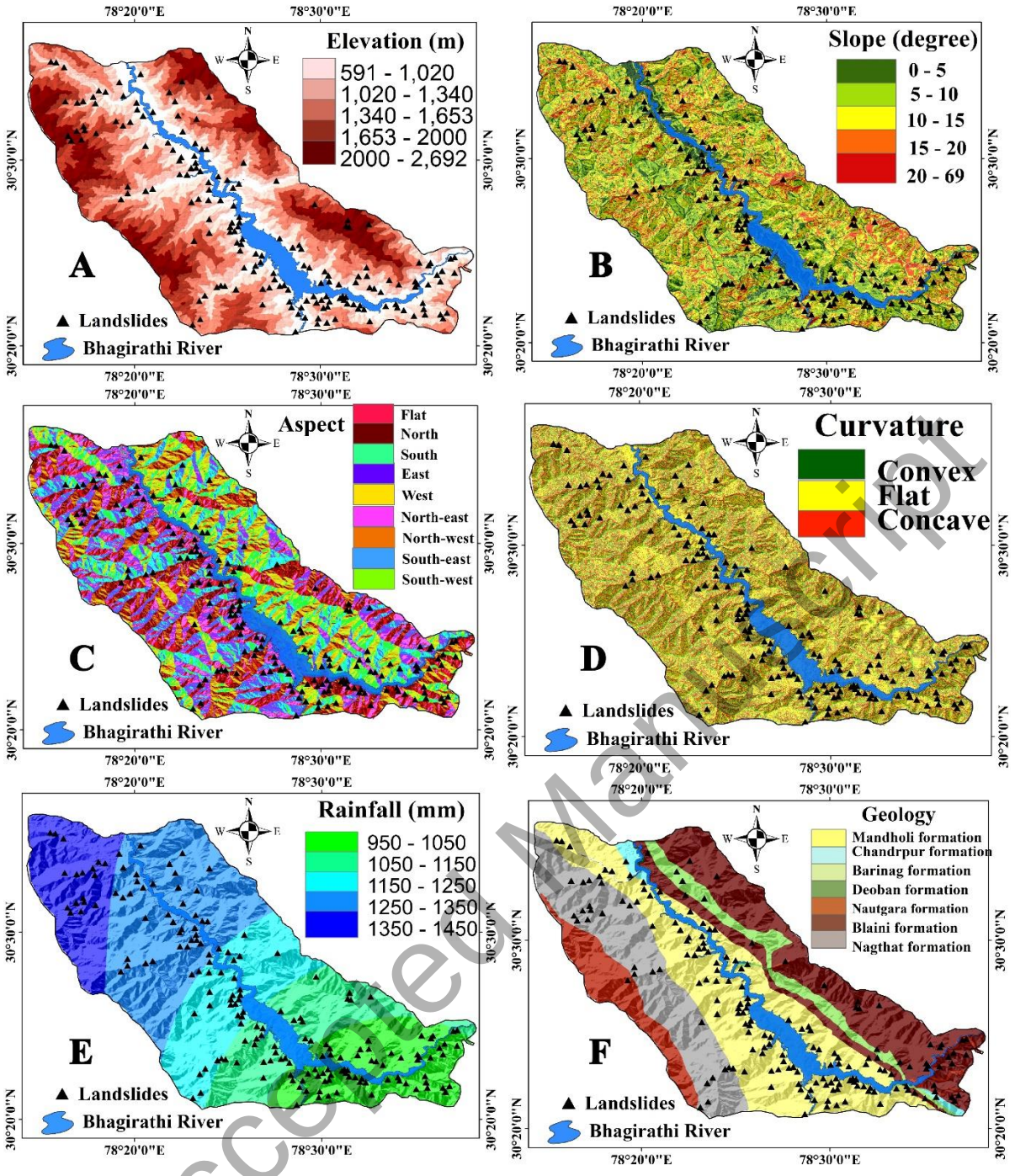
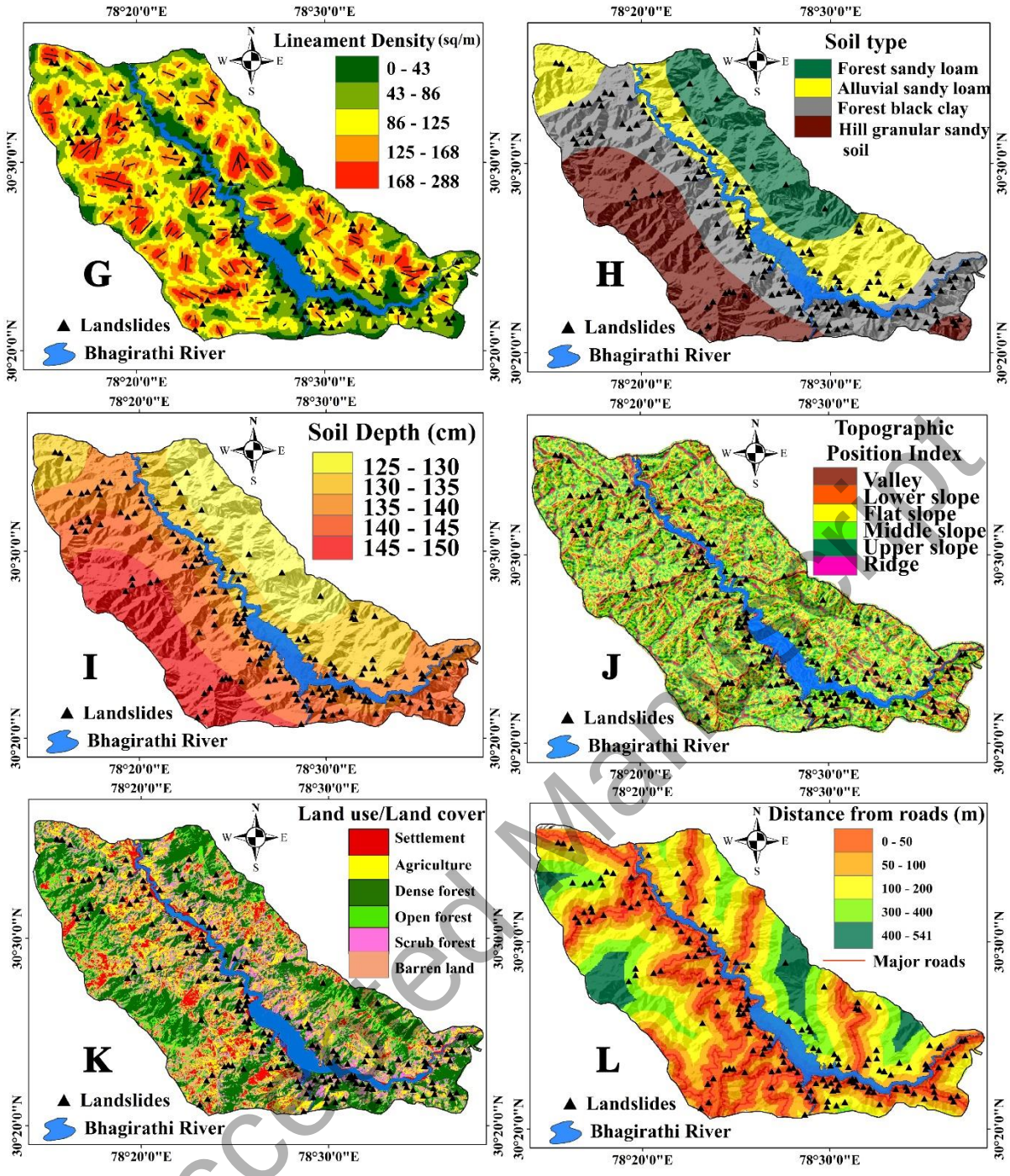



**Fig. 3**. The map of landslide inventory (LIM) with some landslide field photograph (A. a house destroyed in west bank of Bhagirathi river; B. a vegetation zone destroyed by landslide; C. Tehri-Garhwal road blocked due to landslide)

Accepted Manuscript




Accepted Manuscript




Accepted Manuscript

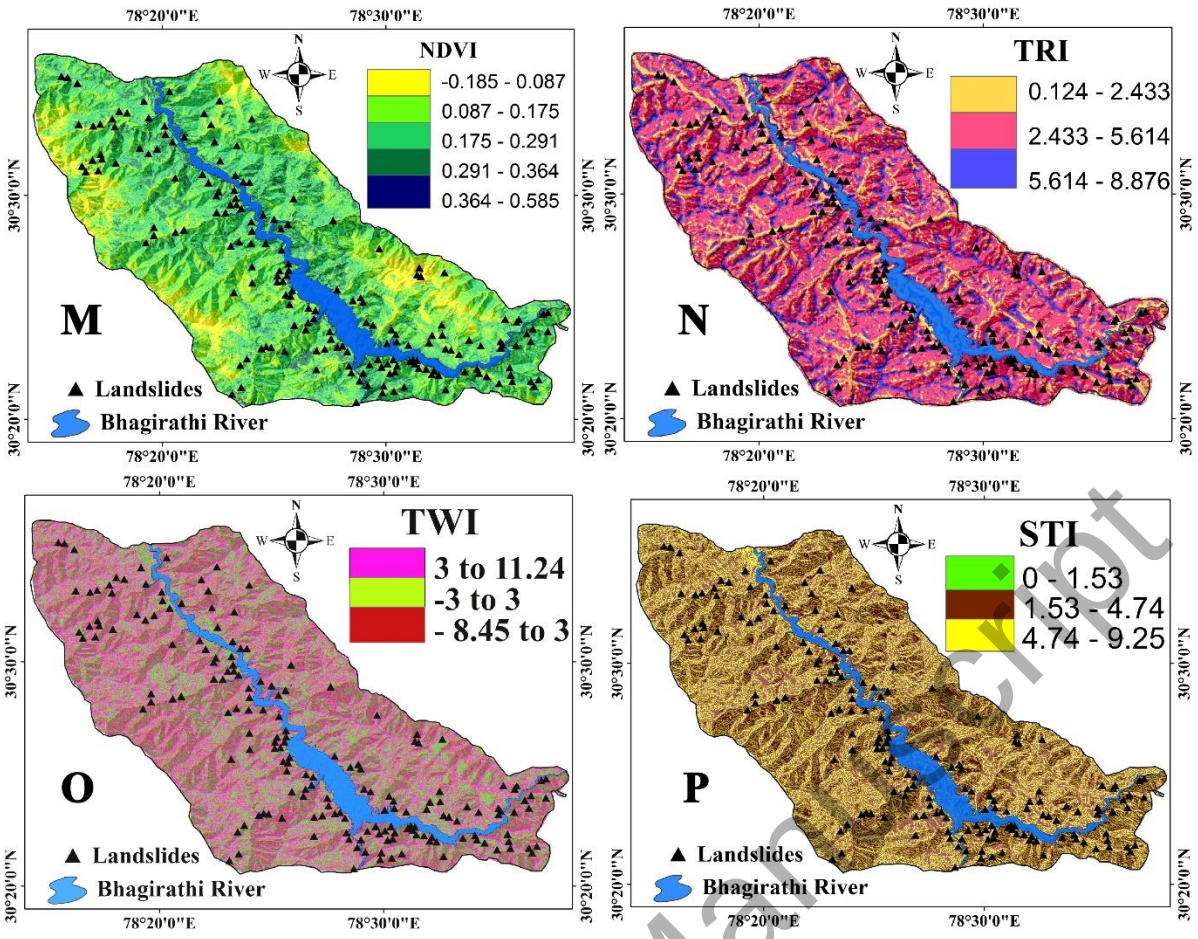
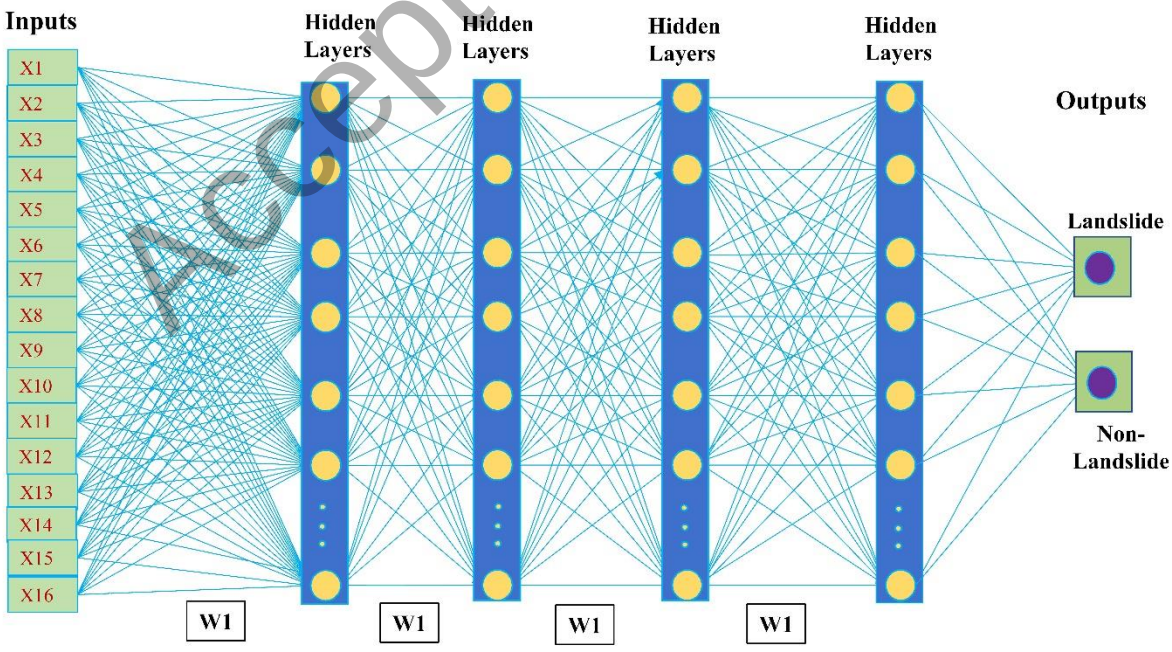



**Fig. 4.** Landslide conditioning factors used in the present study: A. Elevation, B. Slope, C. Aspect, D. Curvature E. Rainfall, F. Geology, G. Lineament density, H. Soil type, I. Soil depth, J. Topographical position index, K. Land use/land cover (LULC), L. Distance from road, M. NDVI, N. Topographic roughness index, O. Topographical wetness index (TWI) and P. Stream power index (STI)




Accepted Manuscript

**Fig. 5.** Theoretical structure of DLNN

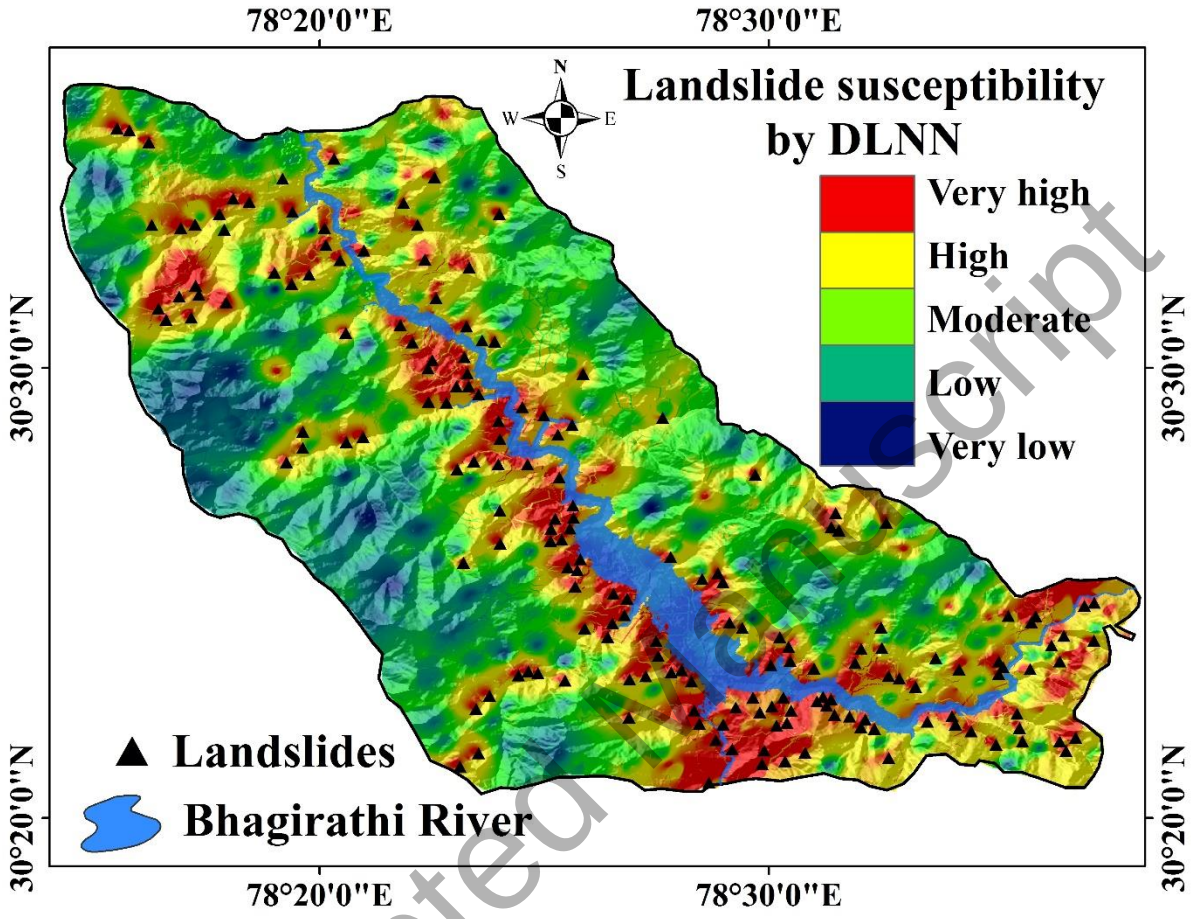



**Fig. 6.** Landslide susceptibility map prepared by DLNN

Accepted Manuscript

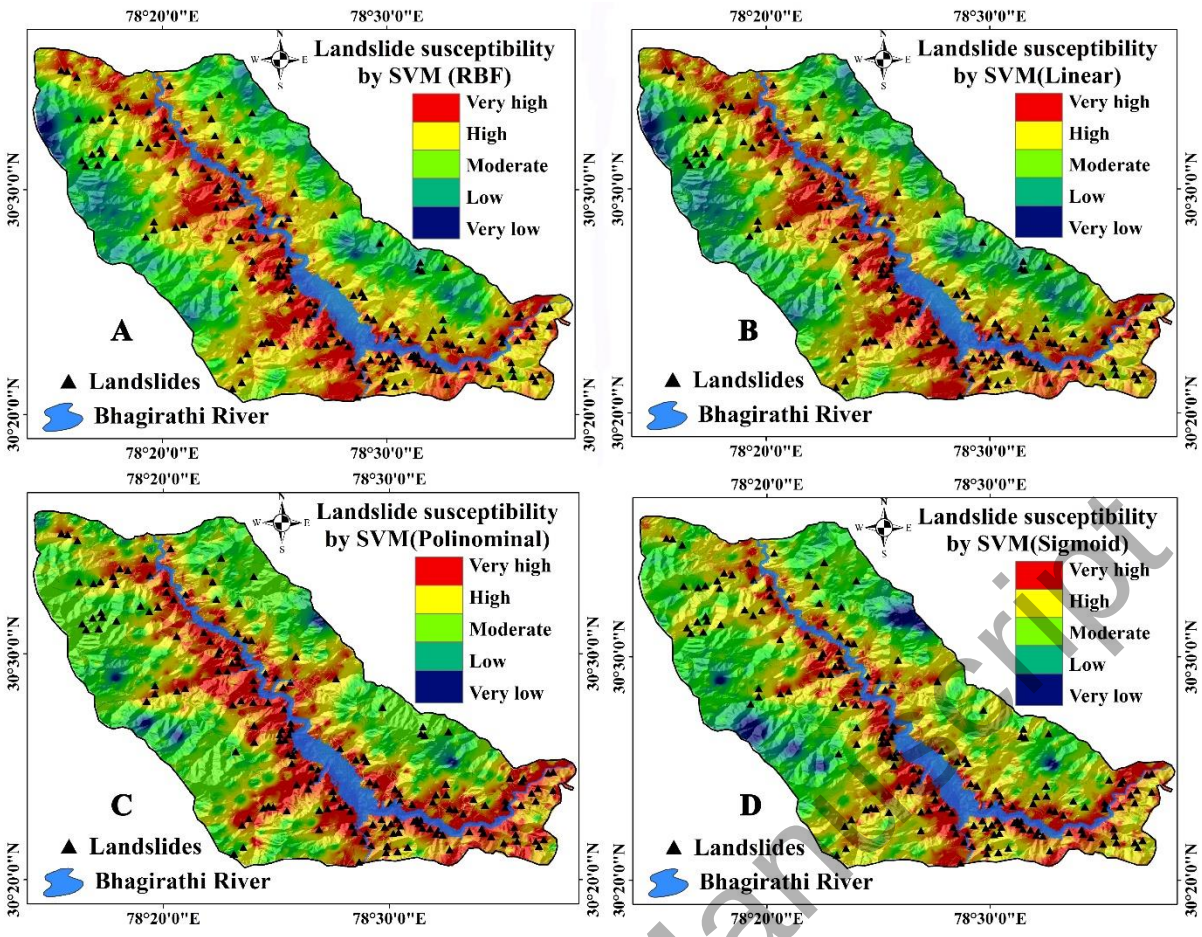



**Fig. 7.** Landslide susceptibility maps prepared by A. SVM (RBF), B. SVM (Linear), C. SVM (Polynomial) and D. SVM (Sigmoid)

Accepted Manuscript

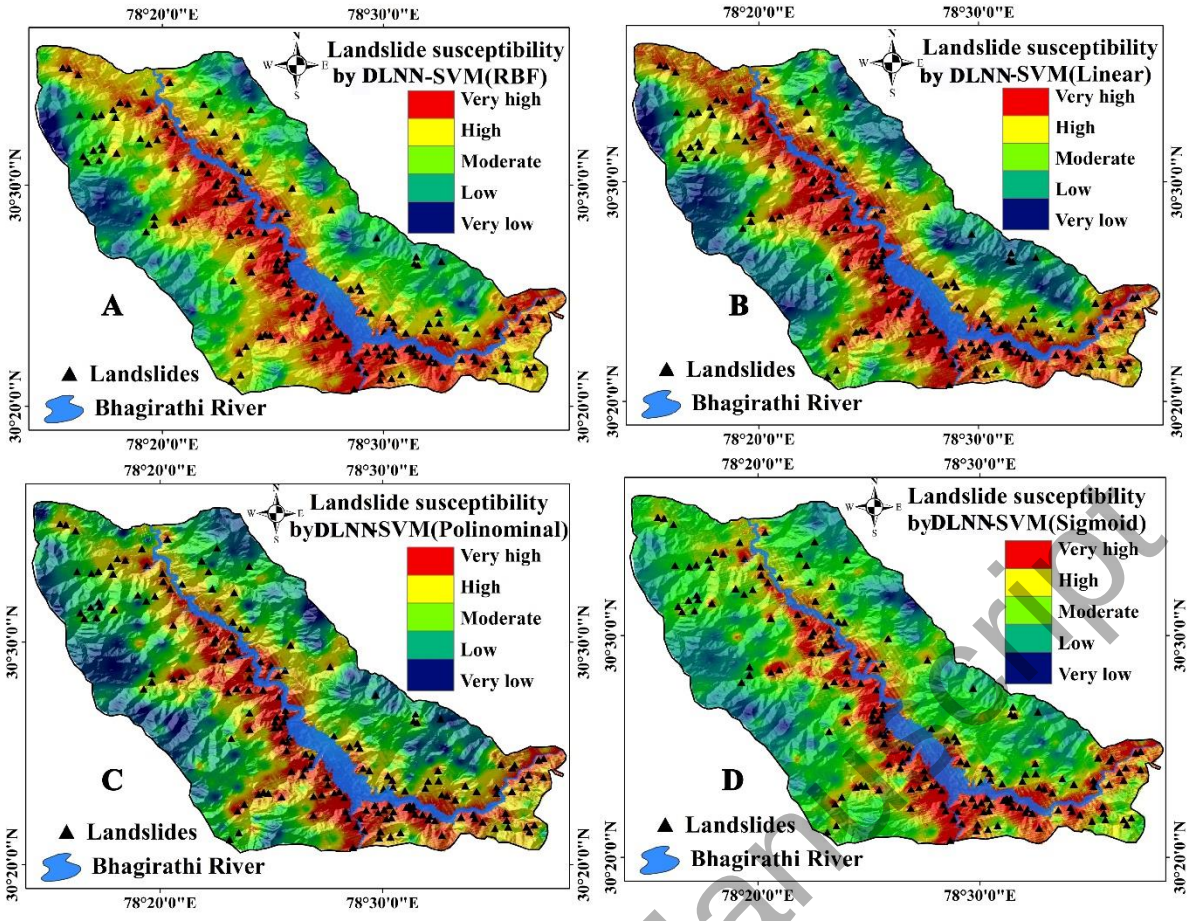



**Fig. 8.** Landslide susceptibility maps prepared by A. SVM (RBF)-DLNN, B. SVM (Linear)-DLNN, C. SVM (Polynomial)-DLNN and D. SVM (Sigmoid)-DLNN

Accepted Manuscript

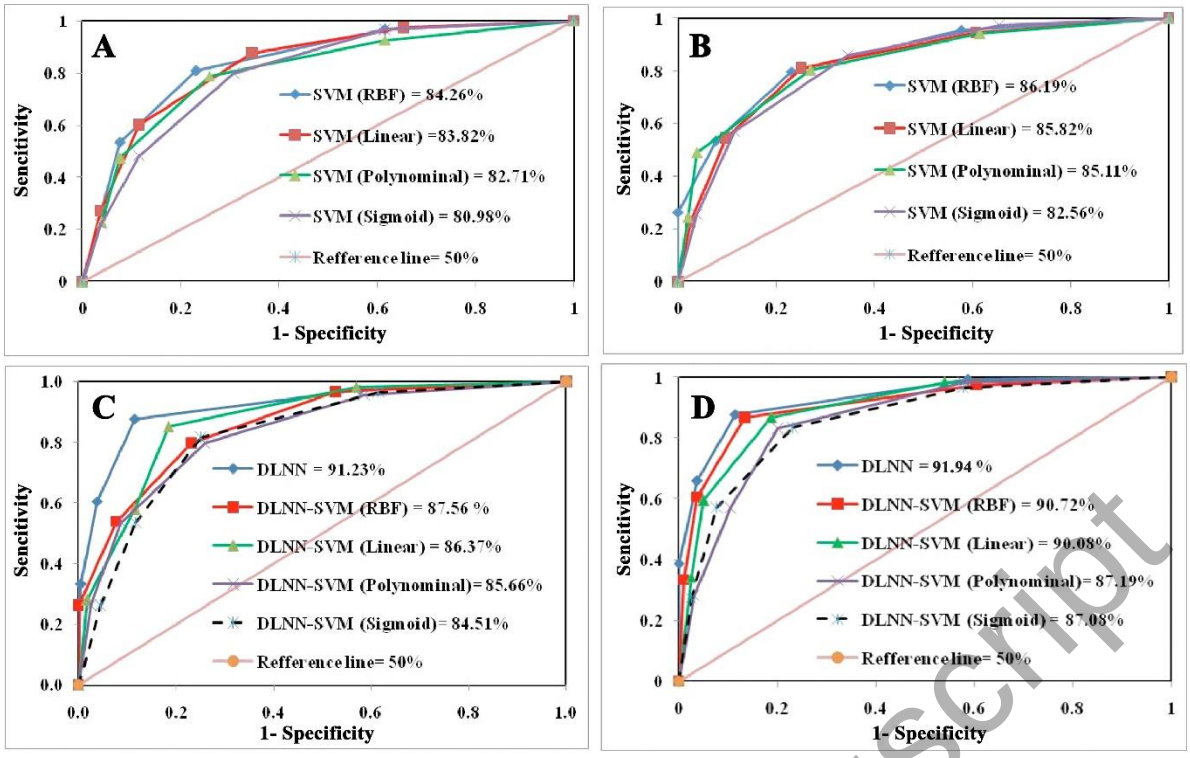



**Fig. 9.** Area under curve using ROC curve for different models

## Tables

**Table 1.** Data sources and properties

| Data | Layer | Source | Date of acquisition | Scale /resolution |
|---|---|---|---|---|
| Geological | Geological map | Geological Survey of India (GSI) | 2005-2006 | 1:250,000 |
| Rainfall map | Rainfall | https://www.worldclim.org/data/monthlywth.html | 1980-2018 | 30s (~1km$^2$) |
| Lineament distance | Lineament | https.//www.bhuvan.nrsc.gov.in | 2005-2006 | 1:50,000 |
| LULC, Settlement, | Landsat 8 | https.//www.Earthexplorer.usgs.gov | | 30 m × 30 m |
| River map, Road map | Topographical sheets (Open series) | Survey of India (SOI) | 2005-2006 | 1:50,000 |
| Altitude, Curvature, Slope, Aspect, TPI and TWI | SRTM DEM | https.//www.Earthexplorer.usgs.gov | 23/09/2014 | 30 m × 30 m |
| Soil texture | Soil map | NBSS & LUP Regional Centre | 2009 | 1:50,000 |
| Soil depth | Soil depth | Field survey | 2019 | 1:50,000 |

Accepted Manuscript

**Table 2.** Sub-classes of LCFs in the present study

| Factor | Sub-class/range | Methods |
|---|---|---|
| Elevation (m.) | 591- 1020; 1020- 1340; 1340-1653; 1653-2000; 2000-2692 | Natural breaks |
| Slope angle(degree) | 0-5; 5-10; 10-15; 15-20; 20-69. | Manual |
| Slope aspect | Flat, North, North-east, East, South-east, South, South-west, West, and North-west. | Azimuth |
| Curvature | Flat, Concave & Convex | Natural breaks |
| Rainfall (mm) | 950-1050; 1050-1150; 1150-1250; 1250-1350; 1350-1450 | Manual |
| Geology formation | Chandrpur, Barinage, Deoban, Nauthgara, Blaini, Nagthat, & Mandholi | Geology units |
| Topographical position index | Valley, Lower slope, Flat slope, Middle slope, Upper slope, & Ridge | Natural breaks |
| Fault density (square/m) | 0-43; 43-86; 86-125; 125-168; & 168-288 | Natural breaks |
| Soil type | Forest sandy loam, Alluvial sandy loam, Forest black clay, & Hill granular sandy soil | Soil type units |
| Soil depth (cm) | 125-130; 130-135; 135-140; 140-145 & 145-150 | Manual |

Accepted Manuscript

| | | |
|---|---|---|
| NDVI | -0.185 -0.087; 0.087- 0.175; 0.175-0.291 & 0.291-0.585 | Natural breaks |
| Land use/land cover | Settlement, Dense forest, Open forest, Scrub forest, Barren land, & Agriculture land | Land use units |
| Distance from roads (m) | 0-50, 50- 100, 100-200, 200-300,300-400, 400-541 | Manual |
| Topographic roughness index | 0.24-2.43; 2.43-5.61 & 5.61-8.87 | Natural breaks |
| Sediment transport index | 0-1.53; 1.53-4.74 & 4.74-9.25 | Natural breaks |
| Topographical Wetness index | -8.45 to -3; -3 to 3 & 3 to 11.24 | Natural breaks |

Accepted Manuscript

**Table 3.** Results of Relief-F tests of LCFs

| **Selected LS factors** | **Relief-F average Merit (AM)** |
|---|---|
| Elevation | 0.0397 |
| Geology | 0.031 |
| Soil type | 0.0216 |
| Soil depth | 0.0205 |
| Distance from roads | 0.0142 |
| Lineament density | 0.0139 |
| Slope | 0.0128 |
| Rainfall | 0.0127 |
| Curvature | 0.0125 |
| Land use/land cover (LU/LC) | 0.0125 |
| Normalized difference vegetation index (NDVI | 0.0093 |
| Aspect | 0.0056 |
| TPI | 0.0042 |
| TWI | 0.0025 |
| TRI | 0.0024 |

| | |
|---|---|
| STI | 0.0024 |

Accepted Manuscript

Table 4. Multicollinearity analysis of landslide conditioning factors used in this study

| Influencing Factor | VIF | TOL |
|---|---|---|
| Aspect | 1.099 | 0.618 |
| Curvature | 1.724 | 0.559 |
| Distance from roads | 5.387 | 0.202 |
| Elevation | 1.803 | 0.394 |
| Geology | 1.649 | 0.622 |
| Lineament density | 1.377 | 0.695 |
| LU/LC | 1.772 | 0.568 |
| NDVI | 1.271 | 0.563 |
| Rainfall | 2.341 | 0.442 |
| Slope | 2.676 | 0.381 |
| Soil depth | 1.748 | 0.581 |
| Soil type | 1.365 | 0.759 |
| STI | 1.248 | 0.574 |
| TPI | 1.464 | 0.589 |
| TRI | 1.244 | 0.586 |
| TWI | 1.242 | 0.593 |

***N.B:*** *VIF: Variance inflation factor; LU/LC; land use/land cover; NDVI: Normalised difference vegetation index; TPI: Topographical position index STI: Sediment transport index; TRI: Topographical ruggedness index; & TWI: topographical wetness index*

Accepted Manuscript

**Table 5.** Areal distribution of susceptibility classes produced by different models

| **Model** | **Classes** | **Area (%)** | **Area ($km^2$)** | **Model** | **Classes** | **Area (%)** | **Area ($km^2$)** |
|---|---|---|---|---|---|---|---|
| **SVM (RBF)** | Very High | 25.59 | 27.48 | **DLNN-SVM (RBF)** | Very High | 26.96 | 28.95 |
| | High | 12.11 | 13.00 | | High | 13.90 | 14.92 |
| | Moderate | 19.85 | 21.32 | | Moderate | 20.99 | 22.54 |
| | Low | 18.14 | 19.48 | | Low | 16.44 | 17.65 |
| | Very low | 24.31 | 26.11 | | Very low | 21.73 | 23.33 |
| **SVM (Linear)** | Very High | 17.87 | 19.19 | **DLNN-SVM (Linear)** | Very High | 22.41 | 24.07 |
| | High | 18.4 | 19.76 | | High | 16.15 | 17.34 |
| | Moderate | 32.22 | 34.60 | | Moderate | 26.60 | 28.57 |
| | Low | 10.85 | 11.65 | | Low | 13.65 | 14.65 |
| | Very low | 20.09 | 21.57 | | Very low | 20.91 | 22.45 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **SVM (Polynomial)** | Very High | 17.33 | 18.61 | **DLNN-SVM (Polynomial)** | Very High | 19.87 | 21.34 |
| | High | 19.99 | 21.47 | | High | 18.07 | 19.40 |
| | Moderate | 36.18 | 38.85 | | Moderate | 31.39 | 33.71 |
| | Low | 13.84 | 14.86 | | Low | 13.74 | 14.76 |
| | Very low | 12.66 | 13.60 | | Very low | 16.78 | 18.02 |
| **SVM (Sigmoid)** | Very High | 18.77 | 20.16 | **DLNN-SVM (Sigmoid)** | Very High | 19.32 | 20.75 |
| | High | 19.21 | 20.63 | | High | 18.64 | 20.02 |
| | Moderate | 27.14 | 29.15 | | Moderate | 29.27 | 31.43 |
| | Low | 18.46 | 19.82 | | Low | 16.10 | 17.29 |
| | Very low | 16.42 | 17.63 | | Very low | 16.60 | 17.83 |
| **DLNN** | Very High | 28.32 | 30.41 | | | | |
| | High | 15.68 | 16.84 | | | | |
| | Moderate | 22.12 | 23.75 | | | | |
| | Low | 14.74 | 15.83 | | | | |
| | Very low | 19.14 | 20.55 | | | | |

**Table 6.** Results of frequency ratio analysis of LCFs

| Factor | Subclass | FR | Factor | Subclass | FR | Factor | Subclass | FR |
|---|---|---|---|---|---|---|---|---|
| **Elevation (m)** | 591 - 1020 | 0.0527 | **Geology** | Deoban formation | 0.0084 | **LU/LC** | River | 0.0039 |
| | 1020 -1340 | 0.0137 | | Chadpur formation | 0.0001 | | Forest | 0.0071 |
| | 1340 -1653 | 0.0057 | | Barinag formation | 0.0094 | | Barren Land | 0.0201 |
| | 1653 -2000 | 0.0083 | | Nautgara formation | 0.0006 | | Scrub | 0.0150 |
| | 2000 -2692 | 0.0096 | | Blaini formation | 0.0055 | | Agricultural Land | 0.0157 |
| **Slope (degree)** | 0 to 5 | 0.0113 | | Nagthal formation | 0.0125 | | Settlement | 0.0149 |
| | 5 to10 | 0.0108 | | Mondholi formation | 0.0254 | **Road distance (m)** | 0-50 | 0.0055 |
| | 10 to 15 | 0.0129 | **Lineament density (sq/m)** | 0-43 | 0.0203 | | 50-100 | 0.0116 |
| | 15 to 20 | 0.0153 | | 43-86 | 0.0163 | | 100-200 | 0.0186 |
| | 20 to 69 | 0.0171 | | 86-125 | 0.0132 | | 200-400 | 0.0108 |
| **Aspect** | Flat | 0.0074 | | 125-168 | 0.0073 | | 400-541 | 0.0002 |
| | East | 0.0112 | | 168-288 | 0.0062 | **NDVI** | -0.185 to 0.087 | 0.0055 |

Accepted Manuscript

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | North | 0.0138 | **Soil type** | Forest sandy loam | 0.0046 | | 0.087 to 0.175 | 0.0116 |
| | Northeast | 0.0178 | | Alluvial sandy loam | 0.0179 | | 0.175 to 0.291 | 0.0186 |
| | northwest | 0.0229 | | Forest black loam | 0.0309 | | 0.291 to 0.364 | 0.0108 |
| | south | 0.0179 | | Hill granular sandy soil | 0.0083 | | 0.364 to 0.585 | 0.0009 |
| | southeast | 0.0111 | **Soil depth (cm)** | 125-130 | 0.0018 | **TRI** | 0.124 to 2.433 | 0.0047 |
| | southwest | 0.0076 | | 130-135 | 0.0034 | | 2.433 to 5.614 | 0.0082 |
| | west | 0.0063 | | 135-140 | 0.0104 | | 5.614 to 8.876 | 0.0389 |
| **Curvature** | Convex | 0.0132 | | 140-145 | 0.0122 | **TWI** | 3 to 11.24 | 0.0851 |
| | Flat | 0.0127 | | 145-150 | 0.0178 | | -3 to 3 | 0.0344 |
| | Concave | 0.0147 | **TPI** | Valley | 0.0075 | | -8.52 to -3 | 0.0113 |
| **Rainfall (mm)** | 950-1050 | 0.0202 | | Lower slope | 0.0089 | **STI** | 0 to 1.53 | 0.0061 |
| | 1050-1150 | 0.0142 | | Flat slope | 0.0064 | | 1.53 to 4.74 | 0.0084 |
| | 1150-1250 | 0.0122 | | Middle slope | 0.0119 | | 4.73 to 9.25 | 0.0129 |
| | 1250-1350 | 0.0135 | | Upper slope | 0.0207 | | | |
| | 1350-1450 | 0.0166 | | Ridge | 0.0131 | | | |

**Table 7.** Results of different accuracy measures

| Models | Datasets | Sensitivity | Specificity | Accuracy (%) | F1 measure | Jaccard | MCC | MAE | RMSE |
|---|---|---|---|---|---|---|---|---|---|
| **SVM (RBF)** | Training | 0.951 | 0.827 | 84.26 | 0.895 | 0.901 | 0.898 | 0.038 | 0.079 |
| | Testing | 0.956 | 0.896 | 86.17 | 0.912 | 0.906 | 0.912 | 0.088 | 0.167 |
| **SVM (Linear)** | Training | 0.943 | 0.877 | 83.82 | 0.908 | 0.878 | 0.901 | 0.042 | 0.137 |
| | Testing | 0.951 | 0.865 | 85.82 | 0.910 | 0.884 | 0.908 | 0.055 | 0.196 |
| **SVM** | Training | 0.866 | 0.798 | 82.71 | 0.892 | 0.887 | 0.872 | 0.139 | 0.286 |
| **(Polynomina)** | Testing | 0.906 | 0.715 | 85.11 | 0.898 | 0.835 | 0.868 | 0.144 | 0.196 |
| **SVM** | Training | 0.865 | 0.798 | 80.98 | 0.882 | 0.881 | 0.865 | 0.126 | 0.294 |
| **(Sigmoid)** | Testing | 0.874 | 0.684 | 82.56 | 0.875 | 0.829 | 0.845 | 0.124 | 0.287 |
| **DLNN** | Training | 0.976 | 0.91 | 91.23 | 0.920 | 0.950 | 0.914 | 0.115 | 0.108 |
| | Testing | 0.984 | 0.872 | 91.94 | 0.954 | 0.922 | 0.939 | 0.037 | 0.039 |
| **DLNN-** | Training | 0.956 | 0.798 | 87.56 | 0.911 | 0.926 | 0.911 | 0.064 | 0.141 |
| **SVM(RBF)** | Testing | 0.965 | 0.807 | 90.72 | 0.954 | 0.889 | 0.884 | 0.036 | 0.061 |
| **DLNN-** | Training | 0.940 | 0.881 | 86.37 | 0.938 | 0.911 | 0.937 | 0.049 | 0.211 |
| **SVM(Linear)** | Testing | 0.882 | 0.901 | 90.08 | 0.956 | 0.939 | 0.884 | 0.112 | 0.245 |

Accepted Manuscript

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **DLNN-SVM** | Training | 0.926 | 0.785 | 85.66 | 0.938 | 0.909 | 0.809 | 0.172 | 0.323 |
| **(Polynomina)** | Testing | 0.893 | 0.784 | 87.19 | 0.935 | 0.875 | 0.884 | 0.163 | 0.275 |
| **DLNN-SVM** | Training | 0.889 | 0.744 | 84.51 | 0.938 | 0.909 | 0.847 | 0.175 | 0.268 |
| **(Sigmoid)** | Testing | 0.876 | 0.712 | 87.08 | 0.935 | 0.875 | 0.884 | 0.191 | 0.328 |

Accepted Manuscript

Carbon Trends 9 (2022) 100209

Contents lists available at ScienceDirect

Carbon Trends

journal homepage: www.elsevier.com/locate/cartre

# N-type diamane: An effective emitter layer in crystalline silicon heterojunction solar cell

Naima, Pawan K. Tyagi*, Vinod Singh

*Department of Applied Physics, Delhi Technological University, New Delhi 110042, India*

ARTICLE INFO

*Article history:*
Received 20 June 2022
Revised 9 August 2022
Accepted 27 August 2022

*Keywords:*
2D material
Diamane
Silicon heterojunction solar cell
AFORS-HET v 2.5 software

ABSTRACT

In the presented work, cell's parameter of heterostructure silicon solar cell modelled as: ITO (front contact)/n-Dn/p-cSi/Ag under the illumination of monochromatic light at standard spectrum AM-1.5 G, has been optimized by using AFORS-HET v 2.5 software. We have used n-type diamane as an emitter layer and the nature of diamane has been considered 3D in spite of 2D as well as electronic nature of diamane is considered isotropic. To ensure that Schottky junction has been formed at interface, the electric contacts have been made along the c-axis so that maximum charge carriers get collected. To obtain the high efficiency, various parameters of n-type diamane as well as p-type c-Si layers have been optimized and the maximum efficiency of 16.84% has been obtained for single layer n-diamane at 100 μm thick silicon wafer. We also investigated the spectral response and dependency of temperature on the performance of exclusively designed solar cell and we obtained the best efficiency 16.84% at 300 K temperature. In order to check the performance on commercially available silicon we have optimized the same solar cell by considering the parameters of commercially available p-type crystalline silicon layer and maximum efficiency 10.41% was achieved. After getting the maximum efficiency 16.84% we further carried out the simulation by optimizing the layer numbers of n-diamane and found decrement in the efficiency up to 15.3% which indicates that, efficiency slightly decreases as layer number increases. We have demonstrated that n-type diamane might be used as an effective emitter layer in crystalline Si heterojunction solar cell.

© 2022 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

## 1. Introduction

Recently, a new allotrope of the carbon named as "Diamane" (Dn) possess $sp^3$ carbon bonding has been synthesized and reported to be two-dimensional (2D) being a physical analogue to "graphene-graphane" [1–4]. Recently, Bakharev et al. [5] confirmed that diamane exhibited 2 D electronic system. It was confirmed by estimating the optical band gap of diamane adopting the approach described in previous studies [6,7] for 2D electronic systems. They found that the experimental optical gap value of 3.3–3.4 eV is consistent with the calculated optical gap of 2.87 eV obtained for F-diamane. Furthermore, Raman spectroscopy was used to confirm the successful synthesis of 2D diamane and the characterization of its electronic structure has also been reported [8,9]. It has thermal conductivity of 1960 W/mK as high as to its bulk counterpart diamond (1000 W/mK), due to which it is considered as promising material for the high temperature electronic applications [1,10]. Furthermore, density functional theory (DFT) calculations reveal that 2D diamane has smaller effective mass ($0.55m_0$) of conduction electrons in comparison of its bulk counterpart i.e., diamond ($0.57m_0$), which is also very promising to use 2D diamane in various electronic applications such as solar cells [1]. 2D materials are best to use in electronic applications as they have large tunability in their bandgaps and they are flexible, strong and extremely thin. It was first predicted by Chernozatonskii et al., in 2009 [11], although its experimental synthesis was reported recently. The detailed verification of its bonding and electronic structure has also been probed by using various characterization techniques like: X-ray photoelectron spectroscopy (XPS), UV photoelectron, Raman, UV–Vis and electron energy loss spectroscopies, transmission electron microscopy and X-ray Diffraction (XRD) [5]. Formerly, the instability of diamane was marked as a big hurdle for its above-mentioned potential applications. Recently, the stability of diamane has been improved by the substitution of N atoms on its surface. The N atom has one more electron than C atom that form lone pair of electrons and behaves as the passivation, which stabilizes the structure and extends the band gap [2]. The first principal method has investigated the configuration

* Corresponding author.
*E-mail address:* pawantyagi@dtu.ac.in (P.K. Tyagi).

https://doi.org/10.1016/j.cartre.2022.100209
2667-0569/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

stability and electronic properties of N-doped diamane [2]. This novel 2D material has direct band gaps which can be controlled in the range from 0.86 to 5.68 eV [4]. Furthermore, due to its large bandgap, the intrinsic carrier concentration stays low even at the high temperature. Furthermore, the substitution of N atoms makes these films highly n-type, and it is found that N-substituted diamanes are harder than fully hydrogenated diamanes [2]. Atomistic simulation demonstrated that diamane carries giant in plane stiffness (715 N/m) as compare of graphene (238 N/m) and graphane (449 N/m) [1,11]. It is noteworthy to mentioned that diamane possess high carriers' mobility, for electrons (2732 $cm^2$/Vs) and for holes (1565 $cm^2$/Vs) in case of fluorinated and hydrogenated diamane, which is superior over those of III-IV semiconductor compounds [12] and reported in 2 D materials. The calculated band gap for fluorinated diamane ($C_4F$) (2.93 eV) confirms its high transparency [13]. It has to be noticed that the work function of 2D diamane can also be turned in the range of 7.24 eV to 0.62 eV after vacuum annealing at 700 °C and the work function of as-grown bilayer graphene was reported to be 4.4 eV [14,5] which also opens up its use in various electronic applications. Above-mentioned unique properties of ultrathin 2D diamane film motivate us to explore the potential application of n-type diamane as an effective emitter layer in Si-based heterojunction photovoltaic cells. The role of emitter layer in the crystalline silicon heterojunction solar cell is not only to allow the maximum photons to surpass across the interface but also to develop a built-in potential at diamane/silicon interface which require to discrete the photogenerated charge carriers. In recent time, diamane is expected to be used in various areas such as nano-optics, nano-electronics, nano-electromechanical systems, mechanical resonators and electronic devices at high temperature [5]. In spite of that, the application of diamane in the solar cell has not been much investigated yet. This also inspired us to use n-type diamane as an emitter layer in the silicon heterojunction solar cell. As diamane layer number increases from one to five, there was a small decrement in the efficiency from 8 to 7.53%. Since, this small efficiency was not worthy to obtain an efficient solar cell therefore, it had become more necessary to optimize the values of parameters of p-cSi as well so that an affordable solar cell with less manufacturing cost can be achieved with better efficiency. Thus, in the present article, solar cell modeled as ITO (front contact)/n-Dn/p-cSi/Ag was optimized under the illumination of monochromatic light with a power density of 100 W/$cm^2$ and spectral width of 10 nm by using latest version 2.5 of the software AFORS-HET. This software generally solves the 1-D semiconductor equations such as continuity equation, transport and Poisson's equation for both holes and electrons by using finite difference at different conditions. In this simulated cell, the n-diamane has been chosen as an emitter layer or conducting electrode while, the p-cSi as an absorbing layer. The variables of both n-diamane and p-cSi layer were optimized to achieve the notable efficiency for n-diamane/p-cSi solar cell. The simulation was performed by considering the isotropic elastic nature of diamane layer. Since, diamane is isotropic in nature therefore the efficiency will remain same for both armchair and zig-zag direction in diamane layer. The purpose of this simulation is to analysis the impact of isotropic nature of diamane and silicon wafer on the functioning of solar cell and to enlighten the cell response.

## 2. Device structure and simulation details

The proposed structure of solar cell is shown in Fig. (1). To design this structure, we have used current version 2.5 of AFORS-HET (Automat FOR Simulation of Heterojunction) software which uses Shockley-Read-Hall recombination statistics for different layers of solar cell [15]. In the proposed structure of solar cell, ITO(TCO)/n-diamane/p-crystalline Silicon/Ag (back contact) displayed in Fig. (1). We have taken ITO as transparent conducting oxide (TCO) as well as anti-reflector to ensure that maximum number of photons get transmit. So, as the light incident on the cell photons get absorb in the absorber layer, as photon absorption increased in the active layer more electron and hole pairs will be generated. In this way, to achieve the maximum electron-hole pairs generation in the absorber layer we intend to use the material of highest transmission efficiency and to fulfill this purpose we used ITO as TCO. ITO can directly be deposited on diamane with the help of electron beam and pulsed layer deposition (PLD) technique transferred on $SiO_2$/Si pattern substrate [16–19]. Microstructure, surface morphology, electrical and optical properties of ITO film can be characterized and analysed by using various techniques like: XRD, scanning electron microscopy, atomic force microscopy, high resolution transmission electron microscopy. The standard transmission line model (TLM) method was used to evaluate sheet resistance ITO/n-Si contacts in HJ silicon solar cells [20]. This deposition technique might not be harmed the diamane as well as $SiO_2$/Si substrate. Ag was chosen as the back contact as shown in Fig. (1). The [n, k] file for front contact as well as back contact was taken as default which are in-built in the software. All the details about the parameters of front and back contacts are tabulated in Table 1. The proposed structure was illuminated under Air Mass 1.5 G, spectral width 10 nm with the power density of 100 mW/$cm^2$ as a light source. Simultaneously, the temperature of the cell was maintained at 300 K. Since, we prefer to take the reflection coefficient of front contact as smaller and the reflection coefficient of back contact as higher for getting higher efficiency. The front contact should be transparent ideally, so that maximum light can pass through the surface and get collected at the absorbing layer. Therefore, in this simulation we have taken the reflection coefficient of front contact as nearly zero whereas, the reflection coefficient of the back contact was taken as 0.05. The generation of electron-hole pairs was described by Lambert-Beer model which is in-built in the software and the reflection from the diamane/Si interface is not considered [15]. Shockley-Read-Hall recombination and/or dangling-bond recombination was considered as recombination model within the software including some defect states within the bandgap. ITO/n-Dn interface is considered as MS-Schottky front interface and the back interface i.e., p-cSi/Ag is also considered as MS-Schottky. To interpret the transportation of charge carriers in diamane layer, drift and diffusion model was applied in the software, where the differential transport equation for both holes and electrons was taken as 2-D in nature with carrier densities in A/cm [15]. At diamane/Si interface, two-dimensional equations are coupled with three-dimensional equations (as we know diamane is 2D and c-Silicon is 3D in nature) to act for the transport and boundary conditions at the interface. In this simulation, we have not assumed any mismatching of lattices between the layers and the movement of minority charge carriers has been considered through the c-axis. Ideally, diamane is 2-dimensional, only if this fulfils the condition $< t > \leq \lambda_F$, where <t> is average thickness of diamane, $\lambda_F$ is the Fermi wavelength

**Table 1**
Information of the parameters used as front and back contact in the designed solar cell.

| Contact variables | Front contact | Back contact |
|---|---|---|
| Material | ITO | Ag |
| Width (m) | $8 \times 10^{-8}$ | $1 \times 10^{-4}$ |
| File | ITO.nk | Ag. nk |
| Metal work function | Yes (flat band) | Yes (flat band) |
| Absorption loss | ITO. Abs | 0 |
| External reflection constant | 0 | 0.05 |
| Surface condition | Plane | Plane |
| Internal reflection constant | 0 | 0 |

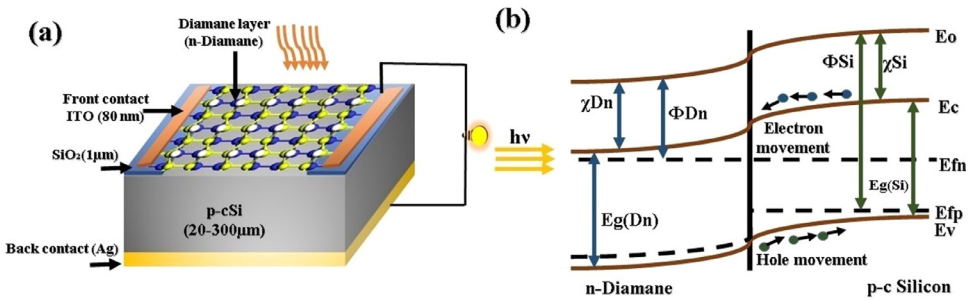



**Fig. 1.** (a) Schematic diagram of the simulated structure, (b) Schematic diagram under forward biasing just after the formation of heterojunction showing the transportation of electrons and holes. $\Phi_{Dn}$, $\Phi_{Si}$ and $\chi_{Dn}$, $\chi_{Si}$ are the diamane and Si work function and the diamane and silicon affinity, respectively.

**Table 2**
Ranges of other parameters chosen for the simulation.

| Input parameters | Diamane | p-cSi |
|---|---|---|
| Thickness (cm) | $0.3 \times 10^{-7}$ – $1.5 \times 10^{-7}$ | 0.002–0.03 |
| Dielectric constant ($\varepsilon_r$) | 4–11 | 11.9 |
| Electron affinity $\chi$ (eV) | 3.3–5.2 | 3.6–4.5 |
| Band gap, $E_g$ (eV) | 0.8–5.6 | 1.12 |
| Optical band gap, $E_g$ (eV) | 2.8 | 1.124 |
| Conduction band density, $N_C$ ($cm^{-3}$) | $1 \times 10^{19}$–$1 \times 10^{21}$ | $3 \times 10^{19}$–$1 \times 10^{21}$ |
| Valence band density, $N_V$ ($cm^{-3}$) | $1 \times 10^{19}$–$1 \times 10^{21}$ | $3 \times 10^{19}$–$1 \times 10^{21}$ |
| Electron mobility, $\mu_n$ ($cm^2$/Vs) | 2732 | 1500 |
| Hole mobility, $\mu_h$ ($cm^2$/Vs) | 743–1565 | 500 |
| Acceptor concentration $N_A$ ($cm^{-3}$) | 0 | $1 \times 10^{16}$ – $7 \times 10^{18}$ |
| Donor concentration $N_D$ ($cm^{-3}$) | $3 \times 10^{14}$ – $1 \times 10^{18}$ | 0 |
| Electron thermal velocity, $V_e$ ($cms^{-1}$) | $3 \times 10^7$ | $1 \times 10^7$ |
| Hole thermal velocity, $V_h$ ($cms^{-1}$) | $3 \times 10^7$ | $1 \times 10^7$ |
| Layer density, ($gcm^{-1}$) | 2.328 | 2.328 |
| Auger electron recombination coefficient, ($cm^6s^{-1}$) | 0 | 0 |
| Auger hole recombination coefficient, ($cm^6s^{-1}$) | 0 | 0 |
| Band-to-band recombination coefficient, ($cm^3s^{-1}$) | 0 | 0 |
| Defect properties | Single acceptor | Single acceptor |
| Total trap density ($cm^{-3}$) | $1 \times 10^{14}$ | $1 \times 10^{14}$ |
| Characteristic energy (eV) | 0.56 | 0.56 |

given as $\sqrt{\frac{2\pi}{n_{2D}}}$, and $n_{2D}$ is the carrier density in $cm^{-2}$[21]. As diamane is synthesized by chemical exposure of bilayer graphene to H or F atoms which results in their attachment to graphene surface, hence we have taken the value of carrier density reported in graphene. By following the condition reported by Sarma et al. [22], even n~$10^5$ $cm^{-3}$ in graphene cannot satisfy the above-mentioned condition additionally, it has also been reported that doping in 2D materials increases the interlayer spacing [21,23]. Therefore, we have considered both graphene as well as diamane as quasi 3D. The quantum confinement can be obtained by lowering the thickness. The thickness of single layer diamane is taken as 0.334 nm and the [n, k] file has been calculated by using the data provided in the previous literature [24]. The value of refractive index (n) and extinction coefficient (k) are estimated to be 1.5 and 2.7, respectively [24]. Because of isotropic nature of diamane, properties like mobility and effective mass of charge carriers are same in both armchair and zigzag directions.

For optimizing the cell parameters and to simulate the cell, the bulk and interfacial defect distribution as well as recombination rate (G) were taken constant in the software. Moreover, the model chosen for ITO/n-diamane and p-crystalline Si/Ag interface was MS-Schottky model and the model chosen for n-diamane/p-cSi interface is drift-diffusion, respectively. The parameters which have to be optimized are: $N_D$ (donor concentration), $N_C/N_V$ (effective conduction and valance band density), energy bandgap ($E_g$), $\chi$ (affinity), $\varepsilon_r$ (dielectric constant), layer thickness, electron and hole mobilities ($\mu_n$, $\mu_h$). The optimization of above-mentioned parameters for diamane was carried out by taking the set values of parameters of p-cSi layer in the range of reported values in the literature and have been tabulated in Table 2. $N_C/N_V$ of diamane have been calculated by using the equation as follows:

$$N_{C/V} = 2\left(\frac{2\pi k_B T m^*_{e/h}}{h^2}\right)^{\frac{3}{2}} \quad (1)$$

Where $k_B$ is the Boltzmann constant, T is the temperature in Kelvin, $m^*_h$ is the hole effective mass, $m^*_e$ is the electron effective mass and h is the Planck's constant [25,26]. Whereas, the thermal velocity ($V_{th}$) of charge carriers has been calculated by using the relation $V_{th} = \sqrt{\frac{3k_BT}{m^*_{e/h}}}$. After getting the best optimized results of diamane's parameters, we have further optimized the various parameters of p-type c-Si wafer such as: $N_A$ (acceptor concentration), $N_C/N_V$ (conduction band and valance band density), $\chi$ (affinity), and wafer thickness by keeping the parameters of n-Dn at their best optimized values. The range of parameters of n-diamane have been taken from the available literatures [4,12,14,16,21,26] and the range of parameters of p-cSi which we have considered here are for (100) oriented single crystalline Si [26–29] tabulated in Table 2.

The schematic of energy band diagram for the proposed structure ITO (TCO)/n-Dn/p-cSi/Ag has been depicted in Fig. 1(b). Since, drift-diffusion effect is considered at the interface therefore, the majority of charge carriers i.e., electrons from the n-diamane and holes from the p-cSi are diffused towards each other until thermal equilibrium is developed. This happens because of concentration gradient at the interface, this phenomenon is called diffusion process. An electric field exists in the depletion region which spread up to some width called depletion width W, in both p and n regions. Both electrons and holes in this depletion region are carried away by this electric field. As a result, a drift current devel-

ops which balances the diffusion current and appears due to the difference in electron and hole densities at the junction. As a result, thermal equilibrium is established and the energy bands start bending near the Fermi level. The built-in potential ($eV_{bi}$) is the reason for charge segregation across the junction in the solar cell as discussed above. This built-in potential ($eV_{bi}$) in both p-cSi and n-diamane can be framed in the form of work-function according to the equations written as follows:

$$e.V_{bi} = \Phi_{BI} = \Phi_{Si-P-type} - \Phi_{Dn,\ N-type}$$

$$= -\chi_{Dn} - kT\ln\left(\frac{N_C}{N_D}\right) + \chi_{Si} + kT\ln\left(\frac{N_A}{N_V}\right) \quad (2)$$

Where $\Phi_{Dn}$ and $\Phi_{Si}$ are the work function of diamane and silicon, respectively. Whereas, $\chi_{Dn}$ and $\chi_{Si}$ are the affinity of diamane and Si, respectively. Moreover, $N_D$ is the donor concentration and $N_C$ is the effective conduction band density for diamane, while $N_A$ is acceptor concentration and $N_V$ is the effective valence band density for p-crystalline Si [26,30].

As n-diamane/p-crystalline Si junction is illuminated, it gives rise to electron-hole pairs (EHP) which gets separated because of the built-in potential. The minority charge carriers flow towards the n-diamane and p-crystalline Si respectively, depicted in Fig. 1(b). When open circuit condition is applied, the photo-generated charge carriers intensify the carrier density in their corresponding fields and because of this intensification of charge carriers, quasi-Fermi level approaches the conduction band in n-diamane and valance band in p-cSi as a result $V_{OC}$ develops. In the same manner under short-circuit condition, these photo-generated charge carriers do not intensify because of the collection of charge carriers at their respective electrodes which results in the $J_{SC}$ that flows in the outer circuit.

## 3. Results

### 3.1. Optimization of n-type diamane layer parameters

To simulate the above designed cell, first we have optimized the doping concentration ($N_D$) of n-type diamane layer ranging from $3 \times 10^{14}$ $cm^{-3}$ to $1 \times 10^{18}$ $cm^{-3}$ while keeping other parameters of diamane and p-type c-silicon at some reasonable values reported in the literatures [4,12,14,16,21,26–29] and the results obtained from this optimization have been plotted in Fig. 2 (a and b). Fig. 2 (b) shows that, as $N_D$ increased from $3 \times 10^{14}$ $cm^{-3}$ to $1 \times 10^{18}$ $cm^{-3}$, $V_{OC}$ found to be constant at 521.1 mV. It might be happened because as we increase the doping concentration above $3 \times 10^{14}$ $cm^{-3}$, mobility decreases due to enhance collisions of charge carriers with ionized doping atoms which results in decrease of carrier's lifetime according to the relation: $\mu_i = \frac{e}{m^*}\tau_i$ here, e is the electronic charge and $m^*$ is the effective mass of electron. Since, mobility and diffusion constant relate by Einstein's relation given by $\frac{D_i}{\mu_i} = \frac{k_B}{e}T$ [25,26] and indicates that as mobility decreases diffusion constant also decreases as it is directly proportional to the mobility of charge carriers and therefore, diffusion length ($L_i = \sqrt{D_i\tau_i}$) reduces. Diffusion length is the average distance an electron can move and recombine, it depends on the recombination time as well as diffusion constant. In this manner, $J_{SC}$ will also decrease and $J_0$ increases following the relations given below:

$$J_{SC} = qG(L_n + L_p) \quad (3)$$

Where $L_n$ and $L_p$ are the diffusion length of electron and hole, respectively. q is the electric charge and G is the electron and hole pair generation rate [26,31].

$$J_0 = eA\left(\frac{D_p P_n}{L_n} + \frac{D_n n_p}{L_p}\right) \quad (4)$$

Where, A is the area of junction [25,26]. This increment in $J_0$ and decrement in $J_{SC}$ results in a reduction of $V_{OC}$ following by the relation:

$$V_{OC} = \frac{nqT}{k} ln\left(\frac{J_{SC}}{J_0} + 1\right) \quad (5)$$

here, $J_0$ is the reverse saturation current. But at the same time, $V_{bi}$ increases according to Eq. (2), results in higher $V_{OC}$. These two processes contradict each other and make the balance to keep $V_{OC}$ at a constant value 521.1 mV. Another possibility for $V_{OC}$ being constant at 521.1 mV is that $V_{OC}$ is independent of $N_D$, if it satisfies the condition: $\Delta\ n_o >> N_D$, where $\Delta\ n_o$ is the additional electron and hole pairs [26,32]. Further, as we increase $N_D$ from $3 \times 10^{14}$ $cm^{-3}$ to $5 \times 10^{15}$ $cm^{-3}$, the conductivity of diamane layer was found to increase and consequently, $J_{SC}$ increased very slightly from 20.87 to 20.99 $mA/cm^2$. On further increasing $N_D$ from $5 \times 10^{15}$ $cm^{-3}$ to $1 \times 10^{18}$ $cm^{-3}$, $J_{SC}$ start decreasing from 20.99 to 20.67 $mA/cm^2$, this happens because at higher $N_D$, diffusion length decreases and as a result $J_{SC}$ decreases. Another reason behind this decrement in the value of $J_{SC}$ from 20.99 to 20.67 $mA/cm^2$ after $N_D$ $5 \times 10^{15}$ $cm^{-3}$ is that on increasing doping concentration, work function will reduce due to this reduction in work function $J_{SC}$ falls off. From Fig. 2(a) it is observed that there is a sudden increase in the value of fill factor (FF) and efficiency ($\eta$) after $N_D = 3 \times 10^{14}$ $cm^{-3}$. This might be happened because of the decrement in the sheet resistance which reduced on increasing doping concentration. After $N_D = 1 \times 10^{16}$ $cm^{-3}$, FF and $\eta$ decrease very slightly and then again starts to increase as $N_D$ increases. Maximum efficiency 7.74% was achieved at $N_D = 10^{18}$ $cm^{-3}$. Therefore, the best optimized value of $N_D$ for n-type diamane layer found to be $1 \times 10^{18}$ $cm^{-3}$.

After optimizing $N_D$ at $10^{18}$ $cm^{-3}$, the next parameter conduction/valence band density ($N_C/N_V$) is optimized by varying in the range of $1 \times 10^{19}$ $cm^{-3}$ to $6 \times 10^{21}$ $cm^{-3}$ for ultra-single diamane wafer. On increasing $N_C/N_V$, the densities of minority charge carriers get reduced. This reduction in minority carriers is possibly because of the smaller period of photo-generated carriers and trapping of the free charge carriers. Due to which, $V_{OC}$ will remain constant at 521.1 mV as we varied $N_C/N_V$ from $1 \times 10^{19}$ $cm^{-3}$ to $6 \times 10^{21}$ $cm^{-3}$. From the Fig. 2(c, d), we noticed that $J_{SC}$ increases linearly on increasing $N_C/N_V$ of diamane layer this increment in $J_{SC}$ is due to the high value of $N_C$ which leads to higher absorption of photons. It is also observed that the FF reduces linearly from 71.88% to 65.12% on varying the $N_C/N_V$ from $1 \times 10^{19}$ $cm^{-3}$ to $6 \times 10^{21}$ $cm^{-3}$. Since FF is reducing consequently, $\eta$ will be reduced from 7.74% to 7.04%. Maximum efficiency 7.74% was achieved at $N_C/N_V = 1 \times 10^{19}$ $cm^{-3}$. Hence, $1 \times 10^{19}$ $cm^{-3}$ is the best optimized value of $N_C/N_V$ for single layer n-diamane.

After optimizing $N_D$ and $N_C/N_V$, the next parameter bandgap ($E_g$) is optimized in the range of 0.86 to 5.6 eV. Here, we have chosen a higher range of bandgap for the simulation by keeping in mind that if the UV light present in solar light could contribute to the higher bandgap. We have taken this range of bandgap in order to see the effect of bandgap on the solar cell's efficiency for all derivatives of diamane as it has various derivatives which possesses a range of bandgap depending on the functional group they contain and many other factors too. But we did not notice any change in the efficiency of solar cell after the bandgap 1.3 eV. It is reported that dielectric constant is inversely proportional to the bandgap energy therefore, on changing the bandgap value dielectric constant also changes. For higher bandgap dielectric constant becomes smaller which consequently increases the recombination loses due to the improvement in the exciton binding energy. Therefore, on further increasing the bandgap there will not be seen any improvement in the value of efficiency of the solar cell and it became constant 7.74% after the bandgap 1.3 eV. In this specific sec-

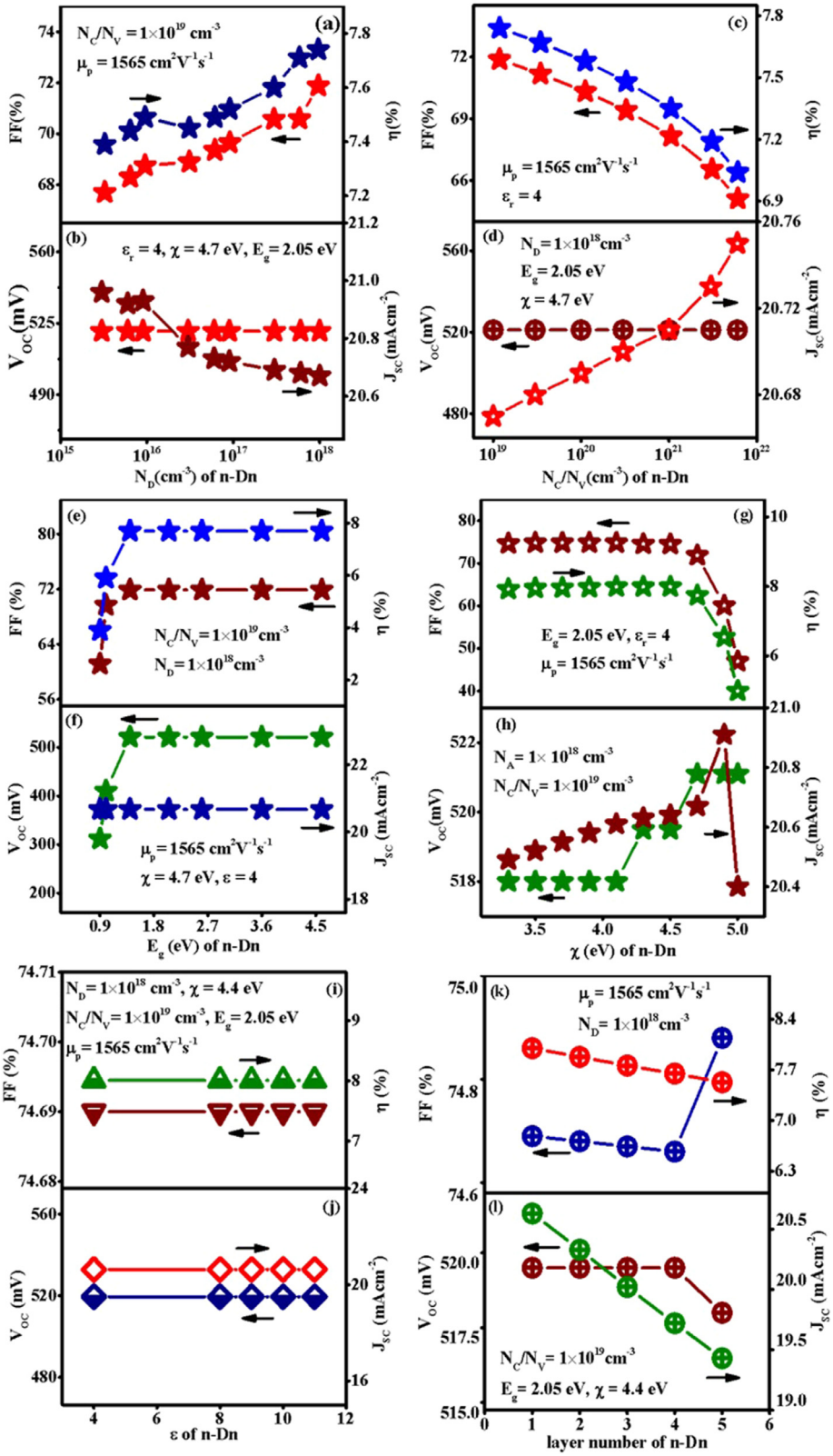



**Fig. 2.** Graphic representation of the performance of designed cell after optimization of different variables of n-diamane, (a, b) $N_D$ (cm$^{-3}$) donor concentration, (c, d) $N_C/N_V$ (cm$^{-3}$) effective conduction/valance band density, (e, f) bandgap $E_g$ (eV), (g, h) $\chi$ (eV) electron affinity, (i, j) dielectric constant $\varepsilon_r$, and (k, l) layer numbers. After fixing the variables of p-cSi at some constant values like: $N_A$ (acceptor concentration) = $1 \times 10^{16}$ cm$^{-3}$, $N_C/N_V$ (effective conduction/valance band density) = $3 \times 10^{19}$ cm$^{-3}$, $\chi$ (affinity) = 4.05 eV and layer thickness = 100 $\mu$m. Remaining variables of both p-cSi and n-Dn are mentioned in Table 2.

tion of manuscript, we discussed only about the effect of varying the bandgap because we exclusively want to see the effect of varying bandgap on the performance of solar cell. Therefore, we did not discuss about the variation in the dielectric constant here. We have discussed about the effect of varying the dielectric constant on the performance of solar cell in another section of the manuscript. As we increased $E_g$, $V_{OC}$ found to be increased and attained a constant value 521.1 mV at 1.3 eV. Whereas, $J_{SC}$ remains constant at 20.67 mA/cm$^2$ as depicted in Fig. 2(f). In Fig. 2(e), it has been shown that FF and $\eta$ increase from 63.96% to 72.09% and 2.79% to 7.74% respectively, for the bandgap ranging from 0.86 to 1.25 eV and then attain a constant value 71.88% for the FF and 7.743% for the efficiency at the bandgap 1.3 eV. Therefore, the bandgap 1.3 eV is the best optimized bandgap for single layer n-diamane.

The optimized results of the electron affinity ($\chi$) in the range from 3.3 eV to 5.2 eV is shown in Fig. 2 (g and h). As $\chi$ increased, very slight increment in the value of $\eta$ from 7.92 to 8% was observed up to the electron affinity 4.4 eV and then a dramatical reduction in the value of efficiency was noticed as shown in Fig. 2(g). The reason behind this small linear increment in the efficiency is that as we increase electron affinity of single layer diamane above 3.3 eV, the barrier height gets decrease which results in the fast movement of minority charge carriers as a result, $J_{SC}$ increased. Thus, $\eta$ was observed to increase linearly up to 4.4 eV. After 4.4 eV, as we increase affinity, the conduction band lowers down and starts making rapid transportation of charge carriers across the interface as a result, there would not be any free charge carrier. Consequently, $J_{SC}$ and FF will be decreased which cause final $\eta$ to decrease. Here, we found that on increasing $\chi$, $V_{OC}$ was improved and attained maximum value 524.2 mV at $\chi = 5.2$ eV and maximum $\eta$ 8% was observed for $\chi = 4.4$ eV. Thus, 4.4 eV is the best optimized value of electron affinity for single layer n-type diamane.

After this we have optimized the value of relative permittivity ($\varepsilon_r$) or dielectric constant. The value of dielectric constant ($\varepsilon_r$) for a bilayer graphene was found to be in the range from 4 to 9 in different experiments [33]. Here we varied the value of relative permittivity ($\varepsilon_r$) in the range from 4 to 11 and the performance of proposed solar cell are depicted in Fig. 2 (i and j). From Fig. 2 (i and j), it is noticed that as the dielectric constant increased from 4 to 5 then may be due to the reduction in the excitation binding energy the recombination losses decreased and therefore the efficiency increased by a small value from 8.006% to 8.007%. But on further increasing the value of dielectric constant the efficiency falls off from 8.007% to 8.006% rather than increasing, as we know higher dielectric constant reduces the exciton binding energy which consequently results in decreased recombination losses. This may be have happened because it is reported that $V_{bi} \propto \frac{1}{\varepsilon_r}$ [25,26,34]. Therefore, an increment in the value of dielectric constant deteriorates $V_{bi}$ which consequently decrease the efficiency [26]. In this manner, these two effects balance each other and no further change in the values of $V_{OC}$, $J_{SC}$, FF and $\eta$ has been observed. In this way, the maximum efficiency 8.007% was achieved at the value 5 of dielectric constant.

After optimizing the relative permittivity ($\varepsilon_r$), we have further optimized the hole mobility of n-Dn layer ranging from 743 to 1565 cm$^2$V$^{-1}$s$^{-1}$ and we noticed that there is no effect of hole mobility on the performance of the photovoltaic cell, because of the ultra-thin layer of diamane.

To understand the role of layers number of n-diamane in the production of charge carriers due to absorption of photons from sunlight by the solar cell with all parameters those we have optimized above at single layer diamane, the above simulated cell was further optimized up to 5 layers of diamane while keeping the parameters of c-Si at the previous constant values that we have taken for previous simulation. It is reported that as layer number increases, both sheet resistance as well as transmittance decrease which results in lowering the electron-hole pair generation rate at the interface that depends on the transmittance of diamane. We found that, as layer numbers of n-diamane increased, $V_{OC}$ decreased very slightly from 519.5 mV to 518 mV and a variation in the value of $J_{SC}$ was found from 20.63 mA/cm$^2$ to 19.14 mA/cm$^2$. This was happened, because of smaller electron-hole pair generation rate. The FF was also found to decreased slightly from 74.69% to 74.66% and then decreased infinitesimally on further increasing the number of layers. Because of lower value of $V_{OC}$, $J_{SC}$ and FF, $\eta$ also decreased from 7.88% to 7.53%. The best optimized efficiency 8% was achieved for single layer n-diamane. Since, we found deterioration in the values of $V_{OC}$, $J_{SC}$, FF and $\eta$ on increasing the layer numbers of n-diamane we can say that thickness of n-diamane does not play an important role in improving the efficiency of the solar cell and the generation of photo carriers. We also know that on increasing the thickness the bandgap decreases which in results increases the dielectric constant which consequently decrease the recombination losses due to decrease in exciton binding energy. Therefore, according to this concept efficiency should get increased on increasing the thickness of n-diamane but this is not seen from the results shown in Fig. 2 (k and l). This may have happened because of the reduction in transmittance on increasing the layer number which compensates the effect of reduced bandgap on increasing the dimensionality of carbon material (diamane). From Fig. 2 (k and l) it is noticed that on increasing the layer number of diamane from 1 to 5 the efficiency changed slightly from 8% to 7.53%. Therefore, we can say that increased number of layers of diamane do not play a significant role in enhancing the efficiency of solar cell. It only provides the required Schottky junction at the interface for the current movement and act as an effective emitter layer in c-Si heterojunction solar cell.

### 3.2. Optimization and simulation of p-crystalline silicon wafer parameters

We further optimized variables of p-crystalline Si wafer while, setting the parameters of n-Dn at best optimized values. To carry out this simulation, first we optimized the acceptor concentration of p-cSi wafer. This will intensify the efficiency of the ITO (TCO)/n-Dn/p-cSi/Ag structured solar cell, the value of $N_A$ for p-cSi wafer was increased from $1 \times 10^{16}$ cm$^{-3}$ to $7 \times 10^{18}$ cm$^{-3}$ and the desired outputs are shown in Fig. 3 (a and b). As we increase $N_A$, the Fermi Level in Si layer moves towards the valence band edge and results in an increase of work function of p-type Si ($\Phi_{p-Si}$), and barrier height according to Eq. (2) which will increase both $V_{bi}$ and $V_{OC}$ according to the equations:

$$V_{OC} = \frac{V_{bi}}{q} - \frac{nkT}{q} \ln\left(\frac{qN_V S_{it}}{J_{SC}}\right) \quad (6)$$

$$V_{bi} = \frac{kT}{q} \ln\left(\frac{N_A N_D}{n_i^2}\right) \quad (7)$$

Where, $S_{it}$ is the interface recombination velocity [35–37]. Therefore, as $N_A$ increases $V_{bi}$ also increases which helps the generated charge carriers in approaching their respective electrodes due to which the values of both $V_{OC}$ and $J_{SC}$ enhanced from 519.5 mV to 697.7 mV and 20.63 mAcm$^{-2}$ to 28.71 mAcm$^{-2}$ respectively, as depicted in Fig. 3(b). FF and $\eta$ were also noticed to increase with $N_A$ from 74.69% to 83.58% and from 8% to 16.74% respectively, as shown in Fig. 3(a). The increment in FF might be related to the decrement in the resistivity of the p-cSi layer [26,38]. The maximum efficiency 16.74% was achieved at $N_A = 7 \times 10^{18}$ cm$^{-3}$. Therefore, $7 \times 10^{18}$ cm$^{-3}$ is the best optimized value of $N_A$ for p-cSi layer Eqs. (6), (7).

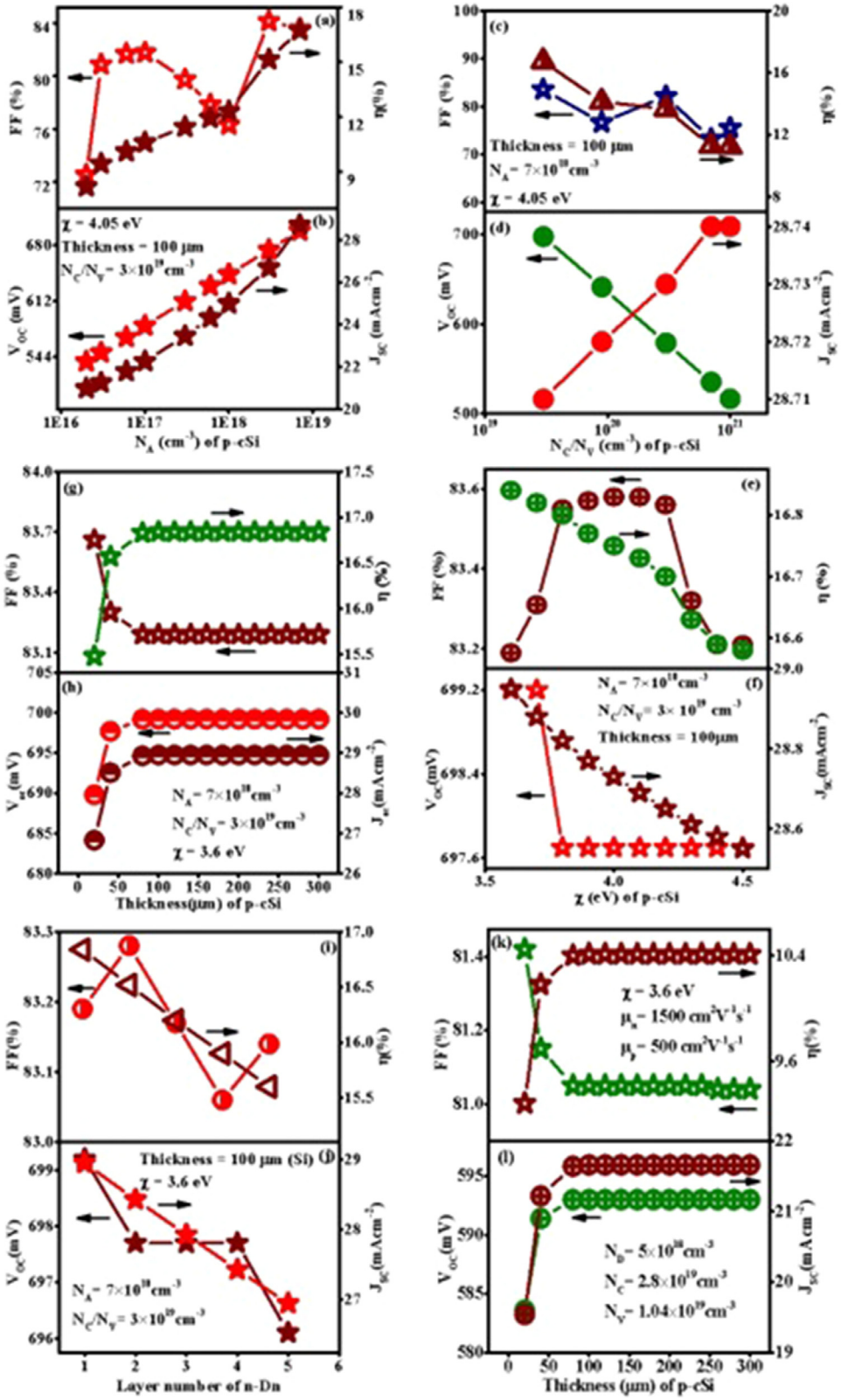



**Fig. 3.** Graphic representation of the performance of the designed cell after simulation of different variables of p-crystalline Si wafer, (a, b) acceptor concentration $N_A$ ($cm^{-3}$), (c, d) $N_C$ /$N_V$ ($cm^{-3}$) conduction/valance band density, (e-f) $\chi$ (eV) affinity, (g, h) wafer thickness (μm), (i, j) layer number of n-Dn at best optimized values of p-cSi wafer and (k, l) thickness of Si layer at commercially available values of parameters p-cSi wafer, whereas the values of parameters of n-Dn are kept at their best optimized values i.e., $N_D = 1 \times 10\ 18$ ($cm^{-3}$), $N_C/N_V = 1 \times 10\ 19$ ($cm^{-3}$), $E_g = 2.05$ (eV), $\chi = 4.4$ (eV), $\mu_n = 2732$ ($cm^2V^{-1}s^{-1}$) and $\mu_p = 1565$ ($cm^2V^{-1}s^{-1}$).

After getting best optimized value of $N_A$, we carried out the simulation further by optimizing $N_C/N_V$ of p-cSi ranging from $3 \times 10^{19}$ $cm^{-3}$ to $1 \times 10^{21}$ $cm^{-3}$ and the outputs are plotted in Fig. 3 (c, d). Here we found that the $V_{OC}$ is decreased from 697.7 mV to 516.4 mV as $N_C/N_V$ increased. As $N_C/N_V$ increases, $V_{bi}$ increases by following Eq. (2). Therefore, the trapping speed of charge carriers inside the Si layer reduce. On more increasing the value of $N_C/N_V$, the mobility and $L_i$ of charge carriers decrease by following the Eqs. (1), (3) and (4). By taking the Eq. (5) in consideration we can say that $J_O$ also increases on increasing $N_C/N_V$. This increment in the value of $J_O$ as well as $V_{bi}$ results in the decrement of $V_{OC}$. The value of $J_{SC}$ was found to increased very slightly from 28.71 $mAcm^{-2}$ to 28.74 $mAcm^{-2}$ as we increased $N_C/N_V$ of p-cSi from $3 \times 10^{19}$ $cm^{-3}$ to $5 \times 10^{20}$ $cm^{-3}$ and then becomes constant. This might have happened because, the bandgap and photons absorption are highly influenced by $N_C/N_V$ [26,39]. Therefore, at $N_C/N_V = 3 \times 10^{19}$ $cm^{-3}$, active silicon layer absorbed more photons which increased the generation rate (G) of electron-hole pair (EHP) and then according to Eq. (3), the higher value of G may balance the lower value of $L_i$ which results in constant value of $J_{SC}$ 28.74 $mAcm^{-2}$. The FF was first found to increase slightly from 83.58 to 83.63% and then decreases up to 75.58% on further increasing the value of $N_C/N_V$ after $8 \times 10^{19}$. The efficiency was found to be decreased linearly from 16.74% to 11.22% on increasing $N_C/N_V$ from $3 \times 10^{19}$ $cm^{-3}$ to $1 \times 10^{21}$ $cm^{-3}$ as depicted in Fig. 3(c). The initial increment in FF may be because of an increment in $J_{SC}$. At higher value of $N_V$, FF decreases abruptly because it depends on $V_{OC}$. In the similar manner, a direct decrement in the value of $\eta$ was observed because of its direct dependency on FF. The maximum value of efficiency 16.74% was found at $N_C/N_V = 3 \times 10^{19}$ $cm^{-3}$. So, $3 \times 10^{19}$ $cm^{-3}$ is the best optimized value of $N_C/N_V$ for p-cSi layer.

After optimizing $N_A$ and $N_C/N_V$ to their best optimized values we further optimized the electron affinity of p-cSi layer in the range from 3.6 to 4.5 eV. For this affinity, $V_{OC}$ was found to slightly increased from 699.2 mV to 697.7 mV on increasing affinity from 3.6 eV to 3.8 eV and then it stayed constant at the value 697.7 mV on further increase in affinity after 3.8 eV. Whereas, the value of $J_{SC}$ was noticed to decrease straight from 28.95 $mA/cm^2$ to 28.55 $mA/cm^2$ as depicted in Fig. 3(e, f). The value of FF first increased from 83.19% to 83.58% on increasing the electron affinity from 3.6 eV to 4.1 eV but on further crossing of the value of electron affinity the FF starts decreasing from 83.58% and reached 83.21% at affinity 4.5 eV. This reduction in FF on increasing the affinity from 4.2 eV to 4.5 eV can be understood by the concept of changing the work function of p-cSi at the interface by the relation: $\Phi_B = \Phi_{p-cSi} - \chi$, where $\Phi_B$ is the barrier height for p-type semiconductor, $\Phi_{p-cSi}$ is the work function of p-cSi and $\chi$ is the affinity of Si due to which the carrier transportation rate in the p-cSi layer would reduce. This reduction will increase the series resistance of both p-cSi wafer and the cell. Therefore, the FF reduces after $\chi > 4.1$ eV. The efficiency of the optimized cell was found to decrease linearly from 16.84% to 16.58% as we increase affinity in the range 3.6 to 4.5 eV. In this way the best optimized cell was obtained at affinity 3.6 eV.

To further analyze the possibility of cost-effective cell, the next variable we optimized is the thickness of p-cSi layer. We have optimized the thickness of p-cSi wafer ranging from 20 to 300 µm. From Fig. 3 (g, h), it is clearly visible that for 20 µm thick p-cSi layer, the efficiency is 15.48%. Higher efficiency 16.84% was obtained for 100 µm thick silicon layer. This is because as the thickness increases number of charge carriers increases consequently, $J_{SC}$ increases. For thicker layer of p-cSi, the lifespan of minority charge carriers gets reduced because of its less drift length. Additionally, the excess photogenerated charge carriers and their recombination rate have been found in proportion in the thicker silicon wafer, which overall results in a constant value of $V_{OC}$ 699.2 mV and $J_{SC}$ 28.95 $mAcm^{-2}$ after thickness 100 µm. The FF and efficiency were also constant after getting a value 83.19% and 16.84% respectively, beyond the thickness 100 µm. Thus, we did not notice any change in the value of $V_{OC}$, $J_{SC}$, FF and $\eta$ on further increasing the thickness of p-cSi layer after 100 µm and we can say that, these saturated values of $V_{OC}$, $J_{SC}$ and FF have produced a maximum efficiency of 16.84%, even on increasing the thickness of p-cSi wafer up to 300 µm. Another reason for not improving the efficiency further on increasing the thickness of p-cSi wafer beyond 100 µm is that in a thick layer of p-cSi more dangling bonds is found which became the recombination centers for the photogenerated charge carriers due to the increased number of defects on the surface of thick silicon layer. That is why, the decreased bandgap on increasing the thickness of layer does not contribute in improving the conversion efficiency of the solar cell. Therefore, large thickness of p-cSi does not find to play an important role in improving the efficiency of the cell. From Fig. 3 (g and h) it is shown that during the optimization of thickness of p-cSi layer the efficiency increases from 15.48 to 16.84% only which is not much recommendable.

After optimization of silicon wafer's parameters, the cell was further simulated to studied the effect of the layer numbers of diamane and the results are plotted in Fig. 3 (i and j), where we found $\eta$ to decrease from 16.84% to 15.6%. For best commercial appreciation of the simulated cell, we have further optimized the thickness of p-cSi wafer by considering the commercially available parameters of Si layer. For this, we have taken the values of acceptor concentration = $5 \times 10^{16}$ $cm^{-3}$, effective conduction band density = $2.8 \times 10^{19}$ $cm^{-3}$, effective valance band density = $1.04 \times 10^{19}$ $cm^{-3}$, affinity = 3.6 eV, electron mobility = 1500 $cm^2/Vs$ and hole mobility = 500 $cm^2/Vs$ from the previously reported data [26,40,41]. We optimized the thickness of silicon layer in the range from 20 to 300 µm and the highest efficiency 10.41% was found for thickness 100 µm as shown in Fig. 3 (k and l). Since, front and back contacts play an important role in the solar cell as the front contact is required to be highly transparent so that maximum photons can reach to the active layer and enhance the conversion efficiency. Therefore, we preferred the front contact with low reflection coefficient or nearly 0. Hence, we have given more preference to ITO (Indium Tin Oxide), as TCO (Transparent Conducting Oxide) in designing of our solar cell to intensify the effective illuminating area, because ITO increases the optical absorption instead of reflection that occurred in other metal contacts and it possesses low sheet resistance, high mobility and it is optically transparent too. Furthermore, in order to check the effect of back contact on the solar cell we have changed the back contacts by Au (where $n$ is taken as 0.27049 and $k$ is taken as 2.7789) and Tin (where $n$ is taken as 1.96 and $k$ is equal to 3.06), these files are present already in the AFORS-HET software and we found that maximum efficiency is achieved in case of Au and Ag because of their larger work function. This indicates that both reflection coefficient and work function of front as well as back constant play significant role in the solar cells. Here, we have given more preference to ITO (Indium Tin Oxide), as TCO (Transparent Conducting Oxide) in designing of our solar cell to intensify the effective illuminating area, because TCO increases the optical absorption instead of reflection that occurred in other metal contacts.

### 3.3. Spectral response, quantum efficiency and temperature dependency

Generally, spectral response is conceptually similar to the quantum efficiency (QE) of the solar cell**.** The QE is defined as the ratio of the number of electrons collected by the cell as output to the number of photons incident on its surface whereas, spectral re-

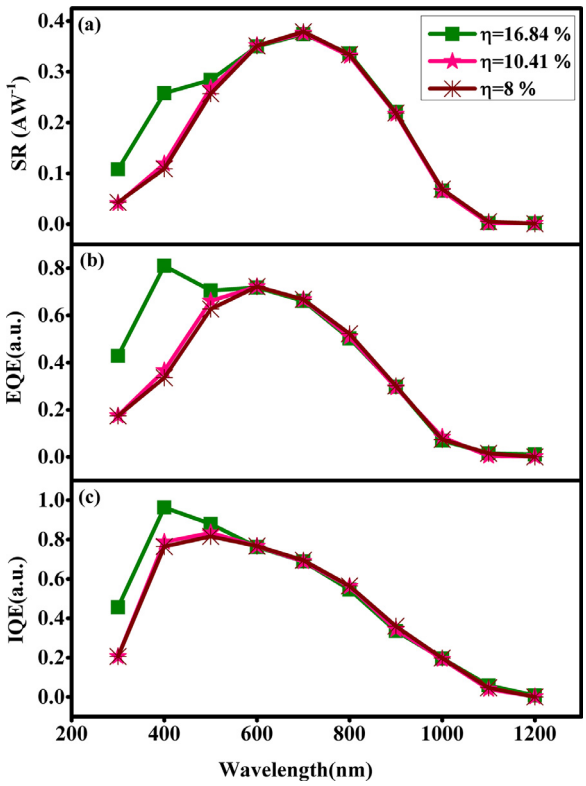



**Fig. 4.** Wavelength dependency of (a) SR (Spectral response), (b) EQE (External quantum efficiency) and (c) IQE (Internal quantum efficiency) of the best simulated solar cell at single layer n-Dn/p-cSi ($\eta$ = 8%, $\eta$ = 16.84% with $N_A = 7 \times 10^{18}$ ($cm^{-3}$), $N_C/N_V = 1 \times 10^{19}$ ($cm^{-3}$) for 100 $\mu$m p-cSi and efficiency = 10.41% with $N_A = 5 \times 10^{16}$ ($cm^{-3}$), $N_V = 1.04 \times 10^{19}$ ($cm^{-3}$) and $N_C = 2.8 \times 10^{19}$ ($cm^{-3}$) at 100 $\mu$m p-crystalline Si).

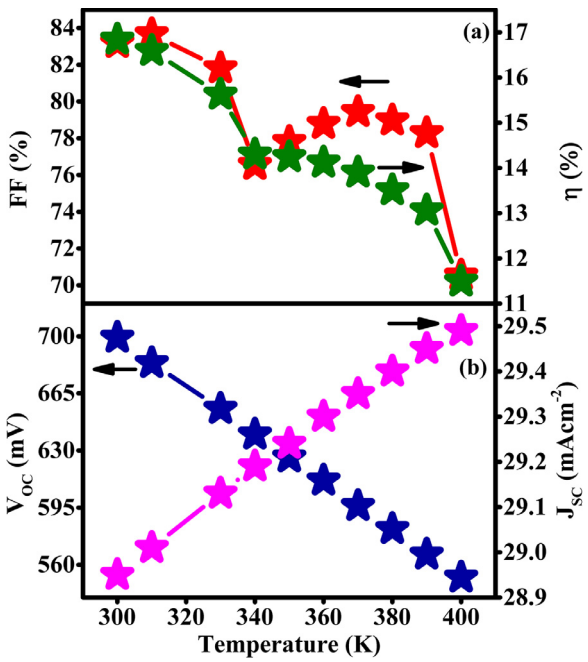



**Fig. 5.** Dependency of the optimized cell response with temperature variation at $\eta$ = 16.84%. (a) efficiency and fill factor, (b) $V_{OC}$ and $J_{SC}$.

sponse is defined as the ratio of current generated by the cell to the optical power incident on its surface [26]. Quantum efficiency and spectral response both are used in cell analysis one uses the photons flux and the other uses the power of light at each wavelength, respectively. QE can be converted into spectral response (SR) by using the formula: SR = $\frac{q\lambda}{hc}$ QE. To analysis the spectral response (SR), external quantum efficiency (EQE) and internal quantum efficiency (IQE) of the optimized cell at efficiencies 8%, 10.41% and 16.84% we performed the simulation in the wavelength range 300–1200 nm and the results are shown in Fig. 4. We observed that, at shorter wavelengths i.e., below 600 nm the diamane layer was found to absorb most of the light, this results in less photons passing to p-cSi layer subsequently small amount of current would generate in the shorter wavelengths and the cell response was found to be very low. In the wavelength range from 600 to 1000 nm most of the photons reach silicon layer where, the energy they constitute found equal to the bandgap energy of silicon layer and this energy then converted into electrical energy inside the silicon layer. Therefore, the optimized cell achieved its ideal response at intermediate wavelengths. Since silicon is an indirect bandgap semiconductor therefore, there will not be any sharp cut off at the wavelength corresponding to its bandgap ($E_g$= 1.12 eV). So, the spectral response falls back to zero at longer wavelengths. From Fig. 4, it is clearly visible that for greater than 100 µm thick silicon layer, spectral response maintained its uniformity. This is because the effective conversion rate remains constant as we increase thickness beyond 100 µm.

The internal and external quantum efficiencies (IQE and EQE) of the optimized cell at efficiencies 8%, 10.41% and 16.84% has been observed and the results are being shown in Fig. 4. Internal quantum efficiency do not consider the reflection and transmission of photon outside the solar cell while, EQE considers the outcomes of optical losses i.e., reflection and transmission. Fig. 4 (b and c) reveals that, at initial wavelength 300 nm EQE and IQE is having smaller value this is because of the optical losses at the shorter wavelengths. The IQE and EQE are maximum at the wavelength 400 nm and above 400 nm both efficiencies start falling off and reaches to zero at 1200 nm. Since, parameter's optimization of silicon as well as diamane was unable to nullify the optical losses on the surface of the solar cell. Hence, maximum EQE and IQE was noticed for the wavelength 400 nm. The higher value of EQE at the wavelength 400 nm suggests that as we optimize doping concentration and effective band density then there would be an increment in the electron-hole pair generation rate and subsequently, collection of surplus charge carriers due to which highest efficiency has been achieved at 400 nm. We found that, IQE and EQE responses of the cell with efficiencies 8%, 10.41% and 16.84% are almost same, which indicates the equal amount of charge carriers gathered for this range of wavelengths parallel to the photons absorbed by the solar cell. At higher wavelength regions, the recombination rate of the generated charge carriers is better therefore, both internal and external efficiencies become negligible at 1200 nm wavelength.

In order to realize the effect of varying temperature on the operation of solar cell we have further simulated the best optimized cell comprises efficiency 16.84% in the environment of 300 to 400 K temperature and the outputs are depicted in Fig. 5 (a, b). We observed that, $V_{OC}$ decreased from 699.2 mV to 552.3 mV while, $J_{SC}$ increased moderately from 28.95 mA/cm$^2$ to 29.49 mA/cm$^2$ as we increase the temperature. Following to Energy-Temperature relation written as: $E_g$ (T) = $E_g$ (0) – ($\frac{\alpha T^2}{\alpha+\beta}$), where $E_g$ (T) is the bandgap of semiconductor at any given temperature T, $E_g$ (0) is the bandgap at zero temperature and ($\alpha$, $\beta$) are the material constant [26,31]. As temperature elevates band gap declines smoothly. We also know that, as temperature increases $J_O$ increases which results in a decrease of $V_{OC}$ [26,31]. Furthermore, at smaller value of $E_g$ large number of absorbed photons have sufficient energy to create e-h pairs which results in an increment in the value of $J_{SC}$. Since

**Table 3**
Overview of the variables of the finest simulated solar cell.

| Cell variables | n-diamane layer optimized cell | n-diamane layer and p-cSi layer optimized cell | p-cSi layer optimized cell for practically available Si parameters |
|---|---|---|---|
| $V_{OC}$ (mV) | 519.5 | 699.2 | 593 |
| $J_{SC}$ (mAcm$^{-2}$) | 20.63 | 28.95 | 21.66 |
| FF (%) | 74.69 | 83.19 | 81.05 |
| $\eta$ (%) | 8 (@ p-crystalline Si wafer at $N_A = 1 \times 10^{16}$ (cm$^{-3}$), $N_C/N_V = 3 \times 10^{19}$ (cm$^{-3}$), $\chi = 4.05$ (eV), and thickness = 100 μm | 16.84 (@ p-crystalline Si wafer at $N_A = 7 \times 10^{18}$ (cm$^{-3}$), $N_C/N_V = 3 \times 10^{19}$ (cm$^{-3}$), $\chi = 3.6$ (eV), and thickness = 100 μm | 10.41 [@ p-cSi wafer at $N_A = 5 \times 10^{16}$ (cm$^{-3}$), $N_C = 2.8 \times 10^{19}$ (cm$^{-3}$), $N_V = 1.04 \times 10^{19}$ (cm$^{-3}$), $\chi = 3.6$ (eV), and thickness = 100 μm] |

$V_{OC}$ decreases, the fill factor and efficiency also decrease and the maximum efficiency $\eta = 16.84\%$ found at temperature 300 K. The optimized results of the designed solar cell are shown in Table 3.

## 4. Conclusions

In this article, we have proposed the structure (TCO) ITO/n-Dn/p-cSi/ (back contact) Ag. The independent consequences of optimizing seven parameters of n-type diamane wafer on the operation of solar cell were investigated and only single parameter: hole mobility ($\mu_p$) was noticed to not influence its working whereas, other parameters were directly affecting the behavior of proposed cell. After optimizing the parameters of diamane as well as silicon, the maximum efficiency of 16.84% was achieved and the formation of Schottky junction at n-Dn/p-cSi interface was found to play major role in the processing of the cell as it helps in carrying current across the junction. This junction is responsible for the current movement across n-Dn/p-cSi interface. As the work function of diamane can be varied in the range 7.24 eV to 0.62 eV and the work function of p-type silicon layer is found in the range 4.9–5.1 eV. Here, we have taken the work function of diamane as 4.4 eV smaller than the work-function of p-type silicon which supports the formation of Schottky junction at the interface. We also compared the optimized values with the reported experimental values and found satisfactory results. Higher thickness of both diamane as well as silicon were not found to contribute in improving the efficiency of the solar cell on the large scale. It has been found that the thickness of front and back contacts does not play an important role on the performance of the modelled heterojunction cell. The results of the modelled cell have also not changed much on changing the metal contacts at front and back side of the cell. Finally, it has been concluded that if we could reduce the sheet resistance of diamane and can increase its work-function by any technique like chemical doping, then few layers n-diamane can work more efficiently as an emitter layer in Si heterojunction solar cell.

## 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.

## CRediT authorship contribution statement

**Naima:** Conceptualization, Formal analysis, Writing – original draft. **Pawan K. Tyagi:** Conceptualization, Supervision, Writing – review & editing. **Vinod Singh:** Writing – review & editing.

## Acknowledgments

The author is thankful to the financial support provided by the Delhi Technological University, new Delhi and authors appreciatively accredited Helmholtz-Zentrum Berlin who provides AFORS-HET simulation software.

## References

[1] Z. Zheng, H. Zhan, Y. Nie, X. Xu, D. Qi, Y. Gu, Single layer diamond – a new ultrathin 2D carbon nanostructure for mechanical resonator, Carbon 161 (2020) 809–815.
[2] T. Pakornchote, A. Ektarawong, U. Pinsook, T. Bovornratanaraks, Modifying electronic and elastic properties of 2-dimensional [110]diamond by nitrogen substitution, C J. Carbon Res. 7 (2021).
[3] T. Pakornchote, A. Ektarawong, W. Busayaporn, U. Pinsook, T. Bovornratanaraks, Roles of nitrogen substitution and surface reconstruction in stabilizing non–passivated single-layer diamond, Phys. Rev. B 102 (2020) 075418.
[4] M. Raeisi, B. Mortazavi, E.V. Podryabinkin, F. Shojaei, X. Zhuang, A.V. Shapeev, High thermal conductivity in semiconducting Janus and non-Janus diamanes, Carbon 167 (2020) 51–61.
[5] P.V. Bakharev, M. Huang, M. Saxena, S.W. Lee, S.H. Joo, S.O. Park, J. Dong, D.C. Camacho-Mojica, S. Jin, Y. Kwon, M. Biswal, F. Ding, S.K. Kwak, Z. Lee, R.S. Ruoff, Chemically induced transformation of chemical vapour deposition grown bilayer graphene into fluorinated single-layer diamond, Nat. Nanotechnol. 15 (2020) 59–66.
[6] R.R. Nair, W. Ren, R. Jalil, I. Riaz, V.G. Kravets, L. Britnell, P. Blake, F. Schedin, A.S. Mayorov, S. Yuan, M.I. Katsnelson, H.M. Cheng, W. Strupinski, L.G. Bulusheva, A.V. Okotrub, I.V. Grigorieva, A.N. Grigorenko, K.S. Novoselov, A.K. Geim, Fluorographene: a two-dimensional counterpart of Teflon, Small 6 (2010) 2877–2884.
[7] V.G. Kravets, A.N. Grigorenko, R.R. Nair, P. Blake, S. Anissimova, K.S. Novoselov, A.K. Geim, Spectroscopic ellipsometry of graphene and an exciton-shifted van Hove peak in absorption, Phys. Rev. B 81 (2010) 155413.
[8] F. Piazza, K. Cruz, Raman evidence for the successful synthesis of diamane, Carbon 169 (2020) 129–133.
[9] F. Ke, L. Zhang, Y. Chen, K. Yin, C. Wang, W. Zhu, H. Wang, Y. Lin, Z. Liu, J.S. Tse, G. Wu, R.C. Ewing, W.L. Mao, J. Wu, H.K. Mao, B. Chen, Synthesis and electronic structure characterization of diamane, Condens. Matter Mater. Sci. 10 (2019) 48550.
[10] L. Zhu, W. Li, F. Ding, Giant thermal conductivity in diamane and the influence of horizontal reflection symmetry on phonon scattering, Nanoscale 11 (2019) 4248–4257.
[11] L.A. Chernozatonskii, P.B. Sorokin, A.G. Kvashnin, D.G. Kvashnin, Diamond-like $C_2H$ nanolayer, diamane: simulation of the structure and properties, Condens. Matter 90 (2009) 134–138.
[12] T. Cheng, Z. Liu, Z. Liu, High elastic moduli, controllable bandgap and extraordinary carrier mobility in single-layer diamond, J. Mater. Chem. C 39 (2020) 13819–13826.
[13] J.T. Robinson, J.S. Burgess, C.E. Junkermeier, S.C. Badescu, T.L. Reinecke, F.K. Perkins, M.K. Zalalutdniov, J.W. Baldwin, J.C. Culbertson, P.E. Sheehan, E.S. Snow, Properties of fluorinated graphene films, Nano Lett. 10 (2010) 3001–3005.
[14] K.J. Rietwyk, S.L. Wong, L. Cao, K.M. O'Donnell, L. Ley, A.T.S. Wee, C.I. Pakes, Work function and electron affinity of the fluorine-terminated (100) diamond surface, Appl. Phys. Lett. 102 (2013) 091604.
[15] R. Stangl, C. Leendertz, J. Haschke, Numerical simulation of solar cells and solar cell characterization methods: the open-source on demand program AFORS-HET, Solar Energy, Intech Open, 2010 ch. 14.
[16] K. Patel, P.K. Tyagi, Multilayer graphene as a transparent conducting electrode in silicon heterojunction solar cells, AIP Adv. 5 (2015) 077165.
[17] M.M. Abd El-Raheem, H.M. Ali, N.M. Al-Hosiny, M.S. Abd el aal, Improment of the optical properties of ITO/$SiO_2$/glass films for photovoltaic applications, J. Non Oxide Glas. 2 (2010) 67–76.
[18] T. Ohshima, T. Matsunaga, H. Kawasaki, Y. Suda, Y. Yagyu, Preparation of ITO thin films by pulsed laser deposition for use as transparent electrodes in electrochromic display devices, Trans. Mater. Res. Soc. Jpn. 35 (2010) 583–587.
[19] S. Zare, H. Izadkhah, C. Vittoria, Deposition of magnetoelectric hexaferrite thin films on substrates of silicon, J. Magn. Magn. Mater. 420 (2016) 245–248.
[20] S.Y. Lien, Characterization and optimization of ITO thin films for application in heterojunction silicon solar cells, Thin Solid Films 518 (2010) S10–S13.
[21] C.K. Borah, P.K. Tyagi, S. Kumar, The prospective application of a graphene/$MoS_2$ heterostructure in Si-HIT solar cells for higher efficiency, Nanoscale Adv. 2 (2020) 3231–3243.
[22] S.D. Sarma, S. Adam, E.H. Hwang, E. Rossi, Electronic transport in two-dimensional graphene, Rev. Mod. Phys. 83 (2011) 407–470.
[23] Z. Lin, B.R. Carvalho, E. Kahn, R. Lv, R. Rao, H. Terrones, Defect engineering

of two-dimensional transition metal dichalcogenides, 2D Materials 3 (2016) 022002.
[24] S. Bouzidi, M. Barhoumi, M. Said, Optical properties of Janus and non-Janus diamanes monolayers using ab-initio calculations, Optik 235 (2021) 166642.
[25] J. Singh, Semiconductor Devices: Basic Principles, 2nd ed., Wiley India (P) Ltd, New Delhi, 2004.
[26] C.K. Borah, P.K. Tyagi, S. Kumar, K. Patel, Few-layer p-type phosphorene sheet: an efficient transparent conducting electrode in silicon heterojunction solar cell, Comput. Mater. Sci. 151 (2018) 65–72.
[27] Q. Wang, M.R. Page, E. Iwaniczko, Y.Q. Xu, L. Roybal, R. Bauer, B. To, H.C. Yuan, A. Duda, Y.F. Yan, Crystal silicon heterojunction solar cells by hot-wire CVD, Conf. Rec. IEEE Photovolt. Spec. Conf. (2008).
[28] U. Gangopadhyay, S. Roy, S. Garain, S. Jana, S. Das, Comparative simulation study between n-type and p-type Silicon Solar Cells and the variation of efficiency of n-type Solar Cell by the application of passivation layer with different thickness using AFORS HET and PC1D, IOSR J. Eng. 8 (2012) 41–48.
[29] S. Xu, X. Zeng, W. Wang, G. Zhou, Y. Hu, S. Wu, Y. Zeng, Simulation and optimization characteristic of novel $MoS_2$/c-Si HIT solar cell, J. Min. Mat. Char. Eng. 5 (2017) 323–338.
[30] M. Lundstrom, Heterostructure fudamentals (electronegatividad de electrones), PDF 1995 p. 43.
[31] K.K. Nair, J. Jose, A. Ravindran, Analysis of temperature dependent parameters on solar cell efficiency using MATLAB, Int. J. Eng. Dev. Res. 3 (2016) 536–541.
[32] A.V. Sachenko, V.P. Kostylyov, M.V. Gerasymenko, R.M. Korkishko, M.R. Kulish, M.I. Slipchenko, I.O. Sokolovskyi, V.V. Chernenko, Analysis of the silicon solar cells efficiency. Type of doping and level optimization, Semicond. Phys. Quantum Electron. Optoelectron. 19 (2016) 67–74.
[33] R. Bessler, U. Duerig, E. Koren, The dielectric constant of a bilayer graphene interface, Nanoscale Adv. 1 (2019) 1702–1706.
[34] Y. Zhao, C. Liang, M. Sun, Q. Liu, D.Li F.Zhang, Z. He, Effect of doping on the short-circuit current and open-circuit voltage of polymer solar cells, AIP 116 (2014) 154506.
[35] D. Hamdani, S. Prayogi, Y. Cahyono, G. Yudoyono, D. Darminto, The effects of dopant concentration on the performances of the a-$SiO_x$:H(p)/a-Si:H($i_1$)/a-Si:H($i_2$)/μc-Si:H(n) heterojunction solar cell, Int. J. Renew. Energy Dev. 11 (2022) 173–181.
[36] A. Belfar, The role of p+-layer dopant concentration, p+-layer band gap and p+-layer thickness in the performances of a-Si: H n – i – p – p+ solar cells with double layer window nanocrystalline silicon, Optik 126 (2015) 5688–5693.
[37] N. Jensen, R.M. Hausner, R.B. Bergmann, J.H. Verner, U. Rau, Optimization and characterization of amorphous/crystalline silicon heterojunction solar cells, Prog. Photovolt. Res. Appl. 10 (2002) 1–13.
[38] M. Belarbi, M. Beghdad, A. Mekemeche, Simulation and optimization of n-type interdigitated back contact silicon heterojunction (IBC-SiHJ) solar cell structure using Silvaco Tcad Atlas, Sol. Energy 127 (2016) 206–215.
[39] Y. Zhou, G. Long, Low density of conduction and valence band states contribute to the high open-circuit voltage in perovskite solar cells, J. Phys. Chem. C 121 (2017) 1455–1462.
[40] X. Miao, S. Tongay, M.K. Petterson, K. Berke, A.G. Rinzler, B.R. Appleton, A.F. Hebard, High efficiency graphene solar cells by chemical doping, Nano Lett. 12 (2012) 2745–2750.
[41] S.M. Sze, Physics of Semiconductor Devices, 2nd ed., A Willey Interscience Publication, Singapore, 1981.

*2022 2nd International Conference on Intelligent Technologies (CONIT)*
*Karnataka, India. June 24-26, 2022*

# Object Detection using Image Dehazing : A Journey Of Visual Improvement

Ritik Tanwar
*Department of Computer Engineering*
*Delhi Technological University*
New Delhi, India
ritiktanwar_2k18co295@dtu.ac.in

Shubham
*Department of Computer Engineering*
*Delhi Technological University*
New Delhi, India
shubham_2k18co344@dtu.ac.in

Shubham Verma
*Department of Computer Engineering*
*Delhi Technological University*
New Delhi, India
shubhamverma_2k18co348@dtu.ac.in

Manoj Kumar
*Department of Computer Engineering*
*Delhi Technological University*
New Delhi, India
mkumarg@dce.ac.in

2022 International Conference on Intelligent Technologies (CONIT) | 978-1-6654-8407-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CONIT55038.2022.9848085

***Abstract*—Object Detection in hazy conditions is very challenging as haze significantly degrades the visibility of images limits visibility especially in outdoor settings. Here we introduce an interesting method to deal with haze that is present in images. Before applying any object detection method on the hazy input image ,it is needed to be dehaze first and recognised later.For dehazing we have used the an Image Dehazing network known as All-in-One Dehazing Network (AOD-net) which is based on reformulation of atmospheric model and generates clean and clear image through a light-weight CNN and for recognition we have used the third version of famous YOLO i.e. YOLOv3 .We test our method on various real time hazy images and compare the object similarity results on hazy image as well as on dehaze image .Along with this we have compared the number of object which are recognised in hazy image and in output clear image.**

***Index Terms*—Image Dehazing , Object Detection , YOLOv3 ,CNN, AOD-net ,ASM**

## I. Introduction

Artificial object recognition is gaining ground with the rise in automation and accuracy level are approaching to that of human recognition. But the majority of studies have been designed to achieve optimal performance in a clean or laboratory-like environment which ignores the most important environmental factor like weather. Due to different external conditions, the quality of many images taken in the outdoor situations are compromised.Outside Weather conditions such as fog, smoke, rain, and snow have impacted the object recognisation models significantly. Such decline in image quality caused by the weather might pose a major threat to the vision system or the camera's monitoring system's performance. To recognise object in adverse hazy situations is still a objection with many recognition techniques which performs badly on this situation so, there is a requirement to build a robust mechanism which automatically removes the haze from hazy image before going through the recognition part . We are using the atmospheric scattering model as in [1] [2] to model the images with haze. The atmospheric scattering model is given by :

$$I(x) = J(x)t(x) + A(1 - t(x)) \tag{1}$$

where , *I(x)* is Observed hazy image , *J(x)* is Original image , *A* is Global Atmospheric Lighting *t(x)* is transmission matrix. *A* also knowns as global atmospheric lighting refers to the natural light of the atmosphere across the entire scene. *t(x)* represents the amount that how much light is received by the camera from the object. It is calculated as follows:

$$t(x) = e^{-\beta d(x)} \tag{2}$$

where , $\beta$ is Scattering Co-efficient and *d(x)* is distance between the object and observer . The idea behind using these equations is that light after reflected from an item is scattered by the atmosphere before it reaches the optical instrument . The quantity of light dispersed named as $\beta$ is determined by atmospheric characteristics and the object's distance *d(x)* from the instrument. The remaining light is transferred towards the instrument from where it reaches the instrument . The value of *t(x)* is in the range (0, 1) since $\beta$ and *d(x)* are both non-negative. The value 0 indicates that no light from the item reaches the camera (all light is dispersed), and value 1 indicates that all of the object's light reaches the camera i.e. no scattering. Greater the value of $\beta$ more will be the scattering of atmospheric light. Furthermore, when the item travels further away from the camera, the amount of light dispersed increases. The light received by the camera may now be estimated using transmission values (for each pixel) and global ambient light. If *t(x)* equals 1, the camera sees the object's light properly (no scattering). If *t(x)* is 0, the camera just observes the ambient light (complete scattering). Otherwise, the camera perceives a linear interpolation between the object's brightness and the ambient light.

Object detection have made a good amount of advancement in recent years all thanks to the concept of convolution neural networks (CNN) and artificial neural networks (ANN). Object



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:56:19 UTC from IEEE Xplore. Restrictions apply.

detection methods may be classified into two groups : one category is based on using region and another is based on using regression. RCNN [3], Fast RCNN [4], and Faster RCNN [5] are some of the object detection techniques that use region proposals. Object recognition techniques that uses regression, on the other hand, include SSD [7] and YOLO [8], among others. These approaches also build region proposal networks (RPN), however at the time of generation, these region suggestions are divided into categories. All of the approaches discussed above have proven to be effective and efficient in terms of object recognition and localization. YOLO merges labels from various datasets into a tree-like structure, although the merged labels are not mutually exclusive. Using a hierarchical design, YOLO9000 [9] improves YOLO's ability to distinguish targets exceeding 9000 categories. YOLOv3 [10] replaces the approach of estimating the cost function with multi-label classification and shows a significant improvement in discriminating small objects.

The paper is divided as follows. Section II explains about Related work. Section III describes about proposed methodology . Section IV describes about the Datasets used in this paper. Section V describes about the Evaluation and Results obtained. Section VI has Conclusion and Future Work.

## II. Related Work

### A. Traditional Methods

[11] is able to eliminate the haze by setting the contrast to its maximum limit. [12]By measuring the light reflected by the scene, an objectively grounded approach was presented. For clean and clear restored pictures, [13] tightened the border limit and used contextual regularisation. [14] suggested a faster approach for the automated recovery of ambient light. [15] The model parameters were supervised learnt after constructing a colour reduction prior for the foggy picture and a linear model of scene depth. In this experiment, the effective dark channel prior (DCP) is utilised.[16] to more accurately construct the transmission matrix for picture dehazing and then uses SURF algorithm for object detection which rapidly locates a point of interest that remains constant regardless of size or rotation, and is appropriate for picture matching.

### B. Deep Learning Methods

CNNs have had a lot of success with computer vision problems, and they've newly been applied to haze removal. [17] used a multi-scale CNN to construct a coarse-scale transmission matrix that was then improved. [18] presented novel state-of-art model known as DehazeNet, which is trainable for medium transmission estimate. It takes a hazy image as input and produces a transmission matrix as output.Then Following the ASM , dehazeNet when combined with empirically determined global atmospheric light, yields a haze-free picture. [19] uses this DehazeNet for removing haze and then uses Faster CNN for object recognition.It accepts an input picture and presents a enclosing cube along with the probity of object being to the specific group.

Above mentioned ways have a familiar belief that for generating a clean and clear picture from a hazy image , we have to measure a correct medium transmission matrix .The ambient light is calculated separately, and the good image is rebuilt with the formula (1). AOD-Net, on the other hand, is constructed on the premise that the physical model might be formulated in a "more end-to-end" manner, with all of its parameters evaluated in a single unified model. AOD-Net will return the clean dehazed image without any intermediate technique to calculate attributes.

For object recognition both SURF algorithm and Faster CNN have some drawbacks.The SURF algorithm has a high dimension of feature descriptor, a huge amount of computation, and low matching accuracy when there is large angle of rotation and large angle of view.Faster CNN requires numerous passes for each image in order to extract all of the targets. Because of different systems function in a sequential order, the performance of the next operation is determined by the performance of the previous operations. To locate and identify the items in a photograph, this method use region proposal networks. RPNs, on the other hand, do not consider the entire picture since they only utilise the parts of it that have a high probability of containing targets.Faster CNN provides accuracy but is time consuming . Therefore to process in less time we have used YOLOv3 . The targetness score in YOLOv3 is obtained by using logistic regression. It calculates the total score for all targets within each boundary box. Because each class is assigned a logistic classifier, YOLOv3 may provide multilabel classification. YOLOv3 utilises Darknet 53 which have fifty-three levels of convolution

## III. Proposed Methodology

The underlying network architecture can be splitted into two different parts : (1) Image Dehazing Network , (2) Object Detection Network. Our method is as follows : First an input hazy image is sent to object detection network i.e. YOLOv3 to recognise and classify object present in them , then the same input image is transformed to dehazed image using image dehazing network which is inspired from AOD-net . The ouput image after dehazing is again sent to object detection network to recognise and classify objects in it. Thereafter we will compare results coming from both the output images from YOLOv3.

Object detection is employed in a variety of areas of human life, including health and education. Because all of these industries are fast evolving, one-stage models must likewise evolve to meet their needs. The next advanced version of YOLO is version 3, which computes the targetness score using logistic regression. It calculates the total score for all targets within each boundry box. Because it utilises a logistic classifier for each class instead of the softmax layer used in YOLO v2, YOLO v3 may provide multilabel classification. Darknet 53 is used by YOLO v3. It has fifty-three convolutional layers. These layers are more detailed than the darknet 19 layers used in YOLO v2.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:56:19 UTC from IEEE Xplore. Restrictions apply.

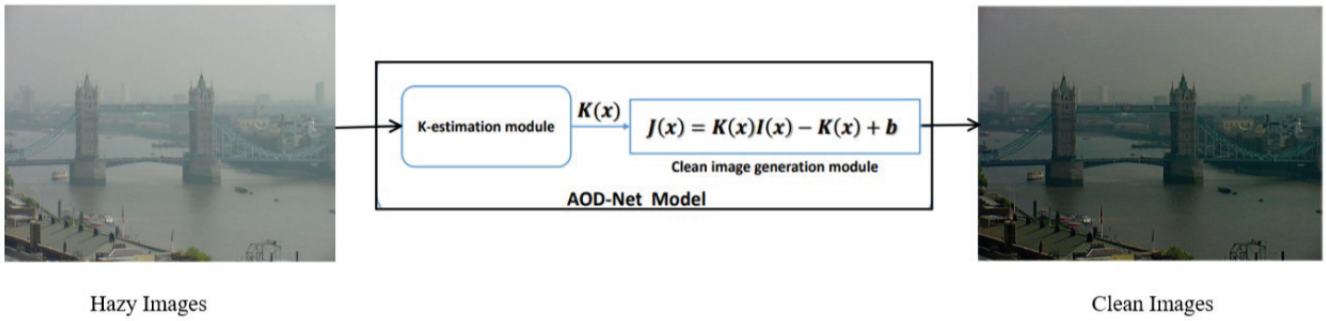



Fig. 1. AOD-net Pipeline

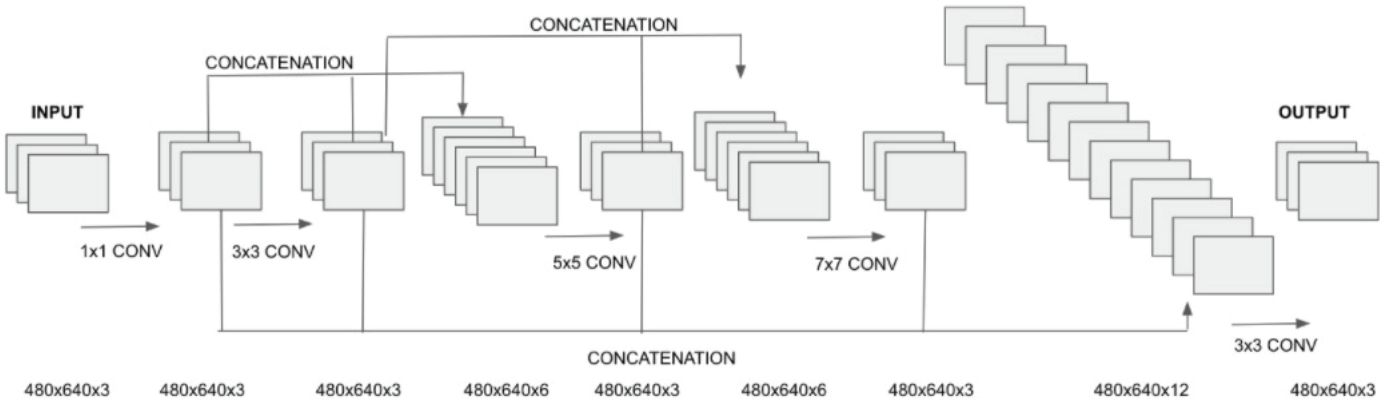



Fig. 2. K estimation module

### A. Image Dehazing Network

In this section the image dehazing netork which we have used to get dehaze image i.e. AOD-net is discussed thoroughly/ First, we'll go through the modified atmospheric scattering model, which is the foundation for the AOD-Net. The AOD-Net structure is then discussed in depth.

*1) Problem Formulation:* Existing dehazing research relies on calculating *t(x)* and *A* separately. The primary issue here is that estimating inaccuracies can accumulates. With the following re-formulation of equations, AOD seeks to estimate both parameters in a unified manner (1):

$$J(x) = K(x)I(x) - K(x) - b \tag{3}$$

where,

$$K(x) = \frac{\frac{1}{t(x)}(I(x) - A) + (A - b)}{I(x) - 1} \tag{4}$$

t(x) and A are combined into a single variable K(x), which is altered by the input I. (x). Once we have measure the value of K(x) or K-estimated module we can generate a clean and clear dehazed image . Therefore , the primary goal now is to estimate K(x) module in such a way that the MSE between the output dehazed image and original image is as minimal as possible. This is accomplished by the use of a CNN.

*2) Network Design:* The used image dehazing network consists two modules.The first module is *K-estimation module* module which uses five convolutional layers to estimate the value of *K(x)* and second module is *clean and clear image generation module* which consists of several element-wise operation of multiplication and addition of layers to recover clear images using the equation (4).

The K-estimation module is heart of this network as it is responsible for indirectly estimation atmospheric light and transmission matrix. The network consists of five convolutional layers with each of layer activsted with ReLu or rectified linear function . For an image of size *W X H X 3*, *K(x)* also has size *W X H X 3*.To produce multi-scale features, each layer has three filters but a distinct kernel size. Additionally, preceding layer activations are concatenated with intermediate layer activations before convolutions to compensate for information loss during convolutions. Fig 2 states the whole architecture of network. The process initiated by sending the input 1st convolutional layer *conv1* of size 1x1 . Inspiring from [17] *concat1* layer concatanates from layers *conv1* and *conv2*. Similarly *concat2, concat3 and so on* are concatenated from *conv2*



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:56:19 UTC from IEEE Xplore. Restrictions apply.

, *conv3* and respectively so on.The intermediate connections and this multi-scale architecture, which captures characteristics at many sizes compensate for the information loss during convolutions. By doing this we will get K-estimation value or *K(x)*.

Following the estimation of *K(x)* we can use the equation (3) to calculate of value *J(x)* which will results in generation of clean and clear image.

### *B. Object Detection Network*

YOLOv3 model for object detection. Object identification is a basic relapse issue for YOLO, which learns enclosed box coord. and rank possibilities from an input image.

*1) YOLOv3 Pipeline:* It uses input set of img. with dimensions Mx416x416x3 . This image is passed to a convolution neural network where the last two dimensions of output are reduced to get an final output volume of (19, 19, 425) . every cell of 19 x 19 grid always gives 425 values .The no. of hook cubes each grid is 5, hence 425 = 5 * 85. 85 = 5 + 80, here five is ($p_c$, $b_x$, $b_y$, $b_h$, $b_w$) and eighty is the class count to detect.The result is a list of bounding cubes together with the classes that have been found. Six numbers are assigned to each enclosed cube ($p_c$, $b_x$, $b_y$, $b_h$, $b_w$, c). Each bounding box is represented by 85 values when c is expanded into an 80-dimensional vector. Finally, Intersection over Union (IOU) is worn to ignore generating overlapping cubes and Non-Max Suppression techniques which assists in overcoming the issue of recognizing an object in an image several times. It is done by choosing boxes with max. probity and ignoring low probity boxes nearby that are less than the predefined cut-off value.

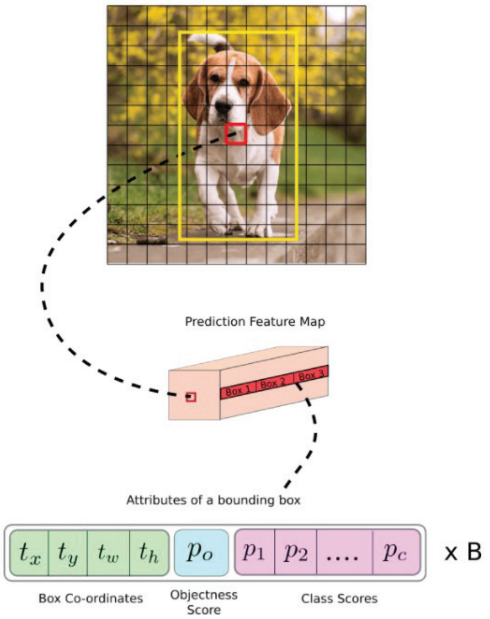



Fig. 3. Image Grid

The formulas which are written below explain how network output transform for obtaining bounding box predictions. Here, $d_x$ and $d_y$ are center coordinates,$d_w$is width and $d_h$ is the height of predicted result. Top left coordinates of the grid are$m_x$and $m_y$. Network outputs are $t_x$,$t_y$, $t_w$ and $t_h$. Anchor dimensions for the box are $p_h$ and $p_w$.

$$d_x = \sigma(t_x) + m_x \tag{5}$$

$$d_y = \sigma(t_y) + m_y \tag{6}$$

$$d_w = p_w^{e^{t_w}} \tag{7}$$

$$d_h = p_h^{e^{t_h}} \tag{8}$$

Each image is divided into S x S grids, with N bounding boxes and certainty guessed for each grid. The bounding box's accuracy, as well as whether the bounding box genuinely encompasses an item, are represented by the confidence (regardless of class). Model also gives the category value for every box in every class during the training phase. To calculate the chance of each class occurring in a predicted box, you can combine the two classes.

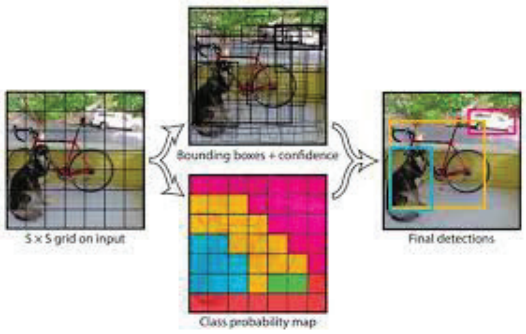



Fig. 4. Yolo Object Detection Flow

As a result, total SxSxN boxes are anticipated. However, the most of these boxes have low confidence scores, and if we establish a threshold, say 30 percent confidence, we can delete the majority of them, as seen in **Fig. 4.**.

*2) YOLOv3 Network Architecture:* YOLOv3 employs use of a Darknet variation that was originally trained on Imagenet and has a 53-layer network. YOLO v3 has a 106-layer fully convolution basic architecture, with 53 extra layers added on top for detection. Using 1x1 disclosure kernels on characteristic mapping of three distinct sizes at three distinct points throughout the net., this third iteration of YOLO disclosure is produced. The detection kernel has the structure 1x1x(Bx(5+C)). where B is the number of bounding boxes that a feature map cell can predict, "5" denotes the four main bounding box properties, and "1" represents the confidence level, and C is the representation of the no. of groups. The kernel size of YOLO v3 trained on COCO is 1 x 1 x 255 since B = 3 and C = 80. Assume we have a 416 × 416 input picture. YOLO v3 model gives prediction at three different scales, which are calculated by downsampling the input image's dimensions by 32 pixels, 16 pixels, and 8 pixels, respectively.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:56:19 UTC from IEEE Xplore. Restrictions apply.

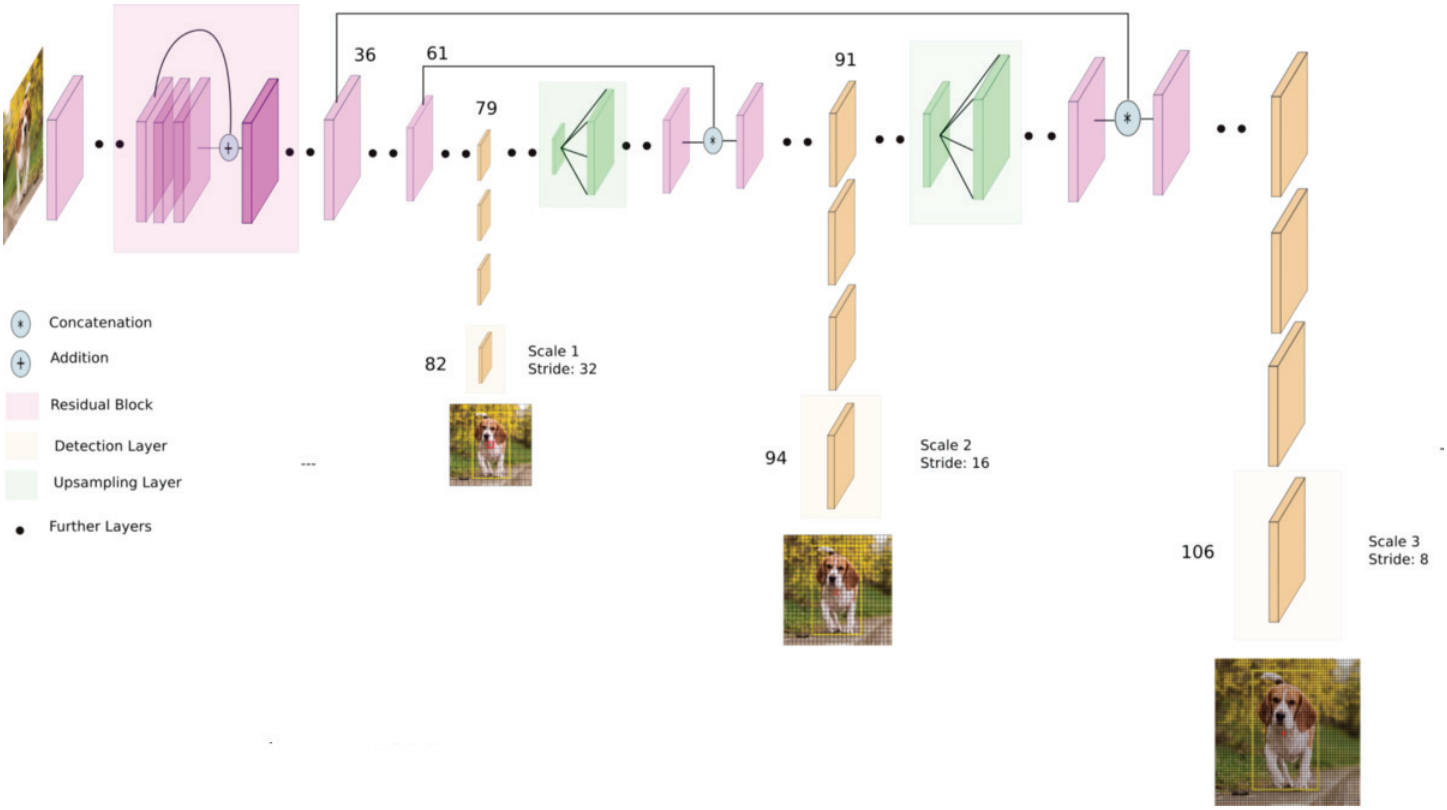



Fig. 5. **The Architecture**. YOLO v3 is based on a Darknet variation that used to have a 53-layer net. trained on Imagenet. In order for recognition, fifty three successive outlets are put on top, giving YOLO v3 a hundred and six layer deep connected underlying architecture.

Initial detection is performed by the 82nd layer. The network down samples the 1st 81 layers of the picture, with the 81st layer having a stride of thirty two. The final feature map will be 13 × 13 if we have a 416 x 416 image. We perform one detection using the 1 x 1 detection kernel, resulting in a 13 x 13 x 255 detection feature map. The feature map from layer 79 is subjected to a few convolution layers before being up sampled by two times to 26 x 26 dimensions. After that, the layer 61 characteristic map is depth combined with this characteristic map. After that, the combined feature maps are fed into numerous 1 x 1 convolution layers, which fuse the values from the former layer (61). The 94th layer then performs the 2nd detection, giving in a disclosure characteristic map of 26x26x255. Before being depth combined with a feature map from layer 36, the feature map from layer 91 is processed to a few number of convolution layers. With effect from the previous layer, a few 1x1 convolution layers are added to include the info (36). We create the final of three feature maps at the 106th layer, providing a 52 x 52 x 255 feature map.

## IV. DATASETS

Artificially hazed pictures are used for the training phase of the Net. For this purpose we are using NYU2 Depth Database [20] which stores depth information of each and every pixel.It contains hazy images in outdoor, urban , wild, indoor, day and night environments . Images were artificially hazed with several values of $\beta$ and $A$ which are selected randomly from uniform distribution based on equation (1) .They used a fog system to generate the haze, then recorded pictures of the identical subject at distinct degrees of haze (a overall of ten) ranging from zero haze to entirely wrapped in haze. Moreover for about 1500 groundtruth images there were 27K hazed images which are obtained.The model was trained on a total of ten epochs. Weight decay was fixed to 0.0001 and weights were set with a regular Gaussian distribution. The gradient norms were set at 0.1. We employed a learning rate of 0.0001 and Adam Optimizer to speed up the gradient descent process by taking into consideration the 'exponentially weighted avg. of the gradients in our implementation.

For training YOLO v3 model we have used COCO dataset [21] which contains contains challenging, high-quality visual datasets for computer vision, mostly state-of-the-art neural networks.The COCO dataset is well-labeled, provides data to train supervised models that are able to identify the common objects in the dataset. Moreover COCO dataset consists of 80 different object categories some of which are people , car , dog , trucker etc.

## V. EVALUATION AND RESULTS

On doing analysis based on various metrics ,and going through some tabular data ,that is created based on our observation and results of the output that we have received from our model,we are in a position to discuss the results that



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:56:19 UTC from IEEE Xplore. Restrictions apply.

Fig. 6. Images with haze levels of 1 to 10

we have been able to observe, the hazy images that we used as input shown some interesting results when dehazed ,the results varied depending on the levels of haziness in the image, for results one can refer to the image comparison in *fig.7.*

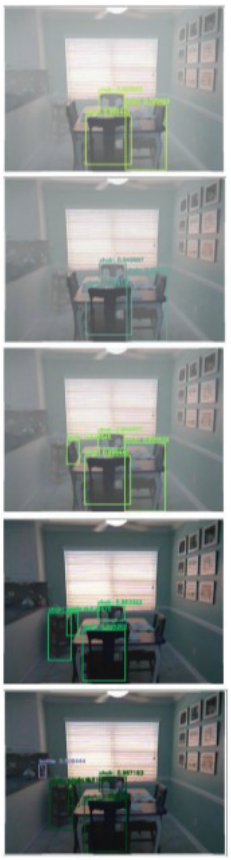


Fig. 7. Number of objects recognised at different hazy levels

From the table 1 and fig.7 we can see a glimpse of what wonders can dehazing model can do to our object Detection model, the results of object recognition get better and are more visible with decreasing haze level. As one can deduce from the

TABLE I
NUMBER OF OBJECTS DETECTED AT VARIOUS HAZE LEVEL.

| Haze Level | Objects Detected in hazy image | Objects detected in Dehazed Image |
|---|---|---|
| 5 | 3 | 5 |
| 4 | 3 | 5 |
| 3 | 4 | 5 |
| 2 | 4 | 5 |
| 1 | 5 | 5 |

observations in the table.1 , that with increasing level of haze in the input image, can make object Detection algorithm less effective . As one can see of haze level 5, the object Detection algorithm can only detect 3 object in an image with haze, as compared to Dehazed Image,where number of object detected are 5, trends follow in case , when we decrease the haze level to 4, that is object Detection algorithm can only detect 3 object in an image with haze, as compared to Dehazed Image,where number of object detected are 5. But as soon as we reach haze level 3 ,we see some improvement as number of detected objects for hazy image is increased from 3 to 4, this results is seen again for haze level 2,but as we reach haze level 1 ,i.e the level at which the haze is negligible or image is dehazed , there we see most number of objects detected so far i.e 5, which is equals to number of object detected for dehazed image, from our observation ,we can conclude that there is significant improvement in object detection functionality of our algorithm as haze decreases. So therefore in dehazed image number of objects recognised is more as compared to that of dehazed image.

In Fig. 8 , we are witnessing the real usage of image-dehazing in object-detection using real world examples, one can easily decipher the riddle of how new objects are detected in Dehazed image as we mention and discuss the number associated with the images in Table 2. So let's begin with Image 1, here we can see that in hazy image ,our algorithm successfully detects the object ,i.e a bed with a accuracy of about 97% ,whereas the object-detection Algorithm detects the same object with an accuracy of about 99% ,that is nearly 2%



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:56:19 UTC from IEEE Xplore. Restrictions apply.

Fig. 8. Object Detection on hazy Image (left-side) and respective Dehaze image (right-side)

TABLE II

| S.No | Objects recognized in hazy image | Similarity percentage of each object | Object recognized in dehazed image | Similarity percentage of each object | Diff in accuracy |
|---|---|---|---|---|---|
| 1 | 1 | Bed 97% | 1 | Bed 99% | 2% |
| 2 | 3 | Chair1 55%<br>Chair2 65%<br>Book 74%<br>Chair3 (Not detected)<br>Chair4 (Not detected)<br>Chair5 (Not detected) | 6 | Chair1 92%<br>Chair2 92%<br>Book 72%<br>Chair3 64%<br>Chair4 80%<br>Chair5 88% | 47%<br>37%<br>2%<br>64%<br>80%<br>88% |

better as compared to the hazy image. Similarly for Image 2 where we can an image with clear multiple objects,our object-detection Algorithm detects 3 objects in case of hazy image, the objects are 2 chairs with accuracy of 65% for near one and 51% for farther one , and a book with an accuracy of 74% , whereas for dehazed image ,the algorithm detects 6 objects including those which are previously detected in hazy image. let's compare results for this observation those two chairs which were detected before are now detected with accuracy of 92% each ,i.e an improvement of about 27% and 41% respectively. also book is also detected with an accuracy of 72%, this abnormal behaviour ca be explained by the fact that is haze is decreased the object appears less like a book and more like bunch of paper or napkins ,which leads to decrease of nearly 2% in its identification.Other than that the algorithm also detected 4 new chairs in the dehazed image with an accuracy of about 92% ,80%,84% and 88% clockwise starting from bottom one. which leads us to the conclusion that our object-detection algorithm (YOLO V3) performs better if we dehaze a hazy image and hence gives better results for real world scenarios.

## VI. Conclusion and Future Work

In conclusion of this paper, we have successfully found a way to improve the extent of object-detection in hazy images also we have shown the quantitative results so obtained, by comparing the results received from an image of hazy environment and results received from dehazed version of the same image, the results were both surprising and insightful. As we come across many scenarios where the resultant image has less accuracy as compared to that of hazy Environment ,that's because Dehazing causes images to darken, which makes recognition more difficult. This has left a unfavourable brunt on the picture's small element or objects.

Actually our main use case is using this in security cameras. We are planning to find a way of finding haze estimation in an image which depending on the threshold value will tell us whether the given image is needed to dehaze or not. This approach will make dehazing network to decide whether to dehaze an image or not when there is a modest amount of haze



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:56:19 UTC from IEEE Xplore. Restrictions apply.

it will directly undergo recognition reducing the computational time of removing haze.Also we will try to include videos haze removal which will reduce the haziness in each an every frame of the video .

## REFERENCES


[1] McCartney, Earl J, "Optics of the atmosphere: scattering by molecules and particles.", New York, John Wiley and Sons,Inc., 1976. 421 pp., 1976..

[2] Narasimhan, Srinivasa G., and Shree K. Nayar, "Vision and the atmosphere", International journal of computer vision 48, no. 3 (2002): 233-254.

[3] Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587. 2014, doi: 10.1109/CVPR.2014.81.

[4] R. Girshick, "Fast R-CNN," Proc. IEEE Int. Conf. Comput. Vis., vol.2015 International Conference on Computer Vision, ICCV 2015, pp. 1440–1448, 2015, doi: 10.1109/ICCV.2015.169.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards RealTime Object Detection with Region Proposal Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2017, doi:10.1109/TPAMI.2016.2577031.

[6] P. Dong and W. Wang, "Better region proposals for pedestrian detection with R-CNN," 30th Anniv. Vis. Commun. Image Process., pp. 3–6,2016, doi: 10.1109/VCIP.2016.7805452.

[7] W. Liu, D. Anguelov, D. Erhan, and C. Szegedy, "SSD: Single Shot MultiBox Detector," ECCV, vol. 1, pp. 21–37, 2016, doi: 10.1007/978-3-319-46448-0.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real- time object detection," IEEE Comput. Soc. Conf.Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 779–788, 2016, doi: 10.1109/CVPR.2016.91.

[9] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR, vol. 2017-Janua, pp. 6517–6525, 2017, doi: 10.1109/CVPR.2017.690.

[10] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement,"arXiv Prepr., 2018.

[11] R. T. Tan, "Visibility in bad weather from a single image", In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1–8. IEEE, 2008.

[12] R. Fattal, "Single image dehazing", ACM transactions on graphics (TOG), 27(3):72, 2008.

[13] G. Meng, Y. Wang, J. Duan, S. Xiang, and C. Pan, "Efficient image dehazing with boundary constraint and contextual regularization", In Proceedings of the IEEE international conference on computer vision, pp.617–624, 2013.

[14] M. Sulami, I. Glatzer, R. Fattal, and M. Werman, "Automatic recovery of the atmospheric light in hazy images," In Computational Photography (ICCP), 2014 IEEE International Conference on, pp. 1–11. IEEE, 2014.

[15] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior", IEEE Transactions on Image Processing, 24(11):3522–3533, 2015.

[16] H. Ann,J.H. Lee , and H.J. Cho, "Research of improved object detection with image dehazing" ,In International Journal of Science and Technology ,pp.76-79 .2018 ,doi: 10.14419/ijet.v7i2.12.11039

[17] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, :Single image dehazing via multi-scale convolutional neural networks", In European Conference on Computer Vision,pp.154–169. Springer, 2016.

[18] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet:An end-to-end system for single image haze removal", IEEE Transactions on Image Processing, 25(11), 2016.

[19] Nalla, Bhanu Teja, Teena Sharma, Nishchal K. Verma, and S. R. Sahoo. "Image dehazing for object recognition using faster RCNN." In 2018 International Joint Conference on Neural Networks (IJCNN), pp. 01-07. IEEE, 2018.

[20] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images", In European Conference on Computer Vision, pp. 746–760. Springer, 2012.

[21] Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In European conference on computer vision, pp. 740-755. Springer, Cham ,doi: 10.48550/arXiv.1405.0312 ,2014




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:56:19 UTC from IEEE Xplore. Restrictions apply.

# Optimised Algorithm for Step Count Estimation Using Sensor Data from Smartphones and Wearables

Apratim Shrivastav
*Delhi Technological University*
New Delhi, India
apratimshrivastav@gmail.com

Utkarsh Kuchhal
*Delhi Technological University*
New Delhi, India
kuchhalutkarsh17@gmail.com

Samarth Singhal
*Delhi Technological University*
New Delhi, India
samarth.a12d@gmail.com

Tanmay Jain
*Delhi Technological University*
New Delhi, India
tanmayj2020@gmail.com

Divyashikha Sethia
*Delhi Technological University*
New Delhi, India
divyashikha@dtu.ac.in

Vidushi Chaudhary
*Samsung R&D Institute*
Noida, India
vidushi.c@samsung.com

Rajiv Nigam
*Samsung R&D Institute*
Noida, India
rajeev-nigam@samsung.com

Ashish Kumar Namdeo
*Samsung R&D Institute*
Noida, India
a.namdeo@samsung.com

2022 IEEE Region 10 Symposium (TENSYMP) | 978-1-6654-6658-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/TENSYMP54529.2022.9864395

***Abstract*—Accurate step detection and counting are essential for many health applications and estimating distances for indoor positioning services. However, the accuracy of the step estimation algorithms in various systems varies due to problems with over or under-counting the steps. Osaka University - Institute of Scientific and Industrial Research (OU-ISIR) gait database is one such dataset where the prior step count estimation algorithm lacks accuracy in prediction. This paper proposes an improved step count estimation algorithm using the OU-ISIR dataset. The algorithm involves normalization and peak detection of the analog signal from accelerometers to achieve near-perfect accuracy. The proposed work also reviews several machine learning and deep learning approaches and compares them to the proposed algorithm. Results indicate that the proposed algorithm achieves higher accuracy than the previous step estimation algorithm and other machine learning and deep learning approaches.**

***Index Terms*—Step Detection, Peak Detection, Machine Learning, Gait Analysis**

## I. Introduction

Over the past decade, the use of mobile phones and wearables have had broad impacts in terms of development in social welfare [1], economic development [2], and accessibility towards services previously inaccessible [3]. The growth of mobile technology has paired with significant improvements in the sensor technology that accompanies these devices, this includes an increase in the number of communication modules Global Positioning System (GPS) and sensor modules such as accelerometer, gyroscope, and magnetometer [4]. It has provided us with more significant resources at our disposal to allow for a deeper research into the technology and an incentive to explore greater number of applications. In recent times, a critical use case of smartphones has been apps and services linked to monitor healthcare and provide users with data that enables them to live healthier lives. Step counting detects the number of walking steps taken by a person. It comes under the umbrella of human activity recognition and forms the foundation for many essential applications involving calculating calories burnt, estimating distances, and foundational gait analysis. By using gait analysis, developers can take advantage of its technology to solve world-class issues, including but not limited to identifying individuals, analyzing Parkinson's diseases, or even recognizing physical activities. Alternate use-cases of step detection involve pairing it with the GPS to allow for increased accuracy for indoor localization of the users. Previously, several researchers have worked on accurate step measurement [5]–[11]. There are several challenges in the process of estimation of step counter computation [12] as follows:

1) *Human behavior*: Humans tend to perform multiple tasks during walking at the same time. This combination of movements disrupts sensor data resulting in noise leading to issues in differentiating steps from other user movements.
2) *Noisy sensor data*: A large number of sensors in modern devices lead to collecting data that contains much noise, posing a challenge to accurate step count estimation

The problem with current step detection programs have been a lack of accuracy and several of them suffer from the

978-1-6654-6658-5/22/$31.00 ©2022 IEEE



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:51:50 UTC from IEEE Xplore. Restrictions apply.

problem of over or under-counting the steps. It has especially been the case with deployed software due to a lack of accurate light algorithms used on wearable technologies and smartphones. The technology remains inaccessible to a vast majority of people. Hence, we believe the broader community will benefit from the research done towards creating light algorithms for step detection. In order to train and evaluate our models, we leverage Osaka University - Institute of Scientific and Industrial Research (OU-ISIR) dataset [13], which uses four sensors:

- Three Inertial Measurement Unit (IMU) sensors, each IMU includes a triaxial accelerometer and a triaxial gyroscope
- One smartphone Motorola ME860, which includes only a triaxial accelerometer

Sensors located around the waist area of the participants were instructed to walk, after which their sensor readings were recorded in the dataset.

The step count estimation algorithm proposed by Salvi *et al.* [14] on peak identification was hampered mainly by the noise in the dataset, varying magnitudes of peaks caused by differing walking speeds, and difficulties in determining thresholds. Existing research that focuses on step detection is computationally demanding and poses difficulty in the real-time calculation, which is a requirement for step counting devices.

The main contributions of this paper are to design and implement a novel algorithm for step count estimation that is blind to the intensity of human walk using the OU-ISIR dataset [13] with close to ideal accuracy. This work also uses machine learning and deep learning approaches to estimate the step counts. Results indicate that the proposed step count estimation algorithm performs better than the other approaches.

The rest of the paper comprises Section II, which provides some insight into the related work done within the field. Section III describes the dataset and explains the different implemented algorithms and the proposed improved step count estimation algorithm. Section IV discusses results and outcomes. The paper finally presents the conclusion and future work in Section V.

## II. Related Work

There have been several prior studies that have used different methods for step counting. Previous work proposed for step count estimation does not achieve accuracy. We reach and solve the problem of detecting steps of different intensities. We have distilled the existing studies into three distinct categories depending on the approach and computational approach.

### A. Thresholding Approach

The thresholding approach judges whether the sensors' data satisfy some predefined thresholds. Step count happens when the value exceeds the set threshold. Rai *et al.* proposed a step count estimation algorithm where the device is tied to the subject's ankle, and it counts a step when the sensor value exceeds some predefined threshold [15]. In comparison, this is a straightforward approach to step detection. It is prone to errors because predicting the single threshold for every gait state is improbable, primarily when using devices in unconstrained environments.

### B. Existing ML and DL Approaches

Several machine learning techniques propose to predict steps using sensor data. Leonhardt *et al.* [16] used the hidden Markov model and achieved a median error of about 1.3%. However, the model needed enormous computational complexity to achieve this accuracy, making it unsuitable for most devices and processes. Several other approaches, including the autocorrelation approach [15], suffer from the same problem of higher computational complexity.

### C. Peak Detection

Peak detection relies on calculating peaks generated by sensory data. Much of prior work dealing with peak detection suffers due to problems with noise, a different magnitude of peaks created due to different walking speeds, and problems with setting thresholds to eliminate false peaks. For example, in references [16], [17] authors use low pass filters to reduce noise but do not account for different walking speeds leading to errors. Salvi *et al.* [14], used windowed peak detection for step detection and achieved an error percentage of 3.54% to 14.04%. However, it suffers from overcounting due to noise and the inability to normalize $x$, $y$ and $z$ values for acceleration into a single signal, resulting in more significant errors with the final number of counted steps.

## III. Dataset and Proposed Method

### A. Dataset

This work uses the OU-ISIR dataset [13], which contains recorded tri-axial accelerometer data collected from 744 subjects (389 males and 355 females) aged 2 to 78 years. Data is individually collected from each of the three accelerometers and stored in different folders. Since more sensors would add to the complexity of the model, we only use the accelerometer sensor for analysis. The model cleans the data to remove sensor data irrelevant to the study. It removes the rows with negative class labels that are not a part of the data-collection activity. Dataset does not include false positives.

### B. Proposed Algorithm

The step count estimation algorithm proposed by Salvi *et al.* [14] does not normalize the acceleration across the three different axis, which is included as a step in our algorithm to prevent the dominance of acceleration in a particular direction. It results in lesser noise reduction in the filtering stage of the algorithm. This paper aims to deal with three primary problems with this algorithm:

1) To figure out the least disruptive way (slightest noise) to normalize acceleration on the three axis.
2) To find the best method to reduce noise and avoid false peaks and additional counting of steps.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:51:50 UTC from IEEE Xplore. Restrictions apply.

3) To calculate steps reliably regardless of walking intensity.

In order to deal with the problems mentioned above, we propose an algorithm as illustrated in Fig. 1. The dataset comprises readings from a tri-axial accelerometer. It is the same accelerometer used commercially in mobile phones and wearable devices. The device outputs the acceleration in its $x$, $y$ and $z$ components. In order to prevent a directional dependent algorithm, the proposed new step count estimation algorithm disregards distinct graphs for each of the three acceleration components *i.e.* $a_x$, $a_y$, $a_z$. The algorithm calculates the magnitude of the three components as indicated in equation (1).

$$|a| = \sqrt{(a_x)^2 + (a_y)^2 + (a_z)^2} \tag{1}$$

where |a| is the magnitude of acceleration and $a_x$, $a_y$ and $a_z$ are $x$, $y$ and $z$ components of acceleration, respectively.

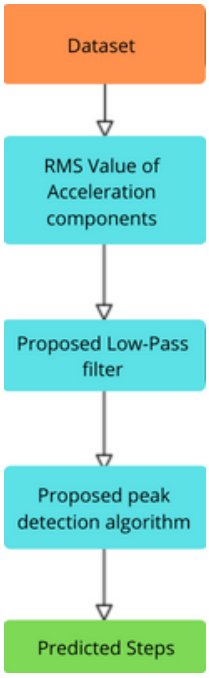



Fig. 1. Methodology for proposed algorithm

This allows us to plot a singular graph for the magnitude of the components of accelerations instead of plotting three distinct graphs for the magnitudes of acceleration. However, the plot still contains a large amount of noise. The peak detection algorithm must ignore the multiple peaks before counting the peaks. This work proposes a customized version of the butter-worth filter (low-pass filter) [18] to solve the problem. We use the low pass filter as it provides a smoother form of a signal removing short-term fluctuations. The model chooses a sampling frequency of 3Hz to provide the best result in noise reduction. As shown in Fig. 2, the resulting signal reduces all noise and detects peaks.

Finally, the algorithm sidesteps the different magnitudes of peaks generated from different walking styles, speeds,

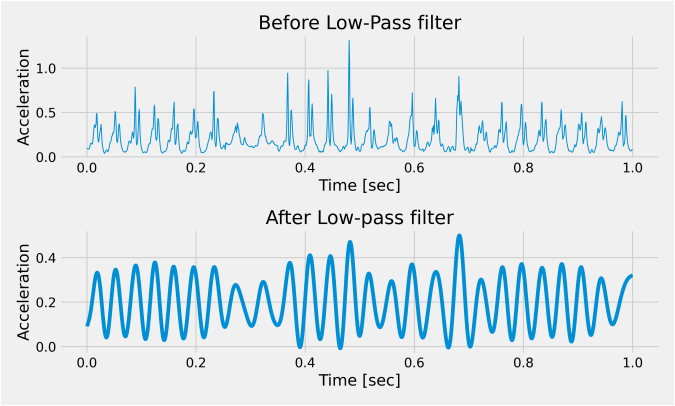



Fig. 2. Before and after applying the low pass filter

and conditions. The proposed step count estimation algorithm aims to localize peaks on the $x$-axis, as shown in Algorithm 1. It avoids the problem of classifying motions for accurate step counting.

**Principle of Proposed Step Count Estimation Algorithm**: *An array 'a' where the 'i'th element of array 'a' represents the normalized value of the accelerometer reading at time 't' with 'dt' interval between the 'i'th and '(i-1)'th element. Our Peak would be detected if a[i-1] < a[i] && a[i] > a[i+1]*

**Algorithm 1** Proposed Step Count Estimation Algorithm

**Require:** X = X axis accelerometer data
Y = Y axis accelerometer data
Z = Z axis accelerometer data
**Ensure:** data = preprocessed data
**while** $i <$ X.length() **do**
  data[i] = $\sqrt{(a_x)^2 + (a_y)^2 + (a_z)^2}$
  $i+=1$
  steps = 0
  data = Filter(data) // to smoothen our plot
  $i$ = 1
  **while** $i+1 <$ data.length() **do**
    **if** data[$i$-1] < data[$i$] && data[$i$] > data[$i$+1] **then**
      steps $+=1$
      Return steps

Fig. 3 illustrates detection of the peaks as indicated by the red dots by the proposed step count estimation algorithm. Each dot represents an increment in the algorithm's counter for number of steps.It helps overcome shortcomings in previous implementations and reach perfect accuracy when applied to the dataset.

### C. ML and DL Approaches

In addition to peak detection techniques, this paper examines various machine learning-based classification approaches using the OU-ISIR dataset [13]. The ML model first removes all the timestamps at which the invalid and null steps occur and retain only the reasonable steps. Fig. 4 illustrates the


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:51:50 UTC from IEEE Xplore. Restrictions apply.

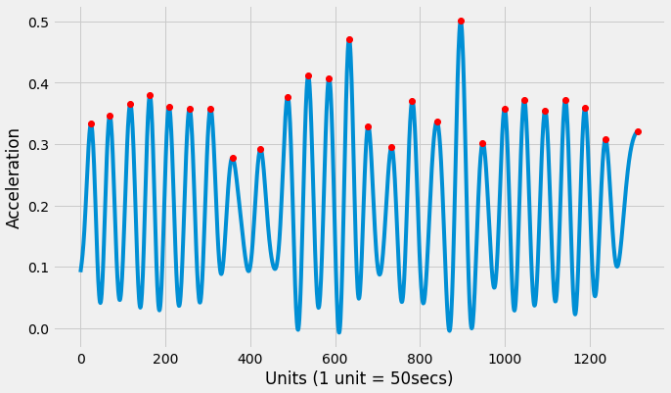



Fig. 3. Peaks detected in the proposed algorithm

methodology followed for machine learning. It comprises an input stage that involves feeding and pre-processing the dataset. The next phase is feature engineering for training. The model calculates 16 features using $a_x$, $a_y$, and $a_z$ each, such as mean, variance, standard deviation, median, RMS, max, min, max value index, min value index, the interquartile range, kurtosis, simple moving average, energy, data entropy, skewness, mean absolute deviation, the relation between $a_x$ and $a_y$, relation between $a_y$ and $a_z$, and the relation between $a_z$ and $a_x$.

The model applies a sliding window approach to calculate these features. Sliding window reduces the time complexity, and we need to calculate the features repeatedly for a set of ranges from the dataset. The model sets the window size as 40 based on the experimentation. It removes those steps which do not have enough data around them for the feature extraction. It extracts the features and stores them up to 5 places of decimal. The model computes the 16 features separately for $a_x$, $a_y$, and $a_z$ as mentioned above.

As the extracted features have different ranges, the model applies normalization to train and test data to convert them to a standard scale. It splits the dataset into training and testing with a ratio of 3:1.

The work compares different machine learning classification models such as random forest, KNN, CNN, SVM, and logistic regression with our proposed step count estimation algorithm. Random forest uses 200 estimators and uses Ginni as the criterion. SVM uses the RBF kernel for classification. For Naive Bayes, we use multinomial Naive Bayes.

This work also compares the results using the CNN approach proposed by Lin *et al.* [19]. The CNN architecture contains three main layers: two separate one-dimensional convolution layers and a dense layer to produce the prediction. It comprises two additional layers to use as the first and last layers in this network. The first reshape layer reshapes the input data into a three-dimensional format suitable for the first convolution. The flattening layer flattens the output of the second convolution layer to change the shape of the feature map to a single dimension and then be input to the fully connected dense layer.

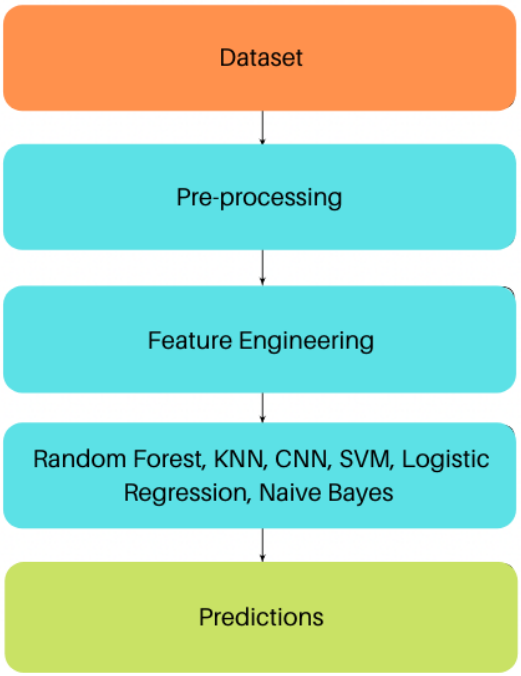



Fig. 4. Methodology of the machine learning algorithm

## IV. Implementation and Result

This work uses distinct methods on the OU-ISIR gait database, and inertial sensor dataset [13] to predict accurate step counting. This work is performed over Google Colab. Table I shows the results for the different algorithms.

TABLE I
Comparison of Accuracy for Different Algorithms

| Technique | Accuracy |
|---|---|
| Proposed Step count estimation algorithm | 99.9 |
| Step count estimation algorithm by Salvi *et al.* [14] | $93.0\% \pm 1.3\%$ |
| Support Vector Machine | 96.53 |
| Random Forest | 96.04 |
| K-Nearest Neighbors | 96.04 |
| Logistical Regression | 91.09 |
| Naive Bayes | 86.39 |
| Convolutional Neural Network | 95.12 |

Results in Table I indicate that the proposed step count estimation algorithm functions optimally in addition to being computationally inexpensive. It performs better than the algorithm proposed by Salvi *et al.* [14]. The alternate machine learning or deep learning-based methods have accuracies between 86.39% to 96.53%, which are less than the proposed step count estimation algorithm.

## V. Conclusion and Future Work

This paper has presented a highly accurate and computationally inexpensive step count estimation algorithm based on triaxial accelerometer data. The proposed algorithm successfully deals with the problem of figuring out the least disruptive way (slightest noise) to normalize acceleration on


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:51:50 UTC from IEEE Xplore. Restrictions apply.

the three-axis. It also reduces noise to eliminate false peaks and overcounting of steps regardless of the intensity of the steps.

In the future, there can be a curation of the dataset to investigate an array of different types of surfaces considering various other edge conditions. Incorporation of advanced deep learning models to achieve better accuracies than our method can be developed.

## VI. Acknowledgement

We thank Samsung India Electronics Private Limited Noida for facilitating Samsung Innovation Campus, Research Lab at Delhi Technological University (DTU), in which we pursued this research work.

## References

[1] D. Björkegren, "The adoption of network goods: evidence from the spread of mobile phones in rwanda," *The Review of Economic Studies*, vol. 86, no. 3, pp. 1033–1060, 2019.

[2] J. C. Aker and I. M. Mbiti, "Mobile phones and economic development in africa," *Journal of economic Perspectives*, vol. 24, no. 3, pp. 207–32, 2010.

[3] K. Locke, L. McRae, G. Peaty, K. Ellis, and M. Kent, "Developing accessible technologies for a changing world: understanding how people with vision impairment use smartphones," *Disability & Society*, pp. 1–18, 2021.

[4] T. T. Ngo, Y. Makihara, H. Nagahara, Y. Mukaigawa, and Y. Yagi, "Similar gait action recognition using an inertial sensor," *Pattern Recognition*, vol. 48, no. 4, pp. 1289–1301, 2015.

[5] F. Gu, K. Khoshelham, J. Shang, F. Yu, and Z. Wei, "Robust and accurate smartphone-based step counting for indoor localization," *IEEE Sensors Journal*, vol. 17, no. 11, pp. 3453–3460, 2017.

[6] X. Kang, B. Huang, and G. Qi, "A novel walking detection and step counting algorithm using unconstrained smartphones," *Sensors*, vol. 18, no. 1, p. 297, 2018.

[7] Y. Feng, C. K. Wong, V. Janeja, R. Kuber, and H. M. Mentis, "Comparison of tri-axial accelerometers step-count accuracy in slow walking conditions," *Gait & posture*, vol. 53, pp. 11–16, 2017.

[8] H. Zhang, W. Yuan, Q. Shen, T. Li, and H. Chang, "A handheld inertial pedestrian navigation system with accurate step modes and device poses recognition," *IEEE Sensors Journal*, vol. 15, no. 3, pp. 1421–1429, 2014.

[9] C. B. Thorup, J. J. Andreasen, E. E. Sørensen, M. Grønkjær, B. I. Dinesen, and J. Hansen, "Accuracy of a step counter during treadmill and daily life walking by healthy adults and patients with cardiac disease," *BMJ open*, vol. 7, no. 3, p. e011742, 2017.

[10] L. Wang, T. Liu, Y. Wang, Q. Li, J. Yi, and Y. Inoue, "Evaluation on step counting performance of wristband activity monitors in daily living environment," *IEEE Access*, vol. 5, pp. 13 020–13 027, 2017.

[11] P. Van Thanh, A.-D. N. Thi, Q. T. T. Thuy, D. C. T. Phuong, V. H. Mau, and D.-T. Tran, "A novel step counter supporting for indoor positioning based on inertial measurement unit," in *International Conference on Integrated Circuits, Design, and Verification (ICDV)*. IEEE, 2017, pp. 69–74.

[12] W. W. Myo, W. Wettayaprasit, and P. Aiyarak, "A more reliable step counter using built-in accelerometer in smartphone," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 2, p. 775, 2018.

[13] M. Johnson, "Activity monitors step count accuracy in community-dwelling older adults," *Gerontology and geriatric medicine*, vol. 1, p. 2333721415601303, 2015.

[14] D. Salvi, C. Velardo, J. Brynes, and L. Tarassenko, "An optimised algorithm for accurate steps counting from smart-phone accelerometry," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 4423–4427.

[15] A. Rai, K. K. Chintalapudi, V. N. Padmanabhan, and R. Sen, "Zee: Zero-effort crowdsourcing for indoor localization," in *Proceedings of the 18th annual international conference on Mobile computing and networking*, 2012, pp. 293–304.

[16] S. Leonhardt, T. Falck, and P. Mähönen, *4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN 2007): March 26-28, 2007 RWTH Aachen University, Germany*. Springer Science & Business Media, 2007, vol. 13.

[17] K.-C. Lan and W.-Y. Shih, "Using smart-phones and floor plans for indoor location tracking-withdrawn," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 2, pp. 211–221, 2014.

[18] R. G. Mello, L. F. Oliveira, and J. Nadal, "Digital butterworth filter for subtracting noise from low magnitude surface electromyogram," *Computer methods and programs in biomedicine*, vol. 87, no. 1, pp. 28–35, 2007.

[19] B. Lin, "Machine learning and pedometers: An integration-based convolutional neural network for step counting and detection," Ph.D. dissertation, Clemson University, 2020.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:51:50 UTC from IEEE Xplore. Restrictions apply.

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/he

Review Article

# Overview of hydrogen production from biogas reforming: Technological advancement

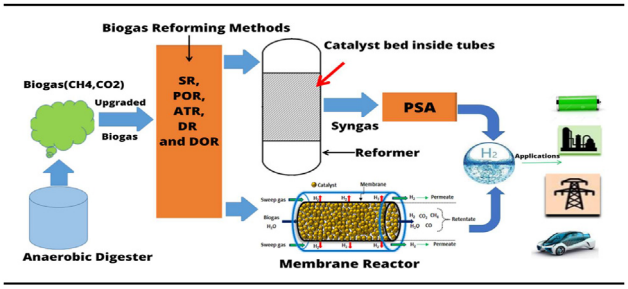
*Ravindra Kumar* [a], *Anil Kumar* [a,b,*], *Amit Pal* [a]

[a] *Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India*
[b] *Centre for Energy and Environment, Delhi Technological University, Delhi 110042, India*

## HIGHLIGHTS

- Biogas reforming is done in order to production of $H_2$.
- Methane requires high temperature for reaction in steam reforming process with catalyst.
- $H_2$/CO ratio is 3 and 2 means $H_2$ yield above 70% and almost 67%, respectively.
- $H_2$ yield around 74%, with $H_2$/CO ratio close to 2.8 in auto thermal reforming.
- Dry reforming process leads to molar ratio $H_2$/CO nearly 1 and $H_2$ yield approx 50%.

## GRAPHICAL ABSTRACT




## ARTICLE INFO

*Article history:*
Received 30 April 2022
Received in revised form
31 July 2022
Accepted 5 August 2022
Available online xxx

*Keywords:*
Biogas
Hydrogen
Reforming
Catalyst
Membrane reactor

## ABSTRACT

In this article, possibilities of biogas reforming techniques for hydrogen production are discussed. The consideration of biogas reforming to produce $H_2$ and fuel cell application from membrane technology is presented. In steam reforming process, methane requires a high temperature for reaction, but a suitable catalyst can manage a higher temperature. The ratio of $H_2$/CO is close to 3, which means higher $H_2$ yield (above 70%). The ratio of $H_2$/CO to nearly 2 and $H_2$ yield almost 67% and also reduces the soot formation for partial oxidation process. In Auto thermal reforming, higher yield of $H_2$ is around 74% with the ratio of $H_2$/CO close to 2.8. The dry reforming process leads to a molar ratio $H_2$/CO of nearly one and $H_2$ yield of approximately 50%. The ratio of $H_2$/CO correspondingly improves and generates $H_2$ yield of approximately 60% for dry oxidation reforming process. For sustainable decentralized power generation in remote and rural areas, large-scale development of $H_2$ energy technology is required. Biogas reforming is an auspicious process for the production of green hydrogen gas as well as for reducing overburden on natural gas. The main benefit of using biogas for $H_2$ production as a renewable energy source is reducing

* *Corresponding author.*
E-mail address: anilkumar76@dtu.ac.in (A. Kumar).
https://doi.org/10.1016/j.ijhydene.2022.08.059
0360-3199/© 2022 Hydrogen Energy Publications LLC. Published by Elsevier Ltd. All rights reserved.

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

excessive burden on natural gas and greenhouse gas emissions. Nowadays, the importance of renewable $H_2$ production has increased due to many reasons such as depletion of fossil fuel reserves, global environmental issues, energy issues, and demand for pure $H_2$.

© 2022 Hydrogen Energy Publications LLC. Published by Elsevier Ltd. All rights reserved.

## Contents

Introduction ... 00
Biogas potential ... 00
Biogas reforming techniques for $H_2$ production ... 00
Steam reforming process (SR) ... 00
Partial oxidation reforming (POR) ... 00
Auto thermal reforming (ATR) ... 00
Dry reforming process (DR) ... 00
Dry oxidation reforming (DOR) ... 00
Current research on conventional processes from biogas reforming based on catalyst ... 00
Steam reforming of biogas (SRB) ... 00
Partial oxidation of biogas (POB) ... 00
Auto thermal reforming of biogas (ATRB) ... 00
Dry reforming of biogas (DRB) and dry oxidation reforming of biogas (DORB) ... 00
Summary ... 00
Membrane reactor (MR) ... 00
Concept of MR ... 00
Arrangement of membrane reactors ... 00
Packed bed membrane reactor (PBMR) ... 00
Fluidized bed membrane reactor (FBMR) ... 00
Modeling outcomes from PBMR and FBMR ... 00
Recommendations for the production of $H_2$ from biogas ... 00
Hydrogen utility in fuel cells ... 00
Recent status of biogas reforming for $H_2$ production from conventional reactors and membrane reactors ... 00
Future prospect of biogas reforming in research direction ... 00
Conclusion ... 00
Declaration of competing interest ... 00
Acknowledgment ... 00
References ... 00

## Introduction

The current energy sector focuses on producing energy via traditional resources like fossil fuels. Growing worldwide demand for energy and environmental alarms related to greenhouse gas emissions has fueled the introduction of alternative power generation possibilities. Biogas reforming is an auspicious process for the production of green hydrogen gas as well as for reducing overburden on natural gas. Currently, Hydrogen production has become the most important research area, especially renewable energies due to globally growing energy demand, depletion of fossil fuels, environmental issues, and progressive fuel cell technology [1,2]. As a renewable energy source created through biological material derived from living organisms and plant-based material used as a fuel termed biomass. Biomass energy systems make an important contribution to sustainable progress and energy systems in developing and developed countries in the coming decades [1].

Biomass could be the highest appropriate renewable energy option compared to other renewable energy sources because of its several benefits and abundantly available worldwide. The waste biomass can be used for generating $CH_4$ from decomposition of organic materials; otherwise, it may contribute to greenhouse effect. Gaseous fuels such as biogas and producer gas could be generated through biomass and agricultural residues like rice straw or wheat straw can be transformed into biogas from anaerobic digestion. In contrast, biogas contains primarily mixture of $CH_4$ and $CO_2$, including other gas traces like $H_2S$, $H_2$, $N_2$, $NH_3$, $H_2O$ and $O_2$ [2].

Biogas can also be attained through landfills known as landfill gas. Biogas composition can vary from site to site, feedstock, and anaerobic digesters used. The most important technology for large-scale production of $H_2$ is to facilitate the

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

reforming of light hydrocarbons (HCs), predominantly methane ($CH_4$) which is the main component of biogas [3]. However, some of these routes produce a combination of CO and $H_2$, also called synthesis gas, widely used in industry [4]. The production of purity $H_2$ via synthesis gas is quite expensive due to the need to remove CO [5].

According to literature on reforming, the research has highlighted cost and energy-saving catalysts. The use of catalysts with higher catalytic activity and stability in reforming route can lead to a decrease in the high temperatures commonly used, which can be accompanied by improvement in the reaction rate and a slowdown in the catalyst deactivation process. Poisoning effect, additional impurities, and carbon deposition (coke) are often severe problems [2].

Generally, hydrogen is generated from various hydrocarbons like methane, methanol or ethanol, and naphtha oil through the traditional steam reforming method. However, according to industrial scale, mostly $H_2$ (above 80%) is produced from SR of natural gas ($CH_4$) using fixed bed reactors (multitubular). In order to small scale, other two alternate options considered with SR such as partial oxidation and auto-thermal reforming, whereas POx occurs less efficiently compared to SR and ATR occurs both exothermic reaction (via POx) and endothermic reaction (via SR) conduct in same reactor [6].

The limitation of equilibrium reactions is the main drawbacks of conventional reforming processes such as SR, POx and ATR. In order to complete conversion of fuel, generates hydrogen with rich gas mixture involving $CO_2$ and other by-products. Thus, to produce pure hydrogen, this chemical process takes place in several reaction units such as high temperature reformers, high temperature (HT) and low temperature (LT) shift reactors, followed by separation units mainly such as pressure swing adsorption (PSA). The number of different process steps reduces system efficiency and makes scale-down uneconomical. A typical illustration of the reaction process of SR is shown in Fig. 1.

Although this method provides high hydrogen yields, obtaining the desired high purity hydrogen requires complex energy integration, expensive heat exchangers (High temperature), and various process units. Among the various technologies associated with hydrogen generation, purification and separation, membrane technology appear to be more auspicious, and membrane separation is progressively gaining attention as a good candidate for replacing existing systems. The specific constraints of thermodynamics and limitations of conventional reactors can be avoided with innovative integrated systems, also known as membrane reactors (MRs). Membrane reactors are technical systems where reaction and separation take place in single unit [7], and compared to conventional configurations wherein the reactor is coupled to a unit of downstream separation, the utility of membrane reactor can offer many potential benefits, such as lower capital costs (caused by reduced size unit), increased yield and selectivity (caused by equilibration shift effect) and decreased downstream cost of separation (integrated separation unit).

The achievement of membrane reactors in order to $H_2$ production depends on few factors such as advancement in thin membrane production with high $H_2$ perm-selectivity and high $H_2$ fluxes, an innovative reactor concept that integrates energy exchange and separation unit, reduces resistance to the transfer of heat and mass, and simplifies shells and seals of membranes.

The main goal of this manuscript is to highlight the numerous possibilities of biogas reforming for hydrogen production based on various conventional and membrane reactors. Catalysts play a key role in hydrogen production methods. Different catalyst compositions in various biogas reforming methods have also been extensively studied. In addition, membrane reactors, their types, and simulation models for the production and purification of hydrogen were investigated. This review presents biogas as renewable energy in hydrogen production and fuel cell applications.

## Biogas potential

Biogas is viable, sustainable, and efficient energy resource because of plentiful supply of low cost feedstocks. The countless applications such as in power sector, heating, fuel, and hydrogen production by chemical processing and biofuels through raw materials. Worldwide, biogas-based power generation capacity has been increasing from 65 GW (year 2010) to 120 GW (year 2019) [145].

The global capacity of biogas plants was approximate 19.5 GW at the end of the year 2019 with growth driven by factor as high prices of fossil fuels, easy access to low cost biomass feedstock and more concerns over global warming and emissions [146,147]. The development of biogas accounted 0.25% energy market and 27% biofuel market worldwide in 2011 [148]. In 2005, biogas digesters (small households) were available to nearly 16 million worldwide, but most were in India and China. In India, the firewood was switched by biogas as an equivalent of 16 million tons (MT) up to 1996, while in China, about 4% of total energy demand is fulfilled by 7 million biogas digesters [148].

Production of biogas and its utilization have numerous benefits. It offers greenhouse reduction, alternative fuel, heat and electricity, fertilizer as by-product, waste recycling, and safety of environment from pollutants. Biogas arrangements could be renovated manure or organic domestic waste materials into gas to light and cook. The waste products like rice, fruits, and vegetables can be transformed into energy rather than disposed of. The waste materials storage and disposal invite insects and pests. Biogas supports waste management and donates to enriched hygiene in rural zones [8].

Biogas can also help dipping deforestation by using less firewood and promoting fertilizers that carry a carbon footprint and affect the soil quality. Biogas consumption drops GHG emissions compared to known fuels (fossil fuels). The combustion of Bio-CNG produces nearly 8–22 g $CO_{2\ eq.}$/MJ, which is 80% lower with compare to fossil fuels. Biogas combustion generally does not generate polyaromatic hydrocarbons (HCs) and hazardous aromatics. Lacking these HCs avoids tenet soot emissions and particulate matters via combustion process. Formation of soot from combustion of biogas/natural gas depends upon combustion conditions, whereas net soot emission is reduced from gas combustion control that is tough to attain along with liquid fuel. Filling stations of biogas (liquefied) are well known in the US, China and Europe [9].

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

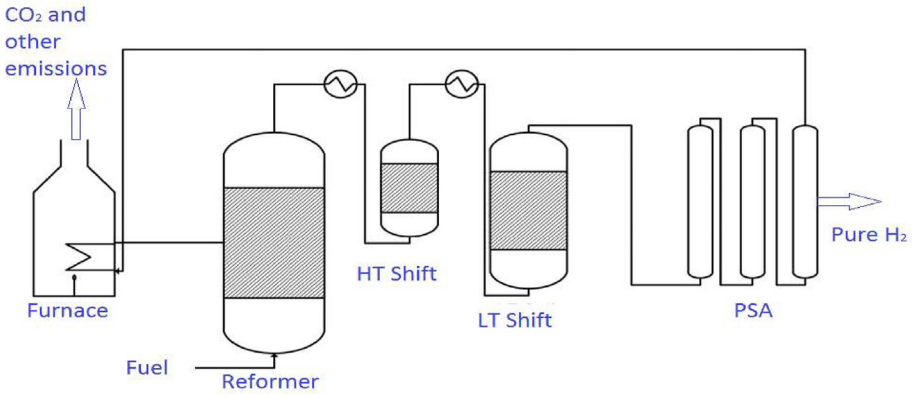



**Fig. 1 – Steam reforming process via conventional method using various reactors [6].**

Biogas is primarily used as fuel for cooking purposes and stationary engines. Biogas potential has not yet been fully explored and used. To promote domestic biogas, the Indian government established the National Biogas Development Project (NPBD) in 1981. The number of biogas plants in India has increased from 1.23 million in 1990 to 4.54 million in 2012, despite the enormous potential of 12.34 million digesters [10].

The installation of biogas plants has increased related to its potential by approximately 26%–37% from 2002 to 2012. As domestic scale (biogas plants) leads to an expected saving of 4.0 million tons of fuel (wood type) every year, as well as producing one Mt of enhanced organic manure [11]. In Europe, the production of biogas has been growing nearly from $9.29 \times 10^9$ L (2009) to $1.65 \times 10^{10}$ L (2016) through the using biogas technology [147].

Biogas is currently used in integrated heat and power (CHP) plants for heat and power generation. In Sweden and Germany, biogas has been used as a transport fuel. It can be turned not only into fuel but also into a valuable commodity as synthesis gas or syngas ($H_2$+CO).

Syngas is used to produce alcohol, alternative fuels, synthetic ammonia-based fertilizers and petroleum-based refinery feedstock [8].

## Biogas reforming techniques for $H_2$ production

The reforming method can be carried out both in tubular type fixed bed reactors and in fluidized bed reactors and continuous type fixed-bed reactors have become widespread due to their relatively low cost and simple design. Generally, all these methods can be performed at low pressures, but in practice, intermediate pressures are used from 10 to 20 bars to increase the efficiency [12].

### *Steam reforming process (SR)*

On an industrial scale, $H_2$ has been produced mainly by steam methane reforming (SMR) over the past 100 years and mainly used in refineries and fertilizer production. The general equation for SMR from biogas or methane is mentioned below [13,143,144].

$$CH_4 + H_2O_{(Steam)} \rightarrow CO + 3H_2 \quad \Delta H_{298K} = +206.2(kJ/mol) \qquad (1)$$

In this process, methane molecule reacts with steam molecule in the presence of suitable catalyst to generate syngas and reaction takes place at high temperature. This process includes steam reforming (SR), water gas shift (WGS) and combined steam reforming reactions [7]. First, methane is reacted with water vapor in SR to form syngas containing carbon monoxide (CO) and hydrogen ($H_2$) in the presence of an appropriate catalyst (Equation (1)). Syngas gas (mixture of CO + $H_2$) is then cooled to a temperature between 300 °C −500 °C and fed into the WGS reactor, where carbon monoxide (CO) reacts with water ($H_2O$) to reduce the CO content and produce carbon dioxide($CO_2$) and hydrogen($H_2$) (Equation (2)) [7]. Additionally, combined or direct steam reforming reaction is sum of SR and WGS reaction (Equation (3)).

Water gas shift reaction:

$$CO + H_2O_{(Steam)} \rightarrow CO_2 + H_2 \quad \Delta H_{298K} = -41.2(kJ/mol) \qquad (2)$$

Combined reaction:

$$CH_4 + 2H_2O_{(Steam)} \rightarrow CO_2 + 4H_2 \quad \Delta H_{298K} = +165(kJ/mol) \qquad (3)$$

The ratio of $H_2$/CO is close to 3, which means higher yield of $H_2$ (above 70%). Methane requires a high temperature (840–950 °C) for reaction, but a higher temperature can be managed by a suitable catalyst [14]. Most industrial steam reforming units use ceramic-based nickel catalysts operating at temperatures of 700–1000 °C and a pressure between 15 and 30 atm [15].

The main detriment of this process is that enormous heat is required to shift equilibrium reaction rightward with the formation of $H_2$ and CO. The reason is that the reaction is endothermic and requires an external heat source [16].

### *Partial oxidation reforming (POR)*

POR is another reforming process with methane reforming to produce $H_2$ with minimum energy costs, and it is done by exothermic process with the utility of air or oxygen.

$$CH_4 + \frac{1}{2}O_2 \rightarrow CO + 2H_2 \quad \Delta H_{298K} = -35.6(kJ/mol) \qquad (4)$$

In this process, methane molecule is partially oxidized to generate a mixture of CO and $H_2$, requires temperature between 700 and 900 °C at atmospheric pressure, ratio of $H_2$/CO

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

to nearly 2 and $H_2$ yield almost 67% and reduces the soot formation [17].

One important characteristic to consider during processing is the molar ratio ($O_2/CH_4$). A higher molar ratio ($O_2/CH_4$) results in complete combustion by fuel, as seen in Equation (5), which is an exothermic reaction type that increases the temperature of reaction, which can initiate the development of hot spots and coke on the catalyst surface in the reactor and decrease the yield of $H_2$ compared with SR process [2].

$$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O \quad \Delta H_{298K} = -801.7(kJ/mol) \tag{5}$$

Nickel-based catalysts are regularly used for POR reactions. Several selective and highly active catalysts for POR methane have been described, such as mixed crystals of NiO–MgO, Ca–Sr–Ti–Ni and Ni–Mg–Cr–La–O. Among the most suitable catalysts for POR reaction, the Ru-based catalyst has excessive coke resistance at very high temperatures. Since ruthenium oxide has good thermal and chemical stability and more resistance to chemical corrosion, it increases the selectivity of $H_2$ production and conversion of methane [19].

### *Auto thermal reforming (ATR)*

To mitigate the disadvantages of the endothermic nature of SR process and low $H_2$ recovery from POR process, auto thermal reforming (Equation (6)) was developed as combined group of the two processes [8].

$$CH_4 + \frac{1}{2}H_2O_{(Steam)} + \frac{1}{4}O_2 \rightarrow CO + \frac{5}{2}H_2 \quad \Delta H_{298K} = 0(kJ/mol) \tag{6}$$

The term 'Autothermal' denotes no need for heat required through the external source for the reaction. It manages the demand of heat of endothermic reaction by releasing heat through the exothermic reactions and can be reflected as an effective process for $H_2$ production [8]. Partial oxidation and methane reforming reactions are injected by steam and oxygen simultaneously. The higher $H_2$ yield is around 74%, with the ratio of $H_2$/CO close to 2.8 [18].

In an ATR reformer, partial oxidation reaction consists in the thermal zone and producing heat to steam reforming occurs in the catalytic zone [12]. The major drawbacks associated with the ATR process are soot formations in combustion zone that feed steam to cause upturns capital and operating cost of reformer [8]. Higher plant efficiencies achieved by biogas reforming via ATR process are 75%, operating temperature range varies from 500 to 700 °C and molar ratio of O/C from 0.80 to 0.90. Generally, maximum temperate is applied for the process since low temperature rises molar ratio $H_2$/CO due to the enlarged content of the non-reacted $CH_4$. Additionally, a low temperature is suitable for further removal of CO [12].

### *Dry reforming process (DR)*

In the DR process, methane reacts with $CO_2$ to generate CO and $H_2$ and reaction temperature increases above 640 °C. The general Equation (7) for dry reforming process is as follows [19].

$$CH_4 + CO_2 \rightarrow 2CO + 2H_2 \quad \Delta H_{298K} = 247(kJ/mol) \tag{7}$$

The reaction of DR is the conversion of methane and carbon dioxide into syngas. In addition to reactions, other chemical reactions can also take place at the same time during the process. A typical side reaction is reverse water gas shift (RWGS) (Eq. (8)) which can reduce the ratio of $H_2$/CO in syngas products. Though it may be undesirable in syngas production, this reaction helps establish the correct $H_2$/CO ratio for higher hydrocarbon production [149].

$$CO_2 + H_2 \leftrightarrow CO + H_2O \quad \Delta H_{298K} = 41(kJ/mol) \tag{8}$$

Particular side reaction contributes formation of coke during the process. The methane ($CH_4$) decomposition presented by Eq. (9), wherein methane decomposed completely toward solid carbon over the catalytic surface and emitted $H_2$. The CO disproportionation is termed as Boudouard reaction, presented in Eq. (10). The hydrogenation of $CO_2$ and CO in Eqs. (11) and (12), respectively, wherein $CO_2$ reacts with $H_2$ to develop solid carbon (C) and water vapor ($H_2O$) [150].

$$CH_4 \leftrightarrow C + 2H_2 \quad \Delta H_{298K} = 74.9(kJ/mol) \tag{9}$$

$$2CO \leftrightarrow C + CO_2 \quad \Delta H_{298K} = -172.4(kJ/mol) \tag{10}$$

$$CO_2 + 2H_2 \leftrightarrow C + 2H_2O \quad \Delta H_{298K} = -90(kJ/mol) \tag{11}$$

$$CO + H_2 \leftrightarrow C + H_2O \quad \Delta H_{298K} = -131.3(kJ/mol) \tag{12}$$

There is a need for an external heat source due to endothermic reaction and also this process is attractive for biogas because of two basic elements involved in raw biogas such as $CH_4$ and $CO_2$. This process leads to molar ratio $H_2$/CO of nearly 1 and yield of $H_2$ of approximately 50% [2]. The main problem with DR is that coke can build up, deactivate the catalyst and blockage the reactor. From previous processes, catalytic materials with $H_2$ selectivity and thermal stability have been used and countless interests to reduce carbon deposition [2]. Although catalysts based on Ru, Rh, and Pt improve the production of $H_2$, they have not received industrial development due to limitations of availability and their high cost, while catalysts based on Ni and Co are widely used [2,12]. Circulating bed reactors and regenerators are a technical process often used to deposit coke on catalysts to avoid catalyst deactivation during continuous operation. The coke formation accumulated on the catalyst surface is removed by incineration, and the generated heat can be used for endothermic dry reforming [12].

### *Dry oxidation reforming (DOR)*

DOR process has been developed by combining two processes as dry and partial oxidation reforming of methane and used for controlling the carbon deposition on catalyst surface. The general equation (Eq. (13)) of DOR is represented by Refs. [2,12]:

$$CH_4 + \frac{1}{2}CO_2 + \frac{1}{4}O_2 \rightarrow \frac{3}{2}CO + 2H_2 \quad \Delta H_{298K} = 123.5(kJ/mol) \tag{13}$$

Combined feed of $O_2$ with $CH_4$ and $CO_2$ offers the following additional benefits: Minimized overall energy consumption, increased conversion of $CH_4$ and enlarged yield of product at low temperatures, increased stability of catalyst and better resistance of deactivation and manage the ratio of $H_2$/CO by adjusting the $O_2$ flow condition [20]. This route additionally

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

diminutions the reaction of energy demand as the exothermic partial oxidation system releases heat and is more useful for dry oxidation process (endothermic). The ratio of $H_2$/CO correspondingly improves and generates $H_2$ yield of approximately 60% [12].

## Current research on conventional processes from biogas reforming based on catalyst

### *Steam reforming of biogas (SRB)*

Hydrogen production process from steam reforming of biogas is considered as a promising alternative route to steam methane reforming with similar efficiency [2] and presented a detailed assessment of hydrogen production method using biogas [21]. Table 1 expresses the catalytic performance of steam reforming biogas (SRB) and their calculations. The noble metals offer a high conversion rate with low temperatures than Ni metal, which offers a high conversion rate with high temperatures.

The main challenge in developing catalysts in hydrogen production via biogas is avoiding deposition of carbon in energetic phase to raise its life. Furthermore, it involves improving the catalyst resistance to poising sulphur content and increasing catalyst of surface area in order to promote activity [8].

Ahmed et al. evaluated catalyst (4 wt% Rh/La−$Al_2O_3$) through SBR in hydrogen production. The catalyst performance was analysed in order to temperature variation from (590−685 °C), the molar ratio of steam to carbon range (1.28−3.85), molar ratio of $CO_2/CH_4$ (0.54−1.50) and the range of GHSV (9810-27000 $h^{-1}$). It was observed that highest conversion of $CH_4$ and lowest conversion of $CO_2$ at 650 °C (in Table 1) at GHSV 9810 $h^{-1}$ and 1.32 S/C low $CO_2$ conversion rate due to low temperature of reaction results in more contribution of WSR reaction. The increasing S/C positively affected $CH_4$ conversion, resulting in higher $H_2$ yield and the marginal effect of GHSV on their process performance [26].

Effendi et al. performed the comparison of fluidized and fixed bed reactors at catalyst (11.5 wt% Ni/$Al_2O_3$), molar ratio of $H_2O/CH_4$ (2.2), GHSV (18,000 $h^{-1}$) and along with temperature 750 °C. The fluidized bed reactor was evaluated to offer (5−15%) better performance than fixed bed reactor. Higher feeding of ratio (gas to steam) was introduced to strong carbon formation due to reduced fluidization in reactor and carbon formation was minimized by increased steam feed. The poor performance of fixed bed produces cold spot formation in the bed of catalyst [29].

Izquierdo et al. estimated the result of various compositions on bimetallic Rh−Ni and Ni catalyst performance in steam biogas reforming at temperature 800 °C, ratio of $CH_4/CO_2$ (1.5) and WHSV (131.6 $g_{gas}/g_{cat}$ $h^{-1}$), whereas catalyst (13 wt% Ni/Ce−Zr−$Al_2O_3$) observed at steam methane ratio of one gives 99.5% of $CH_4$ conversion and 67% of $CO_2$ conversion [30]. In addition, as per Ahmed et al. [26], increasing ratio of S/C was a positive advantage in $CH_4$ conversion, but $CO_2$ conversion was unfavorable, resulting in improvement of WGS over the catalyst (4 wt% Rh/La−$Al_2O_3$).

Despite higher conversion of $CH_4$ and $CO_2$, catalyst presented low $H_2$ yield compared with the catalyst of Ni/Ce−$Al_2O_3$ and Ni/Zr−$Al_2O_3$. The catalysts activity and stability are affected by $H_2S$, which is present in biogas. Sulphur content from $H_2S$ reduced the catalyst activity because it reacts with active metals, limiting access of reactants to active sites. Furthermore, the development of stable metal adsorbate bonds could lead to non-selective side reactions [31]. Additionally, $H_2S$ acts as a toxic gas for metal-based catalysts, but a certain amount of $H_2S$ in the feed increases the reforming activity of the catalyst [32]. Due to the toxic effect, the content of $H_2S$ in biogas sharply reduces the catalytic activity. Therefore, several purification methods have been used to remove $H_2S$, such as membrane technology, adsorption, and biological methods [7].

In contrast, Appari et al. reported catalyst activity loss of 98% with catalyst (15 wt% Ni/$Al_2O_3$) suffered by suphur poising effect [33]. Ashrafi et al. observed that temperate is independent of $H_2S$ concentration and depends on sulphur content over the catalyst's surface in loss of catalyst activity. The poisoning effect of catalyst at the temperature (700 °C) is not recovered by elimination of $H_2S$ content through the feed, but from temperature (800 °C) could be reversed by eliminating $H_2S$ content through the feed. The deactivation of catalyst activity was recovered at (800 °C) and fully deactivated at (700 °C); with increasing temperature $H_2S$ removal increases the catalyst regeneration limit.

**Table 1 – Summary of Catalyst performance calculated by Steam reforming of biogas (SRB).**

| Process | Catalyst | $X_{CH_4}$ | $X_{CO_2}$ | Yield | | Experiment condition | | | | Ref. |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $Y_{H_2}$ | $Y_{CO}$ | S/C | Temp(°C) | $CH_4/CO_2$ | $O_2/C$ | |
| SR | 10 (wt.%) Ni on CeZrLa | 43.0 | 12 | 47.20 | 2.1 | 3.0 | 550 | 1 | – | [22] |
| SR | 5 wt% NiO/$CeO_2$ | 52 | 48 | 72 | 12 | 1 | 873 | 1.5 | – | [23] |
| SR | 15% (Ni) on $Al_2O_3$ | 82 | 8.0 | – | – | 2.0 | 1072 | 1.2 | – | [24] |
| SR | 0.09 (wt%) [Pd(7)−Rh(1)]/($CeZrO_2Al_2O_3$] | 90 | 6.4 | 71.2 | 0.8 | 1.50 | 1072 | 1.5 | – | [25] |
| SR | 4 wt%$RhO_2$/3 wt% $La_2O_3$−$Al_2O_3$ | 93 | 10.5 | 2.62[a] | 0.98[a] | 1.3[a] | 685[a] | 1[a] | – | [26] |
| SR | 1.3 wt %Pd−Rh/$CeZrO_2$−$Al_2O_3$ | 93 | 5 | – | – | 1.2[b] | 800[b] | 1.5[b] | | [25] |
| SR | $Ni_{0.4}Ce_{0.8}O_2$ | 96 | 40 | 72 | 5 | 2 | 700 | 3 | 6.2 | [27] |
| SR | 7.5 (wt%) Ni/$CeO_2$ | 98 | 4.0 | 55 | 29 | 1.0 | 901 | 1.5 | – | [28] |

[a] Experiment was performed using WHSV of 9810 $h^{-1}$. The yield was evaluated based on mol of $H_2$ produced/mol of $CH_4$. The experiments were completed for 40 h.
[b] The experiment was executed for 200 h using WHSV of 20,000 $h^{-1}$.

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

Efficient removal of sulphur content could be costly in biogas applications but it can be economical to operate the reformer with a higher feed of $H_2S$ biogas at higher temperatures. Reformers process at high temperature conduct extensive issues such as hot spot formation, sintering metal crystallites, and decreased surface area of catalyst etc. These issues generate loss in activity, stability, and catalyst life [34]. Catalyst such as Ni based is more active in SR of biogas process from Table 1(numerous catalysts in SRB). In maximum applications, conversion of $CO_2$ was low compared to conversion of methane, while some research also described high conversion of $CO_2$. The high conversion of methane and $CO_2$ has been expressed in higher hydrogen yield.

### *Partial oxidation of biogas (POB)*

Partial oxidation of biogas has several benefits over SRB, including compactness, less sensitivity, and better response time for fuel variants. The reactors of partial oxidation increase interest in energy conversion for high temperatures utilized in fuel cell. It is a faster chemical reaction than SRB due to its exothermic reaction [35].

Hydrocarbon (HC) conversion to syngas via partial oxidation process can occur at high temperatures lacking catalysts. Conversely, the operating temperature will be significantly decreased using catalyst. Most of the studies via partial oxidation catalysts are same as catalyst used in steam reforming process as noble metal and transition metallic elements such as Ni, Fe and Co [36].

Pantaleo et al. described supported catalyst ($La_2O_3$, $CeO_2$ and combined $CeO_2$–$La_2O_3$) with Ni catalyst from Partial oxidation of methane. The support of Ni catalyst is prepared from wet impregnation and co-precipitation. Carbon is formed on $Ni/La_2O_3$ and $Ni/CeO_2$ and carbon is not formed on combined support ($CeO_2$–$La_2O_3$) during test [37].

Chen et al. analysed spiral type Swiss roll reactor via Catalytic partial oxidation of methane (CPOM) and Rh-based catalyst. Three types of biogas are considered for studies: sewage, landfill, and farm biogas. Syngas yield was found at 2.80 mol/mol of $CH_4$) and 31.12% $CO_2$ conversion. The conversion of $CH_4$ (99%) (Table 2), yield of $H_2$ more than 1.49 mol/mol $CH_4$ and ratio of $H_2$/CO were higher than in other studies [38].

Tsyganok et al. evaluated catalyst Pt-, Rh- and Ru-based via joint operation of both dry and partial oxidation of methane ($CH_4$) and supported on Mg–Al mixed oxides. The experiment was conducted at 850 °C and 1 atm, using the ratio of $O_2/CH_4$ (0.46) and GHSV (79 NL $CH_4/(g_{cat}\cdot h)$). The authors investigated Rh-and Ru-based catalysts with high activity performed during the experiment and no carbon deposit during 5 h [39]. While Moral et al. studied syngas production from biogas catalytic partial oxidation and dry reforming used by several Rh-based catalysts with different support $SiO_2$, $CeO_2$ and $\gamma$-$Al_2O_3$. It was measured that the molar ratio $O_2/CH_4$ (0.45) was favorable for higher hydrogen yields and maximum $CH_4$ conversion. The results were obtained by feeding various amounts of $O_2$ and the mixture of biogas in catalytic performance using dissimilar Rh-based catalyst and conducting the experiment with GHSV value of 150 NL $CH_4/(g_{cat}\cdot h)$ and temperature at 700 °C [40].

Rafiq et al. studied through the experimentally and thermodynamic analysis in a plasma assisted gliding arc reactor using Ni catalyst via partial oxidation model biogas as $CH_4$ (60%) and $CO_2$ (40%) and reported highest $CH_4$ conversion as 90%, while the ratio of $CH_4/O_2$ as 3 at temperature 800 °C. The ratio of $CH_4/O_2$ as 2, methane conversion was obtained as 97.5% using Ni catalytic filter in tabular reactor without considering hydrogen. These studies also show that calcium silicate could be a good base material for catalytic filters from partial oxidation of biogas mixture [41].

Table 2 represents the various kinds of catalytic performance from partial oxidation of biogas model. Rh-based catalyst is more useful with various conditions and different parameters.

### *Auto thermal reforming of biogas (ATRB)*

Number of research is limited to auto-thermal reforming of biogas (ATRB). Araki et al. performed ATRB model biogas with catalyst (30 wt% $Ni/Al_2O_3$) and reported the effect of the ratio of S/C and $O_2$/C at the same temperature. Optimum results of S/C (1.5–2.5) and $O_2$/C (0.44–0.56) at 750 °C and investigated 90% conversion of $CH_4$ [46].

Vita et al. investigated catalyst ($Ni/CeO_2$) in auto-thermal reforming of biogas (ATRB). The conversion of $CH_4$ (97.3%) and $CO_2$ (90.5%) with ratio of S/C (0.3) and ratio of $O_2$/C (0.1) was

**Table 2 – Summary of Catalyst performances via partial oxidation of biogas (POB).**

| Catalyst | Performance parameters | $CH_4$ conversion (%) | Temperature (°C) | GHSV | Feed gas | References |
|---|---|---|---|---|---|---|
| Rh | CO selectivity is 52% $H_2$ selectivity is 80% | 90 | 630 | 318,500/h | $CH_4$: $CO_2$: $N_2$ = 50:45:5.0 $O_2$:$CH_4$ = 0.625 | [42] |
| Rh | $H_2$ yield is 1.5 | 79 | 700 | 150 NL/($g_{cat}$ h) | $CH_4$:$CO_2$:$N_2$ = 54:40:6.0 $O_2/CH_4$ = 0.20 and 0.45 | [40] |
| Ni/MgO | $H_2$/CO is 2.6–1.6 | 81 | 700 | 18 L/($g_{cat}$ h) | $O_2/CH_4$ = 1:2 | [43] |
| $NiO/Y_2O_3/ZrO_2$ | $H_2$/CO is 1.01 | 76 | 750 | 0.1073[a] L/min | $O_2$:$CH_4$:$CO_2$ = 0.25:1.5:1.0 | [44] |
| $NaCaCoO_4$ | – | 89 | 785 | 17 000mL/($g_{cat}\cdot h$) | $O_2$:$CH_4$ = 1:1.8–1.91 | [45] |
| Rh | $H_2$ selectivity is 80.93% CO selectivity is 74.23% $H_2$ yield is 1.63 CO yield is 1.24 $H_2$/CO ratio is 1.56 | 99 | 1923 | 10,000/h | $O_2$:$CH_4$:$CO_2$:$N_2$ = 0.6 –0.7:1:0.6–0.91:2.25 –2.70 | [38] |

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

reported (Table 3). The performance of catalyst was tested 150 h with minor sign of deactivation. The catalyst activity was retained due to NiO distribution over the catalyst surface [47].

Izquierdo et al. reported that the catalyst (13 wt% Ni/Ce–Zr–$Al_2O_3$) also utilized in SRB method and ATRB method observed with ratio of S/C (1), $O_2$/C ratio (0.25), and WHSV (161 per hours), see in (Table 3). Bimetallic catalyst (Rh–Ni/Ce–$Al_2O_3$) exposed the maximum yield of $H_2$ in ATR process among the catalysts inspected [48].

Bawornruttanaboonya et al. investigated catalyst (Re–Ni/g-$Al_2O_3$) in auto thermal reforming of biogas. Micro-reactor (Channel type) was compared to novel micro-reactor and noted that novel reactor was optimum condition was $O_2$: $H_2O$: $CO_2$:$CH_4$ (25:5:28:42) % (v/v) at 730 °C and pressure 1 bar [49].

Camacho et al. examined homogeneous (lattice structure) and convention (random form) for $H_2$ production via ATR reaction. Structured catalyst (15–0.05 wt%-Ni-Rh/$MgAl_2O_4$–SiSiC) and monolith ($LiFeO_2$–SiC) were chosen for biogas conversion to $H_2$ gas. The investigations were executed using a ratio of S/C (2), ratio of O/C (1.1), and GHSV (10,000 per hour) through the experimental section and achieved more than 98% methane conversion [50].

Table 3 represents the detail of numerous catalysts in ATRB. It is apparent that catalysts such as Ni based are more active in also ATR of biogas process. The high conversion of methane and $CO_2$ has also been expressed with higher hydrogen yield in the ATR method.

### *Dry reforming of biogas (DRB) and dry oxidation reforming of biogas (DORB)*

Several studies containing dry and oxidative reforming of biogas were investigated apart from reforming SRB and ATRB methods.

Table 4 briefly details different catalyst performances via DRB and DORB and offers extensive results from various catalyst effects under different conditions. This method's main benefit is using greenhouse gases such as $CH_4$ and $CO_2$ [53]. This process is used $CO_2$ for reaction and provides heat source from external due to its endothermic reaction in nature. The key problem of DR process is the development of coke over the catalyst surface during the process due to deficient steam and methane reaction, and this coke formation can blockage the rector. The research on Hydrogen production from DR method has been directed toward thermal stability and growth of carbon tolerant and well-suited noble metallic elements such as Pt, Ru and Rh for DR process, but these metals are more costly. Hence it becomes expensive and unattractive for industries. On the contrary, Ni-based catalyst has been invested for availability and cost factors [8]. The need for high temperature for DR reaction and high carbon content as in reactant makes those metals disposed to deactivate through carbon deposit and sintering. Various researchers investigated bimetallic catalysts in DRB process as Ni-Co [55], Ni–K [56] and Ni–B [57].

Xu et al. communicated the comparison of high-performance catalyst (Ni–Co/$La_2O_3$–$Al_2O_3$) and (Ni/Ce–Zr–$Al_2O_3$) in DRB process and arranged loading of catalyst in wt% (6, 3 and 7) for La, Co and Ni respectively [58]. The higher $CH_4$ conversion and $CO_2$ conversion (94% and 97%) and $H_2$ selectivity 92.7% (Table 4) were detected at a temperature 800 °C along with ratio of $CH_4$/$CO_2$ approx.1 through the period of 500 h, with low 1.9 (wt%) carbon deposition over the catalyst surface.

The authors described the advantages of catalyst Ni/$La_2O_3$–$Al_2O_3$ with adding Co in DRB process [58]. Arbag et al. also express that Co content plays a significant role in the catalyst operation. Large amount of Co (5 wt%) forms $CoAl_2O_4$, which is challenging to reduce and interferes with catalytic activity. For comparison, bimetal (2.5 wt% Ni) -(2.5 wt% Co) from $Al_2O_3$ performed better. Despite the better characteristics, the bimetallic catalyst with a concentration of 2.5 wt% has a higher load. $CH_4$ conversion is increased due to 5 wt% Ni–Co/$Al_2O_3$, Ni–Co metal loading. This indicates a limited contribution of the DR response to $CH_4$ conversion [59].

Several researchers have recently tried increasing the conversion rate in response to biogas reforming. The low steam carbon ratio was maintained to contribute to $CO_2$ in SRB by Appari et al. [33]. Lai et al. performed biogas ($CH_4$-50%: $CO_2$-50%) as equimolar in ODRB. Experiment results from ODR showed that adding a significant amount of oxidant could increase the methane conversion rate, but unnecessary oxidation can reduce the $CO_2$ conversion efficiency [53].

Lau et al. simulated biogas model via oxidative reforming was obtained higher efficiency. In addition, the catalytic activity level of (0.5% Ce)/Ni/$SiO_2$ measures the equilibrium condition and is comparatively stable throughout 24 h of ODRB [54].

**Table 3 – Summary of Catalyst performance calculated by auto-thermal reforming of biogas (ATRB).**

| Process Name | Catalyst composition | $X_{CH4}$ | $X_{CO2}$ | Yield | | Experiment condition | | | | Ref. |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $Y_{H2}$ | $Y_{CO}$ | S/C | TRemperature(°C) | $CH_4/CO_2$ | $O_2$/C | |
| ATR | 25%(Ni) on MgAlCe | 52 | 48 | 72.5 | 12 | 1 | 873 | 1.5 | – | [51] |
| ATR | 1.5 wt% NiO/$CeO_2$ | 97.3 | 90.5 | 96.2 | 1.3[d] | 0.3[d] | 800[d] | 1.5[d] | 0.1 | [47] |
| ATR | 5 wt% Ni on Mg 0.4$Al_2O_3$ | 99 | 15 | – | – | 3 | 1073 | 1.5 | 0.5 | [52] |
| ATR | 13 (wt%)-NiO/Ce–Zr– $Al_2O_3$ | 99 | 34 | 67[e] | 1.5[f] | 1[f] | 800[f] | 1.5[f] | 0.5[f] | [48] |
| ATR | 1 wt%$RhO_2$-13(wt%)- NiO/Ce– $Al_2O_3$ | 99 | 39 | 75[e] | 1.7[f] | 1[f] | 800[f] | 1.5[f] | 0.5[f] | [48] |

[d] The catalyst performance was evaluated with 100 ppm S in the feed for 9 h. The selectivity of $H_2$ was obtained by dividing molar flow rate of $H_2$ divided by molar flow rate of 2 $CH_4$ reacted.
[e] The $H_2$ selectivity was attained by dividing $H_2$ molar flow rate divided by molar flow rate of 2 $CH_4$ reacted. The performance of catalyst was estimated without presence of S.
[f] The catalyst performance was performed for 300 min at WHSV of 1200 per hour over 8 (wt %) Ni supported catalyst.

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

**Table 4 – Catalyst activity examined via DRB and ODRB.**

| Process | Catalyst | $X_{CH_4}$ | $X_{CO_2}$ | $Y_{H2}$ | $Y_{CO}$ | Temp(°C) | $CH_4/CO_2$ | $O_2/C$ | Ref. |
|---|---|---|---|---|---|---|---|---|---|
| [DRB] | 75(Ni-25) Co/($Al_2O_3$)[a] | 25.0 | – | – | 0.6 | 600 | 1 | – | [61] |
| [DRB] | (13 wt%) NiO/(Ce–Zr–$Al_2O_3$) | 60 | 94 | 63 | 1[b] | 800[d] | 1.5[d] | – | [48] |
| [DRB] | 1 (wt%) $RhO_2$-13 wt % NiO/(Ce– $Al_2O_3$) | 64.0 | 85 | 57 | 0.9[b] | 800[d] | 1.5[d] | – | [48] |
| [ODRB] | 20 (wt %) NiO/$Al_2O_3$) | 70.0 | 85 | – | – | – | 1.5 | 1 | [60] |
| [DRB] | 8 (wt%) NiO/20 wt% $CeO_2$– $Al_2O_3$) | 70.0 | 97 | 76 | 60 | 860[f] | 1.5[f] | – | [62] |
| [DRB] | 20 (wt%) NiO/MgO | 80.0 | 85 | | 1.14 | 750 | 150 | 0.1 | [63] |
| [DRB] | 7NiO-3(wt%) $CoO_2$/6 wt% $La_2O_3$– $Al_2O_3$) | 93 | 94 | 95 | 98 | 800[j] | 1[j] | – | [64] |
| [DRB] | 9 (wt%) NiO/$Ce_{0.82}Sm_{0.18}$ $O_{1.91}$ | 98.0 | 97 | 0.8 | – | 800 | 1 | – | [65] |

[a] The catalyst arranged15 wt% metal loading.
[b] The CO yield was calculated on the basis of $H_2$/CO ratio.
[d] The catalyst performance was evaluated at 100 ppm S in the feed for 9 h. The $H_2$ selectivity is obtained by dividing the $H_2$ molar flow rate and divided by 2 M flow rates of $CH_4$ reacted.
[f] The performance evaluation of the catalyst was at WHSV of 1200 per hour for 300 min over 8(wt)% Ni supported catalyst.
[j] The performance of the catalyst was estimated using GHSV of 6000 $cm^3$ $gcat^{-1}$ $h^{-1}$ for 300 h.

Asencios et al. tested the experiment on Ni catalyst support via ODRB and represented the various pH conditions with $Al_2O_3$. Commercially available Ni–$Al_2O_3$ showed the highest activity at 750 °C reaction temperature, but it was deactivated after 3 h. The synthesized Ni–$Al_2O_3$catalyst showed good stability during the 6-h test at more pH conditions close to 7, but obtained only 70% methane conversion rate while the $CO_2$ conversion maintained the equilibrium [60] (Table 4). The catalyst activity is based on the low acidity of the carrier, so less coke is deposited on the catalyst. The low acidity of the substrate reduces the $CO_2$ reaction in the main $Al_2O_3$ core, which leads to an acid-base interaction that promotes this reaction.

## Summary

Hydrogen production through biogas claims to be a sustainable system from environment point of view. However, significant efforts are still required for its generation process to be estimated from a comprehensive environment viable system. Biogas $H_2$ production can be attained through various methods such as SRB, POB, ATRB, DRB and DORB. Table 5 compares the merits and demerits of various reforming processes. Table 6 shows the maturity of hydrogen production with various parameters and conventional reactors. The main key to $H_2$ production via biogas is the development of resistant carbon and sulphur tolerant catalysts. $H_2$ production from SRM is broadly studied and commercially activated. However, the main issue with SRB is the occurrence of $CO_2$ affecting the catalyst performance.

The production of hydrogen from biogas is significantly increasing due to the increasing demand for energy and little literature has been examined. Converting biogas to $H_2$ gas requires the development of small reformers. Wide area synthesis gas applications are used in power generation, chemical industry and energy conversion. Syngas could be directly used in internal combustion engines for energy generation. On the other hand, the direct use of biogas for combustion is disadvantageous due to its high content of $CO_2$ and decreased heating value related to synthesis gas.

The main important aspect of the production of $H_2$ from biogas is the advance of catalysts unaffected by carbon and sulphur content. $H_2$ production with SMR is broadly studied and used commercially. However, the main problem with SRB is the existence of $CO_2$, which affects the catalyst performance. Water gas shift reaction occurs during the SR process and negatively affects $CO_2$ conversion [26].

Low conversion rate of $CO_2$ due to the high activity of the water gas shift changes the composition of the reformation, changes the $H_2$/CO ratio and decreases the quality. In addition, $CO_2$ in feeding as biogas can accelerate the carbon formation reaction and reduce the activity and performance of catalyst. Most issues can be solved by removing present $CO_2$ in biogas, but this removal method is more expensive on a small scale. The existence of $CO_2$ will also be beneficial in decreasing the hot spots produced in the reactor caused by higher temperature operation. Additionally, broadly studied literature via DR method has been reported as deactivation of catalyst and low yield of $H_2$ due to carbon formation. DR and SR processes are endothermic, while dry reforming is more intense and needs an external source of energy supply. Another experiment examined the effect of increasing process efficiency by adding $O_2$ in oxidative dry reforming biogas (ODRB) and removing carbon by oxidation along with $O_2$.

Partial oxidation of biogas appears to have several advantages over SRB, including lower sensitivity and better reaction time to fuel conversion. Partial oxidation reactors are increasingly interested in fuel cells' high-temperature energy conversion. Most research on catalysts for partial oxidation is similar to that used in steam reforming processes such as transition metals and noble metallic elements such as Fe, Ni and Co.

Research on hydrogen production using ATRB is sparse, suggesting more research is needed. The effect of $H_2S$ on the catalytic action of ATR catalysts should be evaluated. Likewise, the impact of the ATR method on $CO_2$ conversion rate was not examined in the study and must be assessed. Composition of molar gas and conversion of $CH_4$ from ATRB process is measured, which does not fully take into account the characteristics of the process. The yield of $H_2$ and CO must be assessed through ATRB process and efficiency should be measured contrary to corresponding chemical stability conditions. Main disadvantage of ATR reformer is the associated decrease in the yield of $H_2$ and low $H_2$/CO ratio.

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

**Table 5 – Merits and demerits of $H_2$ Production from biogas reforming via literature.**

| Technology | Merits | Demerits | Remarks |
|---|---|---|---|
| SR | ✓ Carbon formation is low.<br>✓ Purity of $H_2$ is high.<br>✓ Yield of $H_2$ is high.<br>✓ No need $O_2$ for operation. | ✓ Costly catalyst<br>✓ Need higher temperature<br>✓ Need to add $H_2O$ (vapor)<br>✓ Highest $CO_2$ emissions are high.<br>✓ Air emissions are also high.<br>✓ Need huge size reformer. | ✓ Mostly used in industry applications.<br>✓ Requires removing $CO_2$ from biogas. |
| Partial oxidation | ✓ Operating temperature is relatively low.<br>✓ Energy efficiency is high.<br>✓ Lower methane slip.<br>✓ No need feedstock desulfurization. | ✓Hot spot seems during increasing temperature on catalyst surface.<br>✓ Low Co selectivity can oxidize $CH_4$ completely.<br>✓ Low $H_2$ yield compares to SR.<br>✓ $H_2$/CO ratio is low.v<br>✓ Highly exothermic reaction leading to catalyst deactivation by vformation of hot spots. | ✓ very few industrial applications<br>✓ Unite with other methods.<br>✓ Need air or $O_2$.<br>✓ Catalyst deactivates due to exothermic reaction. |
| ATR | ✓ No requires external heat.<br>✓ Low cost.<br>✓ Energy efficiency is high.<br>✓ Main content can be used as $CH_4$ and $CO_2$ from biogas.<br>✓ Reformer size is compact. | ✓ Need various types of catalysts.<br>✓ Hydrogen yield is low compared to SR.<br>✓ Need control of complex processes.<br>✓ Need air or $O_2$.<br>✓ Separate unit is required for using oxygen. | ✓ Combination of SR and POR process.<br>✓ ATR improves temperature control and decreases hot spot formation and increases temperature control.<br>✓Self-sustaining method. |
| DR | ✓ Conversion efficiency is high.<br>✓ Both greenhouse gases can be used from biogas as $CH_4$ and $CO_2$.<br>✓ Environment friendly using biogas. | ✓ High formation of carbon.<br>✓ Operating temperature is high.<br>✓ Expensive catalyst.<br>✓ Reformer size is large.<br>✓ Commercial experience is low. | ✓ Biogas is much more suitable for this process as both content ($CH_4$ and $CO_2$) are available. |
| ODR | ✓ Formation of carbon is low.<br>✓ Conversion efficiency is high.<br>✓ Improved deactivation resistance and catalyst stability.<br>✓ Decreased total energy involved. | ✓ Production of $CO_2$ is more.<br>✓ Expansive catalyst.<br>✓ Limited availability of catalyst.<br>✓ Another metal is needed for catalyst stability.<br>✓ Complicated control process. | ✓ Industrial application is low.<br>✓ Combination of DR and POR process.<br>✓ Exothermic or endothermic reactions depend on $O_2$ feeding. |

**Table 6 – Conventional process for $H_2$ production from biogas or methane reforming.**

| Process | Reactor type | $CH_4$ conversion (%) | $H_2$/CO | Catalyst | Temperature (°C) | Reference |
|---|---|---|---|---|---|---|
| (SR) | Fixed bed (FB) | 70 | 3.4 | $Ni-Ce_{0.8}Zr_{0.2}O_2$ | 600 | [66] |
| | Fixed bed (FB) | 98 | 3.7 | $Ni-Mg_{17.4}Al_{1.6}O_{20.8}$ | 650 | [72] |
| | Fixed bed (FB) | 90 | 2.7 | $Ru/Al_2O_3$ | 715 | [67] |
| | Fixed bed (FB) | 85 | 2 | $Ni/Al_2O_3$ | 750 | [29] |
| | Fixed bed (FB) | 95 | 2.5 | $Ni/CaO-Al_2O_3$ | 750 | [68] |
| | Fluidized-bed | 96 | 2.2 | $Ni/Al_2O_3$ | 750 | [29] |
| | Fluidized-bed | 98 | 2.1 | $Ni/Al_2O_3$ | 850 | [69] |
| (POR) | Fixed bed (FB) | 100 | 2 | $Ni/Al_2O_3$ | 700 | [70] |
| | Fixed bed (FB) | 95 | 2 | $Ni-CoMgCeO_x/ZrO_2-HfO_2$ | 800 | [71] |
| | Fixed bed (FB) | 85 | 2 | $Pt/CeO_2$ | 800 | [73] |
| | Fixed bed (FB) | 87 | 2 | NiO/MgO | 850 | [112] |
| (ATR) | Fixed bed (FB) | 92 | 3.2 | $Ni/MgAl_2O_4$ | 700 | [113] |
| | Fixed bed (FB) | 95 | 3.5 | $Rh/Al_2O_3$ | 700 | [114] |
| | Fixed bed (FB) | 75 | 1.2 | Ni/NiO–MgO | 750 | [115] |
| | Fixed bed (FB) | 92 | 1.4 | Ni/SBA-15 | 800 | [116] |
| | Fixed bed (FB) | 100 | 2 | $Pt/ZrO_2/Al_2O_3$ | 800 | [117] |
| (DR) | Fixed bed (FB) | 85 | 1.2 | $Ni/CeZrO_2eMg\ Al_2O_4$ | 700 | [118] |
| | Fixed bed (FB) | 67 | 0.7 | La/hydrotalcite | 700 | [119] |
| | Fixed bed (FB) | 70 | 0.9 | $Rh-NiLa/\Upsilon-Al_2O_3$ | 750 | [120] |
| | Fixed bed (FB) | 90 | 1.3 | $Ni/CeO_2-Al_2O_3$ | 860 | [62] |
| (DOR) | Fixed bed (FB) | 86 | 1 | $Rh-NiLa/\Upsilon-Al_2O_3$ | 750 | [120] |
| | Fluidized-bed | 77 | 1.9 | $5Ni/5ZrO_2-SiO_2$ | 750 | [121] |
| | Fixed bed (FB) | 95 | 1.7 | $NdCoO_3$ perovskite | 850 | [122] |
| | Fluidized-bed | 100 | 1 | $Pt-Rh/Ce-ZrO_2-Al_2O_3$ | 900 | [54] |

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

$H_2$ production technology via biogas can be efficiently achieved from several processes like steam reforming (SR) method, partial oxidation (POx) method, auto thermal reforming (ATR) method, Dry reforming (DR) and oxidative dry reforming (ODR) method. Table 5 compares the advantages and disadvantages of many reforming technologies.

## Membrane reactor (MR)

Due to various factors, the membrane technology is not competitive with SR of methane for hydrogen production. In addition, membrane stability cannot yet be fully secure, implying constant maintenance cost. In contrast, membrane technology has not yet been developed for high temperature applications, as degradation occurs when operating at temperatures above 500 °C [151]. However, most pieces of literature have been focused on membrane fabrication for low temperature applications and development progress in high temperature applications has been seen in last few years. The improved membrane stability for high temperature attained by advances in preparation methods, preparation multilayer and Pd based alloy with other materials [152].

A reactor as a fuel processor has numerous major parts such as reforming, combustion, permeating, separation, and purification. The reactor consists of a vessel with a typical closed tubular design or shell type, as shown in Fig. 2. Hydrocarbons are fed into the reactor through another inlet connected to the coil, where the feed is vaporized. The reforming catalyst responds via the vaporized feed tributaries to create a mixed gas flow around the separation section containing an end cap hydrogen sampling membrane supported by tension springs. The hydrogen-rich stream is generated from a portion of the mixed gas stream passing through the membrane, and the remaining mixed gas stream forms a by-product stream. The hydrogen-rich stream passes in the area, penetrating through the tubular membrane. The purification portion is at least partially included in the permeate portion containing the additional catalytic reforming bed. The side stream can be used or stored for other purposes, including fuel for combustion, to meet some or all of the heating needs of the reactor [142].

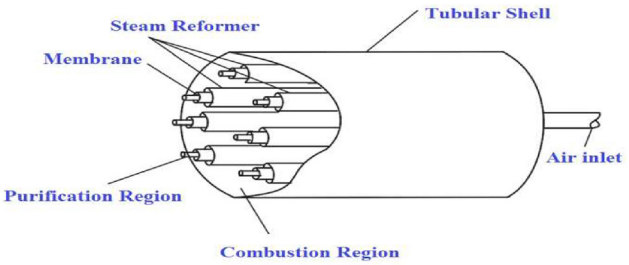



**Fig. 3 – A cross-sectional view of fixed bed (FB) membrane reactor [142].**

Several other implementations use this concept. Fig. 3 schematically shows an alternative reactor design with an enlarged outer metal shell to provide a relatively large combustion portion and one or more membrane tubes [142].

### *Concept of MR*

Conventional steam reformers are widely used industrially for hydrogen production but suffer from several drawbacks such as the formation of carbon, limitation of thermodynamic chemical equilibrium and deactivation catalyst. In addition, the materials used for the reactor walls must be very strong. The reactors operate at high temperatures (650–850 °C) [7].

Membrane reactors have a compact design, in which the chemical reaction and the purification process occur in the same reactor. Membrane reactors also offer low requirement of energy and higher production of hydrogen. Membrane reactors reduce the disadvantages involved with conventional reactors by using selective membrane and the operation of chemical reactions can take place at low temperature (550–600 °C) compared with conventional reactors. In addition, this reactor permits materials to be used at lower temperatures and higher hydrogen yields [7,142]. The assessment of conventional and membrane reactor producing $H_2$ from steam reforming method through the biogas in fuel cell as shown in Fig. 4.

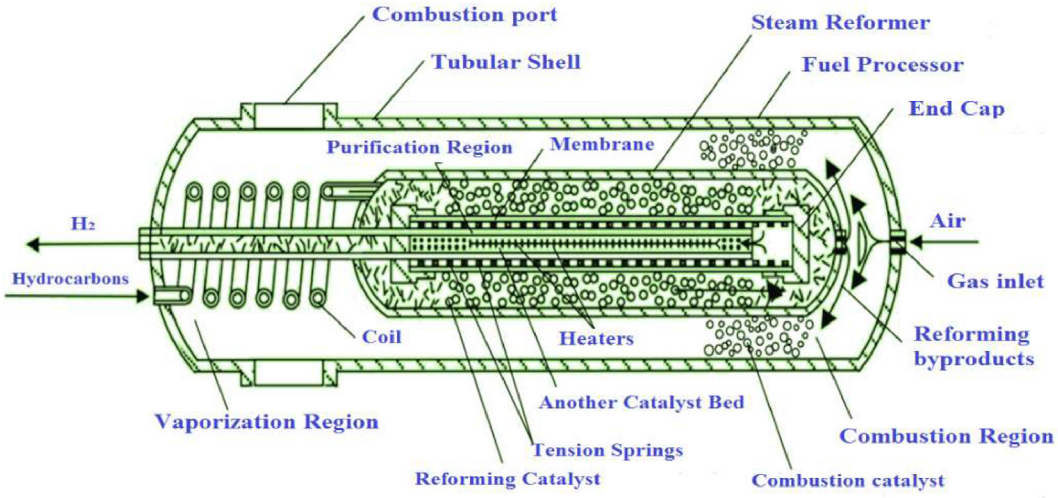



**Fig. 2 – Illustration of fixed bed type membrane reactor [142].**

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

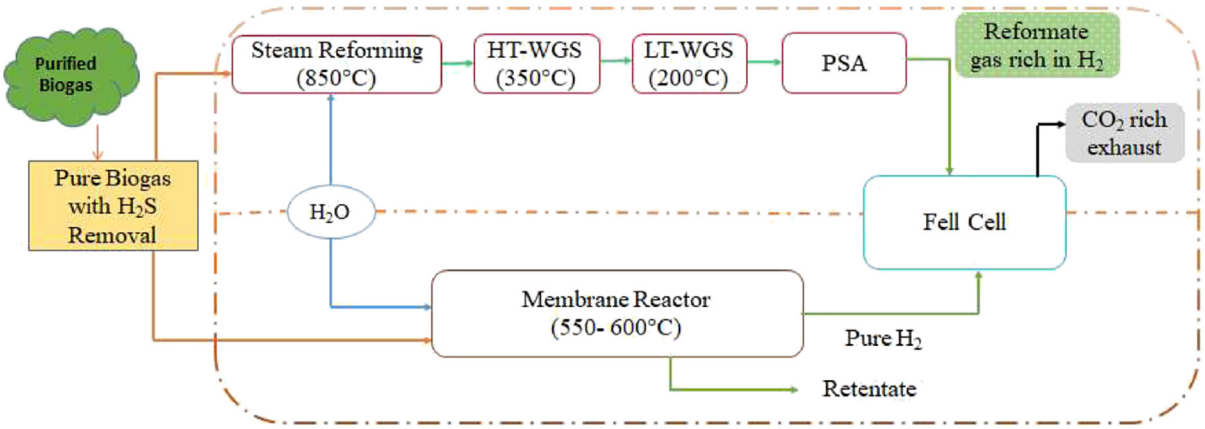



**Fig. 4 – Flow process by conventional and membrane reactors [7,74].**

The membrane is the key role play in membrane reactor for higher hydrogen yield, having more $H_2$ permeation and permitting more hydrogen flow rate. Membrane is a barrier that prevents the passage of non-hydrogen components in the feed gas mixture. The membrane reactors contain a reaction side (region of $H_2$ production) and permeation side (region of $H_2$ separation), as shown in Fig. 5. The membrane must have more selectivity, chemical, low cost and mechanical stability for separating hydrogen from mixture of gases [6].

Membranes can be classified into biological and synthetic membranes, but having several drawbacks involving a pH value range, less operating temperature (<100 °C) and being challenging to clean [75].

Synthetic membranes are mainly divided into organic type (polymer) and inorganic type (matallic and ceramic). The operating temperature is used from 100 to 300 °C in organic membranes, and an operating temperature of more than 200 °C is used in inorganic membranes [7,76]. Fig. 6 shows the mechanism of hydrogen transport through solid Pd-based membranes.

The dense material membranes are classified into supported type and unsupported type membranes. Unsupported type membranes involve drawbacks with brittleness, lower $H_2$ permeability and poising surface. In addition, the membrane area must be improved to ensure high $H_2$ flow rate. Increasing membrane area, as well as increasing membrane thickness, increases material cost [77].

These aspects indicate the need to use composite membranes with porous support and thin selective membranes to improve mechanical steadiness and reduce cost. Pd-based membranes (composite type) selectivity and stability depend on support material quality. Porous material such as stainless steel is regularly used as carrier material due to its properties like good weldability, uniform coefficient of thermal expansion, ease of manufacture, and low cost [78].

Besides the membrane category, various membrane reactors such as micro-reactor, fixed bed and fluidized bed are essential in $H_2$ production. The utility of a fluidized bed reactor has several advantages over a fixed bed as increased heat and mass transfer and eradication of external resistance to mass transfer. The main reasons why this type of reactor is not preferred as rapid decomposition of the catalyst and the complexity of the reactor fabrication [79]. Thus, the fixed bed membrane reactor is the most common.

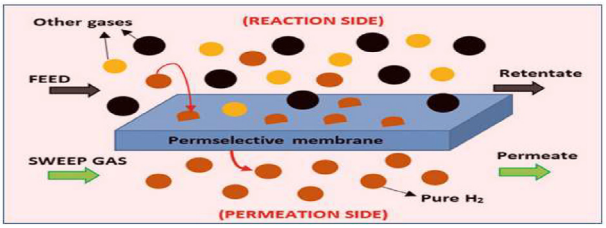



**Fig. 5 – Working principle of the membrane system [7].**

### *Arrangement of membrane reactors*

Technology of membrane reactors is emerging with a variety of $H_2$ purifiers to offer high-purity $H_2$ and recuperate hydrogen via miscellaneous gas tributaries. The literature has presented various categories of membrane reactors in order for hydrogen production. Most of the study has been done in Packed Bed type membrane reactors. However, there is growing interest in new type configurations like fluidized bed membrane reactors, mainly due to better temperature control where mass transfer restrictions can be reduced [6]. Additionally, the membrane geometry can be planner, tubular, spiral wound and plate and frame type. Currently, flat and tubular shapes are most commonly used to separate gas. The planar membranes have been widely used in previous research laboratory, while the tubular membranes are often used on industrial scale. As a result, the ratio of surface area to volume is higher compared to planar membranes [6].

*Packed bed membrane reactor (PBMR)*

The concept of PBMR is the well-known membrane reactor in the literature that has several advantages over the other reactor arrangements, especially in terms of membrane integration and ease of use. The PBMR arrangement generally indicates a system in which the catalyst is introduced into the reactor with a fixed bed configuration with the membrane. In this system, catalyst can be located either in shell or inside membrane, whereas selective gas permeation occurs via reaction side from permeate side (gas separation) or through

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

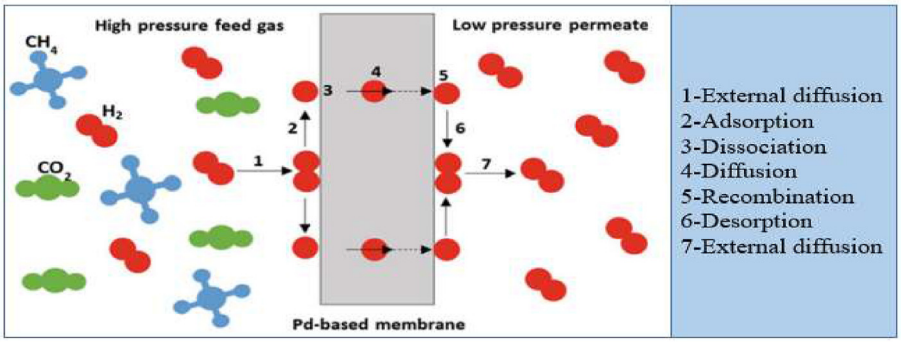



**Fig. 6 – $H_2$ transport mechanism through the dense Pd-based membrane [6].**

membrane side into the catalyst bed. A more thorough description of PBMR design can be seen in Fig. 7.

In this concept, the sweep gas on permeate side makes the most of the driving force in gas separation. The sweep gas via membrane increases the difference of partial pressure, which results in nearly partial pressure being zero with permeate side and is related to increased $H_2$ production. Using a sweep gas also reduces the surface area required for gas separation and is generally attractive if the resulting gas mixture can be used for other processes. Regarding the practice of membrane reactors for applications at high temperatures, reforming steam methane and auto-thermal methane are the two most considered methods. High temperature reactions are preferred over other high temperature reactions and lower temperatures, such as alcohol reforming or water gas shift [140].

Saric et al. investigated SMR with small scale membrane reactor from Pd-based ceramic membrane operating at temperature 580 °C and pressure 28 bars. The membrane thickness was used as 3.8 μm and carried out the steam methane reaction over a nickel catalyst. Methane conversions were nearly (80–95%) and the factor of hydrogen recovery of around (60–80%) during 1100 h of operation, but membrane was presented 12% $CO_2$ concentration on permeate side while content of CO around (55–154 ppm) and also gas leakage at last stage of experiment [80].

Dittmar et al. also studied the reaction of SMR using Pd-based membrane using thickness 12 μm, length 1.9 cm and 6 mm diameter. The performance evaluations of experiment were done in FBR using inner dia. 27.7 mm and 820 mm length with pallets of Ni catalyst. The operation was performed at 600 °C with S/C = 3 and maintained pressure 16 bar on retentate side and 1.3 bar on permeate side, at last, 60% $CH_4$ conversion and $H_2$ recovery around 70% at 650 °C [81].

Marcoberardino et al. performed 1D reactor model for SMR, compared with experimental data using FBMR (fixed bed membrane reactor) and used catalyst Ni–$Al_2O_3$.1% during 5 bar and 600 °C. Modeling approach archived low $CH_4$ conversion as 47.4% and hydrogen recovery factor 28. Dry methane reforming is another method for producing $H_2$ that has been extensively researched in the literature, including membrane technique. In this technique, no required steam ($H_2O$) is fed into the reactor, so no steam is generated at high temperatures, resulting in significant energy savings. However, this process is subject to higher carbon and lower $H_2$ production [82].

Silva et al. used a commercial Pd–Ag alloy using thickness 50 μm to convert methane ($CH_4$) to $CO_2$ in a membrane reactor (packed bed) to produce hydrogen. This study evaluated with conventional packed reactor and used catalyst (Pt/$CeZrO_2$/$Al_2O_3$) for both cases. The main reaction has done the ratio of gas mixture ($CH_4/CO_2$ = 1/1) and rate of feed flow 100 mL/min at 550 °C. The gas separation driving force was formed by feeding Ar (sweep gas) into the membrane, whereas overall pressure (atmospheric) was retained on both sides. The main outcome of this study $H_2$ recovery factor achieved 75% and 60% methane conversion at equilibrium conditions from

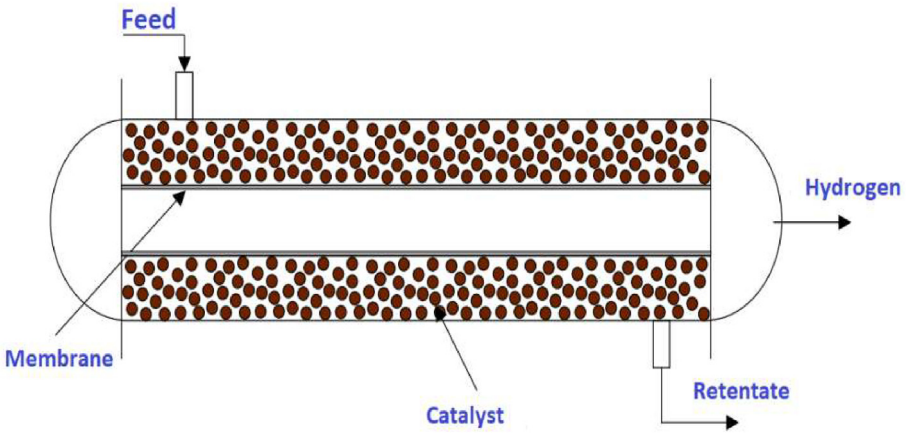



**Fig. 7 – Schematic of fixed bed membrane reactor (FBMR) configuration along with the catalyst confined in shell [140].**

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

conventional reactor and measured countless hydrogen selectivity via membrane deprived of any deterioration for all experiments [83].

Garcia-Garcia et al. similarly performed in a membrane reactor (packed bed) through methane dry reforming method with membrane (Pd-based) and its 5-μm thickness, supported catalyst $Al_2O_3$ with hollow fibre and provided overall surface area 4.4 $cm^2$. The process has been done to compare membrane reactor with Pd-based membrane via stainless steel support, 17 μm thickness and overall surface area 22 $cm^2$ over conventional reactor (Fixed bed reactor). All experiments conducted catalyst (4% $Ru/ZrO_2$–$LaO_3$) and tested activity at various temperature up to 550 °C using a feeding ratio of He: $CH_4$: $CO_2$ = 8:1:1 [84].

The experimental investigation is relatively easy in a packed bed membrane reactor, but it suffers from limitations. The main limitations to this type of configuration reactors such as concentration polarization (limitation of mass transfer bed to wall) [85], huge pressure drop related to catalytic bed (that can be eliminated by using larger particles with a suitably increased internal mass transfer restrictions) and lack of membrane/thermal control freedom. All these disadvantages can limit the further development of membrane reactor technology ( [6].

*Fluidized bed membrane reactor (FBMR)*

FBMR configurations have attracted attention in recent years due to limitations encountered in configurations of membrane reactors (packed bed). This system consists of perm-selective membranes immersed in a fluidized catalyst bed in a bubbling regime. The main advantage of this reactor is decreasing the limitation of bed to wall mass transfer and permitting the operating reactor at a virtually isothermal form (due to catalyst movement). This opportunity can be used to carry out auto thermal reforming of hydrocarbons in membrane reactors. ATR of methane in PBMR is relatively challenging due to hot spot at inlet of reactor. This can melt membrane. Membrane fluidized bed reactors can completely solve this problem. Both auto thermal reforming and $H_2$ recovery can be carried out in the same reactor [86]. The several advantages of membrane supported fluidized bed reactor as listed [87].

- ✓ Pressure drop is negligible, which permits using sizes of tiny particles, resulting in no internal limitation of heat and mass transfer.
- ✓ (Virtual) isothermal process.
- ✓ Membrane flexibility, arrangement of membrane bundles and heat transfer surface area.
- ✓ Advanced fluidization behavior on account of:
  - Decreased axial gas back mixing.
  - Concentrated average bubble size as a result of enriched bubble breakage.

Adris et al. reported the first design of FBMR and proposed Pd-based membranes for selective $H_2$ separation [88]. Over the past period, this design has increased attention, especially after it was shown that Pd-based membranes could sustain perm-selectivities for long periods, even under intense fluidization regimes [89].

Despite the advantages of this arrangement, this concept has been less studied in the literature than fixed bed, and it is suitable for intermediate temperatures. Fuel reforming and alkenes dehydrogenation have been utilized last few years from FBMR with high temperatures. Gallucci et al. described Pd membrane immersed in catalyst bed with air supply and combusted permeated $H_2$ as an exothermic reaction, which offers the necessary heat to reform the process [141]. Medrano et al. also proposed a "Two-Zone" membrane reactor (fluidized bed) reactor. This conception has been proposed and applied to processes in which catalyst deactivation occurs rapidly because of carbon deposition, such as the alkenes dehydrogenation. In this system, the fuel is fed center of bed and the deactivated catalyst is redeveloped over

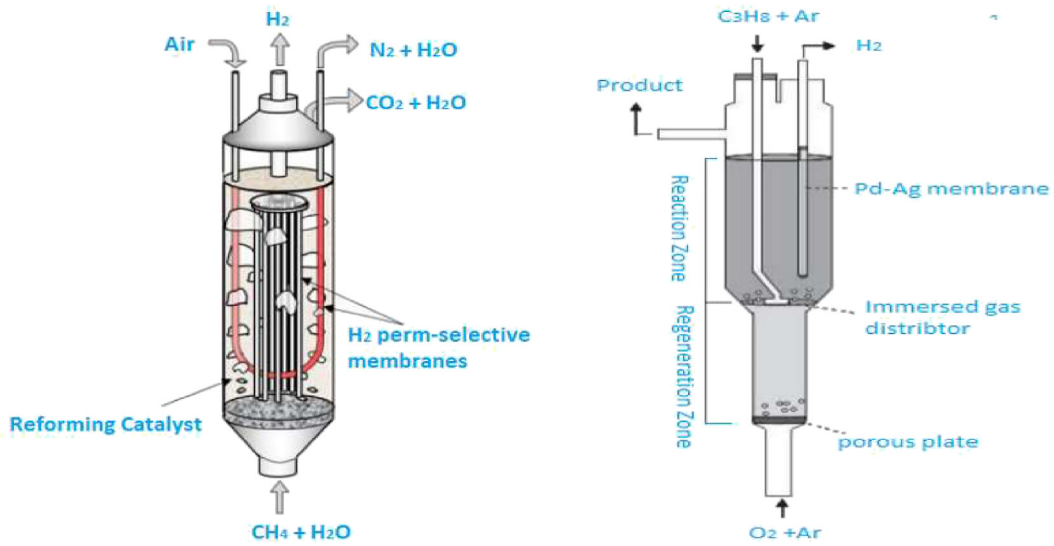



**Fig. 8 – Left side: Fluidized bed membrane reactor configuration with U-shaped membrane to sustain auto-thermal operation [141]; Right side: Two-zone fluidized bed membrane reactor for continuous catalyst reactivation [90].**

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

combustion; carbon deposits in catalyst particles with an insignificant quantity of oxygen supplied through the bottom. This reaction develops a large quantity of heat and that heat is released into the catalyst bed to withstand auto thermal process [90]. A representation of the two new concepts is shown in Fig. 8.

Roses et al. informed the stability of FBMR to steam methane reforming process. The experiment was performed at a temperature of 500–630 °C and the pressure in the reactors was from 2 to 5.3 bar used 3.2 mm dia. of membrane and thickness 3.2 μm of Pd–Ag layer. The test has demonstrated the membranes mechanical strength under the bubbling fluidization situation and the permeated $H_2$ purity remains high throughout entirely parametric studies performed during long period testing by measured CO impurity <10.0 ppm. Tests were run with two different kinds of configurations. Firstly, in a conventional reactor (fluidized bed), then in an FBMR configuration. In all cases, the $H_2$ output was increased [89].

Fernandez et al. also informed the performance of membranes (ceramic supported) for ATR and SMR processes through fluidization conditions at temperatures from 500 to 600 °C. The membranes used in reactor as diameter (1 cm) and length (14.5 cm), respectively. 4 μm thick, Pd–Ag membranes supported on alumina tubes along with an external zirconia layer were arranged with an electroless plating process. The membrane reactor performance was compared via conventional reactor (Fluidized bed) and membrane reactor mounting commercial REB membranes [91]. Outcomes from this study achieved higher hydrogen recovery factors (HRF) with the ceramic supported membranes compared to the commercial membranes (self-supported).

Thus, the fluidized bed membrane reactor represents less studied configuration for high temperature methods, and few studies have been published in recent years. This involves the complex hydrodynamics presence of internals and the difficulty of obtaining reliable models for predicting the fluidized bed membrane reactor behavior. However, the demand for this concept is expected to increase in the next few years as it overcomes many of the drawbacks associated with the design of a packed bed membrane reactor, and it will improve with development of novel ultra-thin membranes. All literature presenting configurations for packed membrane reactor and fluidized bed membrane are given in Table 7.

**Table 7 – Summary of key achievements in configuration of packed and fluidized bed membrane reactors worked at high temperatures.**

| Type of Membrane reactor | Configuration | Process | Catalyst | Thickness (μm) | T (°C) | P (bar) | $X_{FUEL}$ (%) | HRF (%) | Ref. |
|---|---|---|---|---|---|---|---|---|---|
| Packed bed membrane reactor | Pd membrane (Ceramic supported) | SMR | Ni-based | 3.80 | 580 | 28 | 80–95 | 60–80 | [80] |
| | Pd membrane prepared by porous ceramic (NiO-8YSZ) | SMR | Ni-based | 6–10 | 550 | 9 | >60 | – | [92] |
| | Pd membrane (Ceramic supported) | SMR | Pt(Ni)/$CeO_2$ | 5 | 525 | 10 | 90 | | [93] |
| | Pd–Ru membrane (ceramic porous alumina support) | SMR | Ni–Pt/CeZnLa | 4–5 | 550 | 10 | 96 | – | [94] |
| | Pd–Ag alloy provided by REB Research | DR | Pt/$CeZrO_2$/$Al_2O_3$ | 50 | 550 | 1(Ar sweep gas) | >60 | >75 | [83] |
| | Pd based membrane supported on an $Al_2O_3$ hollow fiber support | DR | Ru/$ZrO_2$–$LaO_3$ | 5 | 525 | 1(Ar sweep gas) | >40 | – | [84] |
| | $Pd_{76}Ag_{19}Cu_5$ alloy membrane on a porous stainless-steel disk | DR | Ni-based | 15.4 | 550 | 2 | 65 | 35 | [95] |
| | Pd membrane prepared by ELP on ceramic tube support of Al and Zr | ATR | Ni/$Al_2O_3$ | 5–6 | 650 | 1($N_2$sweep gas) | 99 | – | [96] |
| Fluidized bed membrane reactor | 10 Pd–Ag commercial membranes prepared by ELP | SMR | CPO catalyst | 3.2 | 600 | 3.5 | 55 | – | [89] |
| | 5 Pd–Ag membranes supported by ELP on $ZrO_2$ | SMR | Ru/($Ce_{0.75}Zr_{0.25}O_2$) | 4 | 600 | 1.3(Vacuum) | 89 | 23 | [91] |
| | 5 commercial REB membranes | ATR | Ru/($Ce_{0.75}Zr_{0.25}O_2$) | 4 | 600 | 1.3(Vacuum) | 92 | 20 | [91] |
| | 5 Pd–Ag membranes supported by ELP on $ZrO_2$ | ATR | Ru/($Ce_{0.75}Zr_{0.25}O_2$) | 4 | 600 | 1.3(Vacuum) | 96 | 35 | [91] |
| | Pd–Ag membrane prepared by ELP on metallic Hastelloy X support | ATR | Ni/$CaAl_2O_4$ | 5 | 600 | 4 | >80 | 35 | [97] |

## Modeling outcomes from PBMR and FBMR

It is well recognized that modeling is a tool to optimize, describe, and solve complex problems. The models are categorized based on various configuration reactors, and their complexity depends on the considering number of conservation equations and assumptions. In general, need few basic equations to solve the problems by mass and energy equation, continuity equation and momentum equation. Predicting the behavior of temperature and changes in concentration by first two equations and assessing the pressure change and change in mass due to membrane permeation [98].

All reactor models can be categorized according to the degree of complexity. In particular, the PBMR model can be

divided into 1D (one-dimensional), 2D (two-dimensional) and 3D (one-dimensional) models depending on the number of dimensions. Meanwhile, for FBMR, two types of models based on two-phase or pseudo-homogeneous approaches have been proposed [98].

In general, 1D model is only considered in axial gradient direction and this model is widely used and easy to solve [99]. For example, Simakov and Sheintuch optimized membrane reactor for auto thermal process to $H_2$ production using steam methane reforming reaction in terms of membrane area and heat loss from a pseudo homogeneous one-dimensional model [100].

2D models, such as mass and heat transfer over membranes, are used when radial gradients cannot be neglected [101]. Although this model is more accurate, it is more computationally intensive and takes longer time. Two dimensional (2D) CFD model and highlighted the negative impact of polarization effects on membrane performance are studied by Tiemersma et al. [86]. Oyama and Hakarlioglu studied the packed bed membrane reactor performance, simulated work compared with 1-D and 2-D models and proposed criteria for selecting a model for use in given operating conditions. The 2D model provided better results than the 1D model, which predicts higher conversions at high pressures than those observed experimentally [102].

Roses et al. developed a 2D CFD model, simulating the fuel processor work to hydrogen production via biogas or natural gas based on realistic architecture and its analysis of the possibility of thermally integrated PBMR endothermic with hot steam gas. The significant fuel processor arrangement made favorable to attain very high efficiency, mainly appropriate for incorporation with fuel cells [103].

The 3-D models are intricate and implemented, while nosymmetrical reactors require to be modeled [104].

However, several researchers have studied PBMR and FBMR modeling over the years. Table 8 provides an overview of the mathematical model and the most important results obtained from modeling the methane steam reforming response in two types of reactors.

In summary, it should be noted that no single working model has been developed for performing MSR processes in PBMR and FBMR, and further research is needed to expand the knowledge of modeling these reactors.

## Recommendations for the production of $H_2$ from biogas

It is recognized that biogas has prodigious potential to produce $H_2$ through several reforming methods. However, the choice of a particular process will depend on many factors such as composition of the biogas, the purity required to use the $H_2$, the desired $H_2$ yield, investment availability and others. Reforming technologies exposed in this paper, such as SR, POR and ATR, are already commercialized or will soon be commercialized. New technologies that have not been integrated but are only presently used in research, for example, internal reforming of fuel cells, will take some time to commercialize [3].

Biogas as feedstock cannot be used directly in the traditional reforming technique for producing hydrogen gas due to the involvement of $H_2S$ content in biogas, resulting in catalyst poisoning and diminution in catalyst activity. It means need to remove $H_2S$ for partially treated reforming of biogas. Dry reforming method is interesting for biogas due to need for $CO_2$ content. If the $CO_2/CH_4$ ratio in the biogas is less than 1 (typical case), an additional oxidizer must be added to generate hydrogen or synthesis gas. Thus, adding oxygen according to the DOR process may be a possible method of slightly increasing $H_2$/CO ratio while recovering $H_2$ from biogas [2]. Fig. 9 shows composition and simulating biogas with the ratio of Air-Biogas close to 1, representation of simulation of biogas from dry reforming, plotted C–H–O ternary diagram and located region of carbon inside and outside.

One of the key challenges in practically entire reforming technologies is the development of catalysts that can increase their life by preventing carbon build-up in the active phase. Particularly in the DR process, the problem is even more serious because this process is often impossible due to rapid catalyst deactivation. Now from DOR, a periodic flow of $O_2$ from the catalyst bed promotes the gasification of carbon deposited top of the surface, raising availability of catalytically active sites and temporarily or partially restoring material activity. Though, in reforming cases (DR and DOR), the ratio of $H_2$/CO is nearly one, limiting the possibility of $H_2$ formation, which requires the separation of more CO from the product through the conversion reaction. This makes it challenging to acquire pure $H_2$. In addition, some of the $H_2$ generated in the DR biogas responses with $CO_2$ through a reverse shift reaction to form water ($H_2O$) and carbon monoxide (CO), which diminishes $H_2$ production yields.

Therefore, increasing $H_2$ yield from biogas through the steam reforming process, hereafter shift reactions, appears less elegant but higher practical solution, mainly in terms of diminishing CO and increasing production of $H_2$. In this situation, high obtained ratio of $H_2$/CO is related to 3 and this technique is better controlled and well-known. There are several studies to adapt biogas to SR, despite the higher content of $CO_2$, while maintaining a higher $H_2$/CO ratio close to 2–3. A similar can occur using POR and ATR, where biogas can be used directly without significantly reducing the ratio of $H_2$/CO [124–126].

Thus, ATR seems to be an attractive method for $H_2$ production through biogas regarding $H_2$ yield and energy efficiency. In addition, ATR reactors can be quickly shut down and restarted, leading to the lowest price of any other reforming process in biogas interruption and energy consumption.

Biogas is enhanced bio-methane after biogas purification, which involves composition such as $CH_4$ (93–96%), $CO_2$ (4–7%) and $H_2S$ (below 20 ppm) that permits use in all types of reforming approaches. If low content of $CO_2$ in bio-methane, the higher conversion efficiency of methane for the desired product results in higher $H_2$ yield. However, DR and DOR methods are not suitable as $CO_2$ content has to be added to reactor, which is already $CO_2$ removed in the previous step during biogas purification. Thus, SR, POR and ATR processes are the best-suggested ones [2].

Conventional feedstock biogas reforming processes requires the development of innovative catalysts with their high temperatures and catalytic activity. It must be chemically and thermally stable to prevent carbon (coke) deposition on the

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

**Table 8 – Summary of mathematical models for modeling MSR response via PBMR and FBMR.**

| Configuration of reactor | Classification | Assumptions | Outcomes | Ref. |
|---|---|---|---|---|
| PBMR | 1-D | Isothermal conditions pseudo-homogeneous membrane hydrogen perm-selective | Enhanced MSR reaction regards heat losses and membrane area. | [100] |
| PBMR | 1-D | Plug flow membrane hydrogen perm-selective no-isothermal conditions Ergun's equation | The distribution of reactive/permeating phase can be disused. | [99] |
| PBMR | 1-D | Plug flow membrane hydrogen perm-selective no-isothermal conditions Ergun's equation | The performance of the membrane is affected by hydrogen diffusion as the thickness decreases in a certain area. | [105] |
| PBMR | 1-D | Plug flow membrane hydrogen perm-selective Pseudo-homogeneous Ergun's equation | Complete conversion of methane and high $H_2$ recovery factor is lower for PBMRs than CRs. | [106] |
| PBMR | 1-D | Isothermal conditions pseudo-homogeneous membrane hydrogen perm-selective | Better prediction for permeate flow and composition of PBMR performance. | [107] |
| PBMR | 2-D | Isothermal conditions pseudo-homogeneous membrane hydrogen perm-selective | PBMR performance affected by parameters such as space velocity and temperature reaction. | [101] |
| PBMR | 2-D | Plug flow membrane hydrogen perm-selective no-isothermal conditions Ergun's equation | Showed the membrane thermal stability and its permeability for PBMR. | [108] |
| PBMR | 2-D | Pseudo-homogeneous membrane hydrogen perm-selective | The high permeability of $H_2$ emphasizes the concentration polarization effect. The porosity profile should be considered so as not to overestimate hydrogen penetration. | [86] |
| PBMR | 2-D | 2-D CFD simulation models a system based on a realistic architecture, ideal gas mixing law membrane hydrogen perm-selective | CFD modeling confirms the potential for thermal interaction of membrane hydrogen reformers with hot flue gases in architectures designed for use in PEMFC-based micro-CHP systems. | [103] |
| FBMR | 1-D | Membrane $H_2$ perm-selective dense catalytic bed composed of two phases as dense and bubble phase, reaction takes place at dense phase | The capacity of membrane and its position between catalyst and free board area have been considered as an important design parameter in this configuration. | [109] |
| FBMR | 1-D | Membrane hydrogen perm-selective constant temperature pressure gradients are disregard | The modeling and simulation that have been carried out with Aspen Plus show a decent agreement with outcomes through the literature. | [110] |
| FBMR | 1-D | Membrane hydrogen perm-selective constant temperature pressure gradients are discount | Comparison with CR, FBMR performs biogas steam reforming reaction to achieve much higher hydrogen production. | [111] |

surface and slow down the decontamination process. Traditional reforming methods using feedstock as biogas need to advance catalyst with higher catalytic activity at higher temperatures, which must be stable chemically and thermally to prevent carbon (coke) deposition on the surface and slow down the decontamination process. In terms of treatment of the reforming process, it needs to reactivate outside or inside the catalyst bed and require the periodical replacement due to its permanent deactivation, resulting in slow or stop production. This is the main complication in the reforming process for hydrogen production on a large scale.

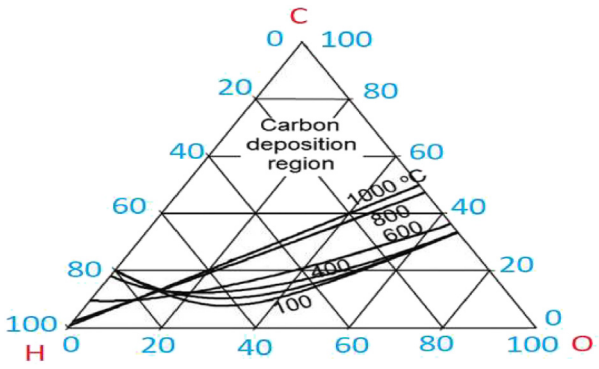



**Fig. 9 – Ternary C–H–O diagram, C–H–O showing the possibility of coke formation for biogas internal reforming in SOFC [2,123].**

## Hydrogen utility in fuel cells

Most studied hydrogen production methods are utilized by hydrogen and syngas in fuel cell. A fuel cell is a conversion of electrochemical devices which convert $H_2$ and $O_2$ into water ($H_2O$) and produce electricity by the process. Unlike batteries, fuel cells do not store energy but directly convert the chemical energy of fuel into electricity without combustion, which is more efficient and produces very low pollution emissions [2]. It is two times more efficient than turbines and internal

combustion engines and has become the most promising conversion device [127].

Generally, there are two fuel cells: proton exchange membrane fuel cells (PEMFCs) and solid oxide fuel cells (SOFC). PEMFCs contain polymer-based electrolytes such as Nafion that provide good resistance to gas permeation. The low operating temperature of the PEMFC allows rapid start-up, and the use of exotic materials required for other types of fuel cells is not required for BoP (Plant Balance) because the cells have no corrosive components [8].

Additional advantages over Alkaline Fuel Cells (AFCs), Phosphoric Acid Fuel Cells (PAFCs) and Molten Carbonate Fuel Cells (MCFCs) include limited operating temperatures and a significant reduction in corrosion problems due to the use of solid electrolytes. In addition, PEMFCs are very small and light compared to other fuel cells, making them perfect when size and portability are leading parameters for applications such as laptops, computers, automobiles, and cell phones [128].

In contrast, SOFC has more versatile fuel cells and greater electrical efficiency than nearly 55%. SOFC could be effectively united with other new technology due to high operating temperatures like microturbines to allow more generation of electricity (efficient more than 70%) [129]. It can also be feasible to provide heat and electricity to coldest region of the world when used with combination of heat and power plants. SOFCs are made of ceramic materials and no moving elements, are free from maintenance and have more reliable systems [130]. It has more benefits, such as a less expensive catalyst, suitable for stationary power generation due to its heavier weight than PEMFCs. Biogas is effectively used to produce $H_2$ intended for fuel cells [2].

The main key benefits of using biogas in fuel cells are as follows.

- ✓ The availability of biogas in huge amounts: Biogas is produced from renewable biomass and animal waste; therefore, the cost of producing this resource is very low.
- ✓ Fuel cost is low as generated from biogas, so fuel cell power cost is also low.
- ✓ Power generated from internal combustion engine via biogas is reduced due to the presence of $CO_2$ content in biogas, but there is no need to remove $CO_2$ to generate energy using fuel cells. Indeed, the presence of $CO_2$ in biogas is useful for the operation of fuel cells, for example, SOFC.
- ✓ More possibility in small scale generation.
- ✓ The operation is in silent mode due to no moving parts in fuel cells.
- ✓ The purification system is simple.
- ✓ Power generation is also possible in remote area.
- ✓ Biogas operated fuel cell power plant (2.8 MW) is running in California, US.

In short, it can be said that biogas can be effectively used in fuel cells. When biogas is used in fuel cells, removal of $CO_2$ is not required. $CO_2$ is advantageous for the production of pure $H_2$. Fuel cells (SOFC) can directly use biogas and produce stable energy [131]. The biggest drawback of this process is that the entire process must be stopped for maintenance in the event of a fuel cell failure. In the hydrogen production process, carbon from the internal reforming process is deposited on the anode's surface and affects the fuel cell's life. PEMFC requires external generation of $H_2$. This option is applicable to supply $H_2$ from SOFC and PEMFC fuel cells.

## Recent status of biogas reforming for $H_2$ production from conventional reactors and membrane reactors

Biogas, created by the decay of organic substances by microorganisms, is a renewable fuel that can be used to produce clean fuels such as hydrogen. There is a growing interest in hydrogen production through biogas. Recently, hydrogen production from biogas using conventional reformers has been extensively studied.

Chattanathan et al. studied biogas reforming for hydrogen production through the biogas dry reforming (BDR) method and examined the concentration of $H_2S$ as percentage of mol. such as 0.5,1 and 1.50 and various temperatures like 650,750 and 850 °C arrange for the conversion of $CH_4$ and $CO_2$. Conversion effect of $CH_4$ and $CO_2$ decreased using small amount of $H_2S$ compared to case where there was no $H_2S$ as well as $H_2S$ concentration increased with (0–0.5% of mol.) as reduced conversion of $CH_4$ and $CO_2$ from (67-19%) and (87-22%) respectively [132].

Izquierdo et al. inspected the various catalysts such as Rh–Ni and Ni supported cylindrical and disc morphologies for few reforming techniques using biogas such as tri-reforming and oxidative reforming and steam reforming in order to hydrogen production. The disc catalyst was more active at lower S/C ratios in steam methane process, while cylindrical catalyst was more active in dry reforming process. Additionally, Rh catalyst showed better performance for all reforming processes and found Rh–Ni catalyst was a better catalyst in cylindrical Zeolite [133].

Cipiti et al. considered a mathematical model for SR reactor using biogas. The molar ratios and inlet temperatures of reactant by reactor performance were studied and analysed parameters; then, the experimental results validated this model using catalyst ($Ni/CeO_2$) and temperature range from 700 to 900 °C. The molar ratio $H_2/CO$ and $CH_4$ conversion change from 4.5 to 2.8 and 65.5%–99.6%, respectively, at temperatures above 700 °C with a ratio of S/C equal to 3. Moreover, when increasing S/C ratio from 1 to 5, $H_2/CO$ and $CH_4$ conversion change from 1.6 to 3.8, at temperature 900 °C [28].

Pawar et al. investigated dry reforming of biogas and inspected catalyst deactivation due to sulphur poisoning and coke formation. This study was examined for various parameters such as temperature (700–800 °C), ratio of $CH_4/CO_2$ (1.5 and 2), and content of $H_2S$ in biogas (0.5 and 10 ppm). The deactivation of catalyst is more active and faster due to sulphur content. The deactivation is irreversible at 700 °C, while the catalyst activation is restored by removing sulphur content from biogas feedstock at 800 °C. In addition, the dry reforming method from biogas under these conditions results in the formation of carbon in the form of multi-walled nanotubes [134].

Tuna et al. studied catalysts (3.1% Ru/g–$Al_2O_3$ and 7.4% Ni/$NiAl_2O_4$/g-$Al_2O_3$) using biogas in steam reformer. This catalyst

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

exhibits higher $H_2$ and CO selectivities, higher $CH_4$ conversion, higher resistance to carbon formation, low deactivation, and no CO or $CO_2$ formation. As the $CO_2$ concentration increased, the $H_2$/CO ratio for catalyst as Ru at 700 °C and Ni at 800 °C reduced from 3 to 2.4 and 2.9–2.4, respectively [135].

Calgora and Perez-Lopez estimated the catalyst (Ni-M-Al) in the dry reforming process using synthetic biogas for $H_2$ production. The investigation has been done of various catalyst compositions as Ni-55%, Al-33% and other elements as 3rd component (La, Ca, M, Mg, Zn, Co, Li). La containing catalysts offer great stability in reforming reactions due to their high sintering resistance and high acidity [136].

Meanwhile, there is a growing interest on $H_2$ production using membrane reactors due to several of the aforementioned advantages over conventional reactors. However, most research in literature is devoted to producing hydrogen from methane through membrane reactors. Research on the production of hydrogen from biogas using membrane reactors is sparse. For example, Silva et al. established a mathematical model of a fixed bed membrane reactor (FBMR) to produce hydrogen from biogas using Pd–Ag hydrogen selective membranes. The results indicated that the yields of $H_2$ and CO below 600 °C are similar with and without hydrogen permeation. $H_2$ permeation at 600 °C, $H_2$ yields, and $CH_4$ conversion were touched 83 and 113.0% higher than the case without $H_2$ permeation [137]. Iulianelli et al. examined membrane ($Pd/Al_2O_3$) in membrane reactor via biogas steam reforming (BSR) for pure hydrogen production. The study found that permeated $H_2$ was approximately 96% at various conditions such as temperature (380 °C), pressure (2 bar) and $H_2O/CH_4$ (3/1), although hydrogen recovery and methane conversion were more than 20% and 15%, respectively. Conversely, at temperature (450 °C), pressure (3.5 bar) and $H_2O/CH_4$ (4/1) increased the hydrogen recovery and methane conversion 70% and 30%, respectively [138]. Nooijer et al. established a 1-D and two-phase model of fluidized bed membrane reactor (FBMR) and studied pure hydrogen production using biogas in the steam reforming process. The model results were validated with experimental values at temperature 435–535 °C, pressure 2–5 bars, and $CO_2/CH_4$ ratio up to 0.90 [139]. Summary of various studies on several methods of $H_2$ production via biogas in the literature as in Table 9.

**Table 9 – $H_2$ production methods via biogas (CR = Conventional Reactor and MR = Membrane Reactor).**

| Methods | Temperature (°C) | Catalyst | $CH_4$ conversion (%) | $H_2$/CO | Composition of Biogas (%) | Type of reactor | Ref. |
|---|---|---|---|---|---|---|---|
| DR | 800 | (Reformax® 250 | 67 | 0.98 | $CH_4$-59, $CO_2$-39,$N_2$-2 | CR (Fixed Bed) | [132] |
| SR,DR,POR | 800 | Rh–Ni based on cylindrical Zeolite L | 86,65,90 | 1.6,0.9,1.1 | $CH_4$-60, $CO_2$-40 | CR (Fixed Bed) | [133] |
| SR | 900 | 7.5 wt%Ni/$CeO_2$ | 99.7 | 2.8 | $CH_4$-60, $CO_2$-40 | CR (Fixed Bed) | [28] |
| DR | 800 | Ni/- $Al_2O_3$ | 72 | 1.2 | $CH_4$-60, $CO_2$-40 | CR (Fixed Bed) | [134] |
| SR | 700,850 | 3.1% Ru/g–$Al_2O_3$ and 7.4% Ni/$NiAl_2O_4$/g-$Al_2O_3$ | 60,60 | 3,2.9 | $CH_4$-60, $CO_2$-40 | CR (Fixed Bed) | [135] |
| SR | 600 | Ni (3.31 wt%)/γ- $Al_2O_3$) | 47 | – | $CH_4$-74, $CO_2$-26 | MR(Fixed Bed) | [137] |
| SR | 380–450 | Ni (25 wt%)/$Al_2O_3$ | 15–34 | – | $CH_4$-60, $CO_2$-40 | MR(Fixed Bed) | [138] |
| SR | 435–535 | – | 30–70 | – | $CH_4$-53, $CO_2$-47 | MR(Fluidized Bed) | [139] |

## Future prospect of biogas reforming in research direction

Broader $H_2$ production from biogas conversion technologies will require additional efforts on a larger scale. Based on literature survey, the recommendations or suggestions from several investigators are the following recommendations for possible research directions.

- Design and development of reactor and process used in biogas reforming offer effective hydrogen production with less energy consumption.
- Assessing the return on investment and operating costs of industrial scale production facilities requires large-scale studies of advanced processes.

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

- It has been found that operating parameters such as catalyst characteristic and performance, optimum temperature, presence of impurities and concentration determine the limits of achieving efficient hydrogen production.
- Modeling of CFD and COMSOL multiphysics are open research areas that can be utilized to validate existing experimental results.
- Some researchers prefer biological methods of $H_2$ production to chemical methods due to the possibility of using sunlight, $CO_2$ and organic waste as substrates to transition to a green environment in moderate conditions requiring more consideration.

## Conclusion

Large-scale development of $H_2$ energy technologies is needed for sustainable decentralized power generation in rural and remote areas. Biogas has abundant potential for inexpensive and environmentally friendly $H_2$ production, especially in developing countries like India. Biogas reforming is an auspicious process for the production of green hydrogen gas as well as for reducing overburden on natural gas. Currently, Hydrogen production has become the most important research area, especially renewable energies due to globally growing energy demand, depletion of fossil fuels, environmental issues, and progressive fuel cell technology.

Hydrogen production through the chemical reforming method appears to be more preferred in the coming decades than other costly and not fully developed processes. The chemical reforming methods include steam reforming (SR), partial oxidative reforming (POR), auto-thermal reforming (ATR), dry reforming (ATR), and dry oxidative reforming (DOR). However, a traditional reformer with a separation unit for purification and two separate water-gas shift reactors requires a lot of space, complex control, and lower system efficiency. Thus, an innovative membrane reactor that allows simultaneous hydrogen production and separation in a single reactor is more attractive than traditional reforming. The following conclusions are prepared by evaluation of literature review.

- The main problems with biogas reforming methods, mentioned in most of the published studies, are associated with deactivation of catalyst due to coke formation on surface of catalyst and poisoning effect with sulphur-containing materials, which can lead to reduction of hydrogen production.
- The biogas refinement phase is very important and needs to eradicate the destructive effects of ($H_2S$) hydrogen sulfide, regardless of the reforming process.
- The catalyst must be designed for longer life to prevent carbon formation on the active catalyst's surface in all reforming methods used for the hydrogen generation reactions. The Catalysts (Nickel-based) are main ideal catalysts for all reforming processes due to low cost.
- Auto-thermal reforming methods can be more attractive because of their high hydrogen yield and conversion efficiency. Though, this process has a difficult management arrangement.
- The biogas reforming methods such DR and DOR have proven to be more attractive for hydrogen production. However, these methods are undesirable due to carbon establishment on catalyst surface and the need for an additional oxidizing agent.
- Most of the $H_2$ production through biogas is carried out via traditional reformers based on steam reforming processes. Though, it is also probable to use an innovative membrane reactor to yield hydrogen through the biogas at a lower temperature in a single stage (i.e., reforming and separation process occur in one place). Further research is needed to develop materials and designs for membrane reactors to produce hydrogen through biogas.
- There is a growing interest on $H_2$ production using membrane reactors due to several of aforementioned advantages over conventional reactors. Additionally, most of the literature research is devoted to producing hydrogen from methane and biogas by membrane reactors.
- In high-temperature fuel cells such as SOFCs, the tolerance for CO generated during internal biogas reforming process decreases above 50 ppm, leading to carbon deposition on the catalytic electrodic, which requires better control of the chemical composition of incoming biogas.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

Authors are highly grateful to Department of Mechanical Engineering and Center for Energy and Environment, Delhi Technological University Delhi (India), for providing basic facilities to compile this work.

## REFERENCES

[1] Berndes G, Hoogwijk M, Van den Broek R. The contribution of biomass in the future global energy supply: a review of 17 studies. Biomass Bioenergy 2003;25(1):1–28.
[2] Alves HJ, Junior CB, Niklevicz RR, Frigo EP, Frigo MS, Coimbra-Araújo CH. Overview of hydrogen production technologies from biogas and the applications in fuel cells. Int J Hydrogen Energy 2013;38(13):5215–25.
[3] Holladay JD, Hu J, King DL, Wang Y. An overview of hydrogen production technologies. Catal Today 2009;139(4):244–60.
[4] Göransson K, Söderlind U, He J, Zhang W. Review of syngas production via biomass DFBGs. Renew Sustain Energy Rev 2011;15(1):482–92.
[5] Li J, Smith KJ. Methane decomposition and catalyst regeneration in a cyclic mode over supported Co and Ni catalysts. Appl Catal Gen 2008;349(1–2):116–24.
[6] Gallucci F, Fernandez E, Corengia P, van Sint Annaland M. Recent advances on membranes and membrane reactors for hydrogen production. Chem Eng Sci 2013;92:40–66.

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

[7] Nalbant Y, Colpan CO. An overview of hydrogen production from biogas. Accelerating the Transition to a 100% Renewable Energy Era 2020:355–73.
[8] Nahar G, Mote D, Dupont V. Hydrogen production from reforming of biogas: review of technological advances and an Indian perspective. Renew Sustain Energy Rev 2017;76:1032–52. https://doi.org/10.1016/j.rser.2017.02.031.
[9] Jothi G, Pugalendhi S, Poornima K, Rajendran G. Management of root-knot nematode in tomato Lycopersicon esculentum, Mill., with biogas slurry. Bioresour Technol 2003;89(2):169–70.
[10] Lohan SK, Dixit J, Kumar R, Pandey Y, Khan J, Ishaq M, Kumar D. Biogas: a boon for sustainable energy development in India′ s cold climate. Renew Sustain Energy Rev 2015;43:95–101.
[11] Rao PV, Baral SS, Dey R, Mutnuri S. Biogas generation potential by anaerobic digestion for sustainable energy development in India. Renew Sustain Energy Rev 2010;14(7):2086–94.
[12] Gregorie EF, Lamb JJ, Lien KM, Pollet BG, Burheim OS. Hydrogen and biogas. In: Micro-optics and energy. Cham: Springer; 2020. p. 131–55.
[13] Di Marcoberardino G, Foresti S, Binotti M, Manzolini G. Potentiality of a biogas membrane reformer for decentralized hydrogen production. Chemical Engineering and Processing-Process Intensification 2018;129:131–41.
[14] Shanmugam K, Tysklind M, Upadhyayula VK. Use of liquefied biomethane (LBM) as a vehicle fuel for road freight transportation: a case study evaluating environmental performance of using LBM for operation of tractor trailers. Procedia CIRP 2018;69:517–22. https://doi.org/10.1016/j.procir.2017.11.133.
[15] Pen MA, Gomez JP, Fierro JG. New catalytic routes for syngas and hydrogen production. Appl Catal Gen 1996;144(1–2):7–57.
[16] Cruz PL, Navas-Anguita Z, Iribarren D, Dufour J. Exergy analysis of hydrogen production via biogas dry reforming. Int J Hydrogen Energy 2018;43(26):11688–95.
[17] Patinvoh RJ, Osadolor OA, Chandolias K, Horváth IS, Taherzadeh MJ. Innovative pretreatment strategies for biogas production. Bioresour Technol 2017;224:13–24.
[18] Nikolaidis P, Poullikkas A. A comparative overview of hydrogen production processes. Renew Sustain Energy Rev 2017;67:597–611.
[19] Nawfal M. Valorisation catalytique du biogaz pour une énergie propre et renouvelable. Université du Littoral Côte d'Opale; 2015, January.
[20] Chen L, Zhu Q, Hao Z, Zhang T, Xie Z. Development of a Co–Ni bimetallic aerogel catalyst for hydrogen production via methane oxidative $CO_2$ reforming in a magnetic assisted fluidized bed. Int J Hydrogen Energy 2010;35(16):8494–502.
[21] Yang L, Ge X, Wan C, Yu F, Li Y. Progress and perspectives in converting biogas to transportation fuels. Renew Sustain Energy Rev 2014;40:1133–52.
[22] Angeli SD, Turchetti L, Monteleone G, Lemonidou AA. Catalyst development for steam reforming of methane and model biogas at low temperature. Appl Catal B Environ 2016;181:34–46.
[23] Chang ACC, Lee KY. Biogas reforming by the honeycomb reactor for hydrogen production. Int J Hydrogen Energy 2016;41(7):4358–65.
[24] Chiodo V, Maisano S, Zafarana G, Urbani F. Effect of pollutants on biogas steam reforming. Int J Hydrogen Energy 2017;42(3):1622–8.
[25] Roy PS, Park CS, Raju AS, Kim K. Steam-biogas reforming over a metal-foam-coated (Pd–Rh)/($CeZrO_2$– $Al_2O_3$) catalyst compared with pellet type alumina-supported Ru and Ni catalysts. J CO2 Util 2015;12:12–20.
[26] Ahmed S, Lee SH, Ferrandon MS. Catalytic steam reforming of biogas–effects of feed composition and operating conditions. Int J Hydrogen Energy 2015;40(2):1005–15.
[27] Lin KH, Chang HF, Chang ACC. Biogas reforming for hydrogen production over mesoporous Ni2xCe1– $xO_2$ catalysts. Int J Hydrogen Energy 2012;37(20):15696–703.
[28] Cipiti F, Barbera O, Briguglio N, Giacoppo G, Italiano C, Vita A. Design of a biogas steam reforming reactor: a modelling and experimental approach. Int J Hydrogen Energy 2016;41(27):11577–83.
[29] Effendi A, Zhang ZG, Hellgardt K, Honda K, Yoshida T. Steam reforming of a clean model biogas over Ni/$Al_2O_3$ in fluidized-and fixed-bed reactors. Catal Today 2002;77(3):181–9.
[30] Izquierdo U, Barrio VL, Lago N, Requies J, Cambra JF, Güemez MB, Arias PL. Biogas steam and oxidative reforming processes for synthesis gas and hydrogen production in conventional and microreactor reaction systems. Int J Hydrogen Energy 2012;37(18):13829–42.
[31] Cimino S, Lisi L, Russo G. Effect of sulphur during the catalytic partial oxidation of ethane over Rh and Pt honeycomb catalysts. Int J Hydrogen Energy 2012;37(14):10680–9.
[32] Laosiripojana N, Charojrochkul S, Kim-Lohsoontorn P, Assabumrungrat S. Role and advantages of $H_2S$ in catalytic steam reforming over nanoscale $CeO_2$-based catalysts. J Catal 2010;276(1):6–15.
[33] Appari S, Janardhanan VM, Bauri R, Jayanti S. Deactivation and regeneration of Ni catalyst during steam reforming of model biogas: an experimental investigation. Int J Hydrogen Energy 2014;39(1):297–304.
[34] Ashrafi M, Pfeifer C, Pröll T, Hofbauer H. Experimental study of model biogas catalytic steam reforming: 2. Impact of sulfur on the deactivation and regeneration of Ni-based catalysts. Energy Fuel 2008;22(6):4190–5.
[35] Chaubey R, Sahu S, James OO, Maity S. A review on development of industrial processes and emerging techniques for production of hydrogen from renewable and sustainable sources. Renew Sustain Energy Rev 2013;23:443–62.
[36] Meng B, Zhang H, Zhao Z, Wang X, Jin Y, Liu S. A novel LaGa0. 65Mg0. 15Ni0.20$O_3$–δ perovskite catalyst with high performance for the partial oxidation of methane to syngas. Catal Today 2016;259:388–92.
[37] Pantaleo G, La Parola V, Deganello F, Calatozzo P, Bal R, Venezia AM. Synthesis and support composition effects on $CH_4$ partial oxidation over Ni–CeLa oxides. Appl Catal B Environ 2015;164:135–43.
[38] Chen WH, Lin SC. Biogas partial oxidation in a heat recirculation reactor for syngas production and $CO_2$ utilization. Appl Energy 2018;217:113–25.
[39] Tsyganok AI, Inaba M, Tsunoda T, Suzuki K, Takehira K, Hayakawa T. Combined partial oxidation and dry reforming of methane to synthesis gas over noble metals supported on Mg–Al mixed oxide. Appl Catal Gen 2004;275(1–2):149–55.
[40] Moral A, Reyero I, Alfaro C, Bimbela F, Gandía LM. Syngas production by means of biogas catalytic partial oxidation and dry reforming using Rh-based catalysts. Catal Today 2018;299:280–8.
[41] Rafiq MH, Hustad JE. Experimental and thermodynamic studies of the catalytic partial oxidation of model biogas using a plasma-assisted gliding arc reactor. Renew Energy 2011;36(11):2878–87.
[42] Chen WH, Lin SC. Characterization of catalytic partial oxidation of methane with carbon dioxide utilization and excess enthalpy recovery. Appl Energy 2016;162:1141–52.

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

[43] Meshkani F, Rezaei M, Andache M. Investigation of the catalytic performance of Ni/MgO catalysts in partial oxidation, dry reforming and combined reforming of methane. J Ind Eng Chem 2014;20(4):1251–60.
[44] Asencios YJ, Rodella CB, Assaf EM. Oxidative reforming of model biogas over NiO–$Y_2O_3$–$ZrO_2$ catalysts. Appl Catal B Environ 2013;132:1–12.
[45] Dedov AG, Loktev AS, Komissarenko DA, Mazo GN, Shlyakhtin OA, Parkhomenko KV, Moiseev II. Partial oxidation of methane to produce syngas over a neodymium–calcium cobaltate-based catalyst. Appl Catal Gen 2015;489:140–6.
[46] Araki S, Hino N, Mori T, Shimizu T, Hikazudani S. Reforming reactions of model biogas over honeycomb supported Ni catalysts. J Jpn Petrol Inst 2009;52(3):120–7.
[47] Vita A, Pino L, Cipitì F, Laganà M, Recupero V. Biogas as renewable raw material for syngas production by tri-reforming process over $NiCeO_2$ catalysts: optimal operative condition and effect of nickel content. Fuel Process Technol 2014;127:47–58.
[48] Izquierdo U, Barrio VL, Requies J, Cambra JF, Güemez MB, Arias PL. Tri-reforming: a new biogas process for synthesis gas and hydrogen production. Int J Hydrogen Energy 2013;38(18):7623–31.
[49] Bawornruttanaboonya K, Devahastin S, Mujumdar AS, Laosiripojana N. Comparative numerical evaluation of autothermal biogas reforming in conventional and split-and-recombine microreactors. Int J Hydrogen Energy 2018;43(51):22874–84.
[50] Camacho YM, Bensaid S, Lorentzou S, Vlachos N, Pantoleontos G, Konstandopoulos A, Werzner E. Development of a robust and efficient biogas processor for hydrogen production. Part 2: experimental campaign. Int J Hydrogen Energy 2018;43(1):161–77.
[51] Lino AVP, Assaf EM, Assaf JM. Hydrotalcites derived catalysts for syngas production from biogas reforming: effect of nickel and cerium load. Catal Today 2017;289:78–88.
[52] Luneau M, Gianotti E, Meunier FC, Mirodatos C, Puzenat E, Schuurman Y, Guilhaume N. Deactivation mechanism of Ni supported on Mg-Al spinel during autothermal reforming of model biogas. Appl Catal B Environ 2017;203:289–99.
[53] Lai Ming-Pin, Lai Wei-Hsiang, Horng Rong-Fang, Chen Chen-Yu, Chiu Wei-Cheng, Su Siou-Sheng, Chang You-Ming. Experimental study on the performance of oxidative dry reforming from simulated biogas. Energy Proc 2012;29:225–33.
[54] Lau CS, Tsolakis A, Wyszynski ML. Biogas upgrade to syn-gas ($H_2$–CO) via dry and oxidative reforming. Int J Hydrogen Energy 2011;36(1):397–404.
[55] Foo SY, Cheng CK, Nguyen TH, Adesina AA. Syngas production from $CH_4$ dry reforming over Co–Ni/$Al_2O_3$ catalyst: coupled reaction-deactivation kinetic analysis and the effect of $O_2$ co-feeding on $H_2$: CO ratio. Int J Hydrogen Energy 2012;37(22):17019–26.
[56] Nagaraja BM, Bulushev DA, Beloshapkin S, Ross JR. The effect of potassium on the activity and stability of Ni–MgO–$ZrO_2$ catalysts for the dry reforming of methane to give synthesis gas. Catal Today 2011;178(1):132–6.
[57] Fouskas A, Kollia M, Kambolis A, Papadopoulou C, Matralis H. Boron-modified Ni/$Al_2O_3$ catalysts for reduced carbon deposition during dry reforming of methane. Appl Catal Gen 2014;474:125–34.
[58] Xu J, Zhou W, Li Z, Wang J, Ma J. Biogas reforming for hydrogen production over nickel and cobalt bimetallic catalysts. Int J Hydrogen Energy 2009;34(16):6646–54.
[59] Arbag H, Yasyerli S, Yasyerli N, Dogu G, Dogu T. Enhancement of catalytic performance of Ni based mesoporous alumina by Co incorporation in conversion of biogas to synthesis gas. Appl Catal B Environ 2016;198:254–65.
[60] Asencios YJ, Elias KF, Assaf EM. Oxidative-reforming of model biogas over NiO/$Al_2O_3$ catalysts: the influence of the variation of support synthesis conditions. Appl Surf Sci 2014;317:350–9.
[61] Sengupta S, Ray K, Deo G. Effects of modifying Ni/$Al_2O_3$ catalyst with cobalt on the reforming of $CH_4$ with $CO_2$ and cracking of $CH_4$ reactions. Int J Hydrogen Energy 2014;39(22):11462–72.
[62] Bereketidou OA, Goula MA. Biogas reforming for syngas production over nickel supported on ceria–alumina catalysts. Catal Today 2012;195(1):93–100.
[63] Asencios YJ, Bellido JD, Assaf EM. Synthesis of NiO–MgO–$ZrO_2$ catalysts and their performance in reforming of model biogas. Appl Catal Gen 2011;397(1–2):138–44.
[64] Xu J, Zhou W, Li Z, Wang J, Ma J. Biogas reforming for hydrogen production over a Ni–Co bimetallic catalyst: effect of operating conditions. Int J Hydrogen Energy 2010;35(23):13013–20.
[65] Zimicz MG, Reznik BA, Larrondo SA. Conversion of biogas to synthesis gas over NiO/$CeO_2$–$Sm_2O_3$ catalysts. Fuel 2015;149:95–9.
[66] Roh HS, Eum IH, Jeong DW. Low temperature steam reforming of methane over Ni–Ce (1–x) Zr (x) $O_2$ catalysts under severe conditions. Renew Energy 2012;42:212–6.
[67] Avraam DG, Halkides TI, Liguras DK, Bereketidou OA, Goula MA. An experimental and theoretical approach for the biogas steam reforming reaction. Int J Hydrogen Energy 2010;35(18):9818–27.
[68] Kolbitsch P, Pfeifer C, Hofbauer H. Catalytic steam reforming of model biogas. Fuel 2008;87(6):701–6.
[69] Effendi A, Hellgardt K, Zhang ZG, Yoshida T. Optimising $H_2$ production from model biogas via combined steam reforming and CO shift reactions. Fuel 2005;84(7–8):869–74.
[70] De Rogatis L, Montini T, Cognigni A, Olivi L, Fornasiero P. Methane partial oxidation on NiCu-based catalysts. Catal Today 2009;145(1–2):176–85.
[71] Choudhary VR, Mondal KC, Choudhary TV. Partial oxidation of methane to syngas with or without simultaneous steam or $CO_2$ reforming over a high-temperature stable-NiCoMgCeOx supported on zirconia–hafnia catalyst. Appl Catal Gen 2006;306:45–50.
[72] Fonseca A, Assaf EM. Production of the hydrogen by methane steam reforming over nickel catalysts prepared from hydrotalcite precursors. J Power Sources 2005;142(1–2):154–9.
[73] Corbo P, Migliardini F. Hydrogen production by catalytic partial oxidation of methane and propane on Ni and Pt catalysts. Int J Hydrogen Energy 2007;32(1):55–66.
[74] Di Marcoberardino G, Manzolini G, Guignard C, Magaud V. Optimization of a micro-CHP system based on polymer electrolyte membrane fuel cell and membrane reactor from economic and life cycle assessment point of view. Chemical Engineering and Processing-Process Intensification 2018;131:70–83.
[75] Xia Younan, Lu Y, Kamata Kaori, Gates Byron, Yin Yadong. Macroporous materials containing three-dimensionally periodic structures. Chemistry of Nanostructured Materials 2002:69–100.
[76] De Falco M, Salladini A, Palo E, Iaquaniello G. Reformer and membrane modules (RMM) for methane conversion powered by a nuclear reactor. Nuclear Power Deploy. Oper. Sustain. 2011.
[77] Helmi A, Gallucci F, van Sint Annaland M. Resource scarcity in palladium membrane applications for carbon capture in

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

integrated gasification combined cycle units. Int J Hydrogen Energy 2014;39(20):10498–506.
[78] Li A, Grace JR, Lim CJ. Preparation of thin Pd-based composite membrane on planar metallic substrate: Part I: pre-treatment of porous stainless-steel substrate. J Membr Sci 2007;298(1–2):175–81.
[79] Alavi M, Iulianelli A, Rahimpour MR, Eslamloueyan R, De Falco M, Bagnato G, Basile A. In: Basile A, Dalena F, Tong J, Nejat Veziroglu T, editors. Chapter. 8. Fixed bed membrane reactors for ultrapure hydrogen production: Modelling approach. Hydrogen Production, Separation and Purification for Energy; 2017. p. 231–57.
[80] Sarić M, van Delft YC, Sumbharaju R, Meyer DF, de Groot A. Steam reforming of methane in a bench-scale membrane reactor at realistic working conditions. Catal Today 2012;193(1):74–80.
[81] Dittmar B, Behrens A, Schödel N, Rüttinger M, Franco T, Straczewski G, Dittmeyer R. Methane steam reforming operation and thermal stability of new porous metal supported tubular palladium composite membranes. Int J Hydrogen Energy 2013;38(21):8759–71.
[82] Di Marcoberardino G, Sosio F, Manzolini G, Campanari S. Fixed bed membrane reactor for hydrogen production from steam methane reforming: experimental and modeling approach. Int J Hydrogen Energy 2015;40(24):7559–67.
[83] Silva FA, Hori CE, da Silva AM, Mattos LV, Múnera J, Cornaglia L, Lombardo E. Hydrogen production through $CO_2$ reforming of $CH_4$ over $Pt/CeZrO_2/Al_2O_3$ catalysts using a Pd–Ag membrane reactor. Catal Today 2012;193(1):64–73.
[84] García-García FR, Soria MA, Mateos-Pedrero C, Guerrero-Ruiz A, Rodríguez-Ramos I, Li K. Dry reforming of methane using Pd-based membrane reactors fabricated from different substrates. J Membr Sci 2013;435:218–25.
[85] Caravella A, Barbieri G, Drioli E. Concentration polarization analysis in self-supported Pd-based membranes. Separ Purif Technol 2009;66(3):613–24.
[86] Tiemersma TP, Patil CS, van Sint Annaland M, Kuipers JAM. Modelling of packed bed membrane reactors for autothermal production of ultrapure hydrogen. Chem Eng Sci 2006;61(5):1602–16.
[87] Deshmukh SARK, Heinrich S, Mörl L, van Sint Annaland M, Kuipers JAM. Membrane assisted fluidized bed reactors: potentials and hurdles. Chem Eng Sci 2007;62(1–2):416–36.
[88] Adris AM, Lim CJ, Grace JR. The fluidized-bed membrane reactor for steam methane reforming: model verification and parametric study. Chem Eng Sci 1997;52(10):1609–22.
[89] Roses L, Gallucci F, Manzolini G, van Sint Annaland M. Experimental study of steam methane reforming in a Pd-based fluidized bed membrane reactor. Chem Eng J 2013;222:307–20.
[90] Medrano JA, Julián I, Herguido J, Menéndez M. Pd-Ag membrane coupled to a two-zone fluidized bed reactor (TZFBR) for propane dehydrogenation on a Pt-Sn/MgAl2O4 catalyst. Membranes 2013;3(2):69–86.
[91] Fernandez E, Coenen K, Helmi A, Melendez J, Zuñiga J, Tanaka DP, Gallucci F. Preparation and characterization of thin-film Pd–Ag supported membranes for high-temperature applications. Int J Hydrogen Energy 2015;40(39):13463–78.
[92] Kume T, Ikeda Y, Iseki T, Yakabe H, Tanaka H, Hikosaka H, Ito M. Performance evaluation of membrane on catalyst module for hydrogen production from natural gas. Int J Hydrogen Energy 2013;38(14):6079–84.
[93] Patrascu M, Sheintuch M. Design concepts of a scaled-down autothermal membrane reformer for on board hydrogen production. Chem Eng J 2015;282:123–36.
[94] Kyriakides AS, Voutetakis S, Papadopoulou S, Seferlis P. Optimization of an experimental membrane reactor for low-temperature methane steam reforming. Clean Technol Environ Policy 2016;18(7):2065–75.
[95] Sumrunronnasak S, Tantayanon S, Kiatgamolchai S, Sukonket T. Improved hydrogen production from dry reforming reaction using a catalytic packed-bed membrane reactor with Ni-based catalyst and dense Pd Ag Cu alloy membrane. Int J Hydrogen Energy 2016;41(4):2621–30.
[96] Yan Y, Cui Y, Zhang L, Li L, Zhang J, Chen Y, Lin C. Experimental investigation of methane auto-thermal reforming in hydrogen-permeable membrane reactor for pure hydrogen production. Int J Hydrogen Energy 2016;41(30):13069–76.
[97] Medrano JA, Fernandez E, Melendez J, Parco M, Tanaka DAP, van Sint Annaland M, Gallucci F. Pd-based metallic supported membranes: high-temperature stability and fluidized bed reactor testing. Int J Hydrogen Energy 2016;41(20):8706–18.
[98] Iulianelli A, Liguori S, Wilcox J, Basile A. Advances on methane steam reforming to produce hydrogen through membrane reactors technology: a review. Catal Rev 2016;58(1):1–35.
[99] Caravella A, Di Maio FP, Di Renzo A. Optimization of membrane area and catalyst distribution in a permeative-stage membrane reactor for methane steam reforming. J Membr Sci 2008;321(2):209–21.
[100] Simakov DS, Sheintuch M. Model-based optimization of hydrogen generation by methane steam reforming in autothermal packed-bed membrane reformer. AIChE J 2011;57(2):525–41.
[101] Marin P, Patiño Y, Díez FV, Ordóñez S. Modelling of hydrogen perm-selective membrane reactors for catalytic methane steam reforming. Int J Hydrogen Energy 2012;37(23):18433–45.
[102] Oyama ST, Hacarlioglu P. The boundary between simple and complex descriptions of membrane reactors: the transition between 1-D and 2-D analysis. J Membr Sci 2009;337(1–2):188–99.
[103] Roses L, Manzolini G, Campanari S. CFD simulation of Pd-based membrane reformer when thermally coupled within a fuel cell micro-CHP system. Int J Hydrogen Energy 2010;35(22):12668–79.
[104] Smith RB, Muruganandam L, Shekhar SM. CFD analysis of water gas shift membrane reactor. Chem Eng Res Des 2011;89(11):2448–56.
[105] Hüppmeier J, Baune M, Thöming J. Interactions between reaction kinetics in ATR-reactors and transport mechanisms in functional ceramic membranes: a simulation approach. Chem Eng J 2008;142(3):225–38.
[106] Patel KS, Sunol AK. Modeling and simulation of methane steam reforming in a thermally coupled membrane reactor. Int J Hydrogen Energy 2007;32(13):2344–58.
[107] Patrascu M, Sheintuch M. On-site pure hydrogen production by methane steam reforming in high flux membrane reactor: experimental validation, model predictions and membrane inhibition. Chem Eng J 2015;262:862–74.
[108] De Falco M. Pd-based membrane steam reformers: a simulation study of reactor performance. Int J Hydrogen Energy 2008;33(12):3036–40.
[109] Mahecha-Botero A, Chen Z, Grace JR, Elnashaie SSEH, Lim CJ, Rakib M, Shirasaki Y. Comparison of fluidized bed flow regimes for steam methane reforming in membrane reactors: a simulation study. Chem Eng Sci 2009;64(16):3598–613.
[110] Ye G, Xie D, Qiao W, Grace JR, Lim CJ. Modeling of fluidized bed membrane reactors for hydrogen production from steam methane reforming with Aspen Plus. Int J Hydrogen Energy 2009;34(11):4755–62.

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

[111] Saebea D, Authayanun S, Patcharavorachot Y, Arpornwichanop A. Enhancement of hydrogen production for steam reforming of biogas in fluidized bed membrane reactor. 2014.
[112] Ruckenstein E, Hu YH. Methane partial oxidation over NiO/MgO solid solution catalysts. Appl Catal Gen 1999;183(1):85–92.
[113] Mosayebi Z, Rezaei M, Ravandi AB, Hadian N. Autothermal reforming of methane over nickel catalysts supported on nanocrystalline MgAl2O4 with high surface area. Int J Hydrogen Energy 2012;37(2):1236–42.
[114] Simeone M, Salemme L, Allouis C. Reactor temperature profile during autothermal methane reforming on $Rh/Al_2O_3$ catalyst by IR imaging. Int J Hydrogen Energy 2008;33(18):4798–808.
[115] Wu J, Yuan XZ, Martin JJ, Wang H, Zhang J, Shen J, Merida W. A review of PEM fuel cell durability: degradation mechanisms and mitigation strategies. J Power Sources 2008;184(1):104–19.
[116] Barrai F, Jackson T, Whitmore N, Castaldi MJ. The role of carbon deposition on precious metal catalyst activity during dry reforming of biogas. Catal Today 2007;129(3–4):391–6.
[117] Souza MM, Schmal M. Autothermal reforming of methane over $Pt/ZrO_2/Al_2O_3$ catalysts. Appl Catal Gen 2005;281(1–2):19–24.
[118] Eltejaei H, Bozorgzadeh HR, Towfighi J, Omidkhah MR, Rezaei M, Zanganeh R, Ghalam AZ. Methane dry reforming on $Ni/Ce0.75Zr0.25O_2–MgAl_2O_4$ and $Ni/Ce0.75Zr0.25O_2–\gamma$-alumina: effects of support composition and water addition. Int J Hydrogen Energy 2012;37(5):4107–18.
[119] Serrano-Lotina A, Martin AJ, Folgado MA, Daza L. Dry reforming of methane to syngas over La-promoted hydrotalcite clay-derived catalysts. Int J Hydrogen Energy 2012;37(17):12342–50.
[120] Lucrédio AF, Assaf JM, Assaf EM. Reforming of a model biogas on Ni and Rh–Ni catalysts: effect of adding La. Fuel Process Technol 2012;102:124–31.
[121] Zhang X, Lee CSM, Hayward DO, Mingos DMP. Oscillatory behaviour observed in the rate of oxidation of methane over metal catalysts. Catal Today 2005;105(2):283–94.
[122] Choudhary VR, Mondal KC. $CO_2$ reforming of methane combined with steam reforming or partial oxidation of methane to syngas over $NdCoO_3$ perovskite-type mixed metal-oxide catalyst. Appl Energy 2006;83(9):1024–32.
[123] Shiratori Y, Ijichi T, Oshima T, Sasaki K. Internal reforming SOFC running on biogas. Int J Hydrogen Energy 2010;35(15):7905–12.
[124] Zhu J, van Ommen JG, Knoester A, Lefferts L. Effect of surface composition of yttrium-stabilized zirconia on partial oxidation of methane to synthesis gas. J Catal 2005;230(2):291–300.
[125] Trimm DL. Catalysts for the control of coking during steam reforming. Catal Today 1999;49(1–3):3–10.
[126] Araki S, Hino N, Mori T, Hikazudani S. Autothermal reforming of biogas over a monolithic catalyst. J Nat Gas Chem 2010;19(5):477–81.
[127] Kirubakaran A, Jain S, Nema RK. A review on fuel cell technologies and power electronic interface. Renew Sustain Energy Rev 2009;13(9):2430–40.
[128] Pérez LC, Koski P, Ihonen J, Sousa JM, Mendes A. Effect of fuel utilization on the carbon monoxide poisoning dynamics of Polymer Electrolyte Membrane Fuel Cells. J Power Sources 2014;258:122–8.
[129] Kandepu R, Imsland L, Foss BA, Stiller C, Thorud B, Bolland O. Modeling and control of a SOFC-GT-based autonomous power system. Energy 2007;32(4):406–17.
[130] Rostrup-Nielsen J. Steam reforming of hydrocarbons. A historical perspective. In: Studies in surface science and catalysis, vol. 147. Elsevier; 2004. p. 121–6.
[131] Staniforth J, Kendall K. Biogas powering a small tubular solid oxide fuel cell. J Power Sources 1998;71(1–2):275–7.
[132] Chattanathan SA, Adhikari S, McVey M, Fasina O. Hydrogen production from biogas reforming and the effect of $H_2S$ on $CH_4$ conversion. Int J Hydrogen Energy 2014;39(35):19905–11.
[133] Izquierdo U, Barrio VL, Bizkarra K, Gutierrez AM, Arraibi JR, Gartzia L, Cambra JF. Ni and RhNi catalysts supported on Zeolites L for hydrogen and syngas production by biogas reforming processes. Chem Eng J 2014;238:178–88.
[134] Pawar V, Appari S, Monder DS, Janardhanan VM. Study of the combined deactivation due to sulfur poisoning and carbon deposition during biogas dry reforming on supported Ni catalyst. Ind Eng Chem Res 2017;56(30):8448–55.
[135] Tuna CE, Silveira JL, da Silva ME, Boloy RM, Braga LB, Pérez NP. Biogas steam reformer for hydrogen production: evaluation of the reformer prototype and catalysts. Int J Hydrogen Energy 2018;43(4):2108–20.
[136] Calgaro CO, Perez-Lopez OW. Biogas dry reforming for hydrogen production over Ni-M-Al catalysts (M= Mg, Li, Ca, La, Cu, Co, Zn). Int J Hydrogen Energy 2019;44(33):17750–66.
[137] Silva FSA, Benachour M, Abreu CAM. Evaluating hydrogen production in biogas reforming in a membrane reactor. Braz J Chem Eng 2015;32:201–10.
[138] Iulianelli A, Liguori S, Huang Y, Basile A. Model biogas steam reforming in a thin Pd-supported membrane reactor to generate clean hydrogen for fuel cells. J Power Sources 2015;273:25–32.
[139] de Nooijer N, Gallucci F, Pellizzari E, Melendez J, Tanaka DAP, Manzolini G, van Sint Annaland M. On concentration polarisation in a fluidized bed membrane reactor for biogas steam reforming: modelling and experimental validation. Chem Eng J 2018;348:232–43.
[140] Gallucci F, Medrano JA, Fernandez E, Melendez J, Van Sint Annaland M, Pacheco-Tanaka DA. Advances on high temperature Pd-based membranes and membrane reactors for hydrogen purification and production. J. Membr. Sci. Res 2017;3(3):142–56.
[141] Gallucci F, Annaland MVS, Kuipers JAM. Autothermal reforming of methane with integrated $CO_2$ capture in a novel fluidized bed membrane reactor. Part 1: experimental demonstration. Top Catal 2008;51(1–4):133.
[142] Lu N, Xie D. Novel membrane reactor concepts for hydrogen production from hydrocarbons: a review. Int J Chem React Eng 2016;14(1):1–31.
[143] Braga LB, Silveira JL, Da Silva ME, Tuna CE, Machin EB, Pedroso DT. Hydrogen production by biogas steam reforming: a technical, economic and ecological analysis. Renew Sustain Energy Rev 2013;28:166–73.
[144] Hajjaji N, Martinez S, Trably E, Steyer JP, Helias A. Life cycle assessment of hydrogen production from biogas reforming. Int J Hydrogen Energy 2016;41(14):6064–75.
[145] Kabeyi MJB, Olanrewaju OA. Biogas production and applications in the sustainable energy transition. J Energy 2022. 2022.
[146] Chen ST, Kuo HI, Chen CC. The relationship between GDP and electricity consumption in 10 Asian countries. Energy Pol 2007;35(4):2611–21.
[147] Achinas S, Achinas V, Euverink GJW. A technological overview of biogas production from biowaste. Engineering 2017;3(3):299–307.

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

[148] Hahn H. Economic and ecological evaluation of biogas plant configurations for a demand oriented biogas supply for flexible power generation. 2015.
[149] Abdullah B, Abd Ghani NA, Vo DVN. Recent advances in dry reforming of methane over Ni-based catalysts. J Clean Prod 2017;162:170–85.
[150] Gao Y, Jiang J, Meng Y, Yan F, Aihemaiti A. A review of recent developments in hydrogen production via biogas dry reforming. Energy Convers Manag 2018;171:133–55.
[151] Spallina V, Pandolfo D, Battistella A, Romano MC, Annaland MVS, Gallucci F. Techno-economic assessment of membrane assisted fluidized bed reactors for pure $H_2$ production with $CO_2$ capture. Energy Convers Manag 2016;120:257–73.
[152] Gallucci F, Fernandez E, Medrano JA, Pacheco Tanaka DA, van Sint Annaland M. Pd-based membranes for high temperature applications: current status. 2016.

Please cite this article as: Kumar R et al., Overview of hydrogen production from biogas reforming: Technological advancement, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2022.08.059

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/362933866

# Performance Evaluation of Four Meta-Heuristic MPPT Algorithms for a DC-DC Converter based PV System under Dynamic Weather Conditions (DWC)

**Conference Paper** · July 2022

DOI: 10.1109/ICICCSP53532.2022.9862414

CITATIONS
0

READS
9

**3 authors**, including:

Vinay Kumar
Delhi Technological University
**2** PUBLICATIONS **0** CITATIONS

SEE PROFILE

All content following this page was uploaded by Vinay Kumar on 26 August 2022.

The user has requested enhancement of the downloaded file.

# Performance Evaluation of Four Meta-Heuristic MPPT Algorithms for a DC-DC Converter based PV System under Dynamic Weather Conditions (DWC)


Vinay Kumar
*Department of Electrical Engineering*
*Delhi Technological University*
*(Formerly Delhi College of Engineering)*
Shahbad Daulatpur, Rohini, Delhi, India
vk390677@gmail.com
vinaykumar_2k18ee240@dtu.ac.in

Yash Dagar
*Department of Electrical Engineering*
*Delhi Technological University*
*(Formerly Delhi College of Engineering)*
Shahbad Daulatpur, Rohini, Delhi, India
yashdagar22899@gmail.com
yashdagar_2k18ee248@dtu.ac.in

Prakash Chittora
*Department of Electrical Engineering*
*Delhi Technological University*
*(Formerly Delhi College of Engineering)*
Shahbad Daulatpur, Rohini, Delhi, India
prakashchittora@gmail.com
prakashchittora@dtu.ac.in



***Abstract*—Solar Photovoltaic (PV) Systems are one of the most favorable and advantageous power generation frameworks used worldwide to fulfill the global requirements/needs for Renewable & Clean Energy. The output power characteristics generated by the PV Systems are non-linear and are strongly dependent upon the variable environmental parameters such as the Solar Irradiation (W/m$^2$) and Temperature (ºC). Each PV System has its own ($V_{MPP}, I_{MPP}$) point defined at a specific Irradiation (W/m$^2$) and Temperature (ºC) value at which the power is optimal. Thus, in order to ensure the continuous operation of the PV System at the Maximum Power-Point (MPP), an intelligently smart Maximum Power-Point Tracking (MPPT) mechanism is mandatorily implemented along with the Photovoltaic (PV) System. This paper exhibits a comparative study on the behaviors of the four MPPT algorithms namely: Cuckoo Search Algorithm (CSA), Flower Pollination Algorithm (FPA), Grey Wolf Optimization Algorithm (GWOA), and the Particle Swarm Optimization Algorithm (PSOA) under Dynamic Weather/Environmental conditions based on fluctuations in Temperature (ºC) and Solar Irradiation (W/m$^2$) for a DC-DC Step-Up Converter based Photovoltaic (PV) System that is modeled & implemented under Matlab/Simulink Environment.**



***Keywords*—Photovoltaic (PV) System, Maximum Power-Point Tracking (MPPT), Dynamic Weather Conditions (DWC), Cuckoo Search Algorithm (CSA), Flower Pollination Algorithm (FPA), Grey Wolf Optimization Algorithm (GWOA), Particle Swarm Optimization Algorithm (PSOA), Matlab/Simulink Software tool.**


## I. Introduction

In today's technologically sophisticated global world, the demand and acknowledgment for clean and renewable Energy/Power/Electricity are continuously expanding at an ever-increasing exponential pace, owing to the continued hostile & destructive consequences of global warming and climate change. In comparison with classical Non-Renewable energy sources such as Fossil Fuels and Natural Gases, the Solar Radiant Energy is considered and thought to be a very important and favorable energy source among them, because of it being a clean, pollution-free, reliable, sustainable, and infinitely abundant in nature, Solar Energy poses zero threat to the Environment and the existing Ecosystem. The major drawback of this energy source and PV system technology is the dependency and reliance of the PV Output Power on naturally unpredictable & Dynamic Weather/Climatic & Environmental Conditions. In this paper we have focused on a comprehensive analysis and comparative study of the four Meta-Heuristic MPPT algorithms namely: Cuckoo Search Algorithm (CSA) [1], Flower Pollination Algorithm (FPA) [2], Grey Wolf Optimization Algorithm (GWOA) [3], and Particle Swarm Optimization Algorithm (PSOA) [4] for a Vikram Solar Eldora PV Module (60 Cells & 250 W) under the effect of Dynamically abrupt variations in weather parameters such as Temperature (ºC) and Solar Irradiation (W/m$^2$) for a sample time of 10 seconds. The behavior of each MPPT is intuitively studied under different climatic scenarios: (A) Standard Weather Conditions, (B) Dynamic Weather Conditions: under fixed Temperature and varying Irradiation, and part (C) Dynamic Weather Conditions: under fixed Irradiation and varying Temperature values.

This paper has been organized and presented under the following structure: in which the Section (I): Introduction - Introduces the readers to the goal and setting of the paper, the Section (II) talks about the general working of the PV System, modeling of the single diode PV Cell, Electrical Specifications of the employed Vikram Solar Eldora (VSPB.60.250.03) PV Module, and the DC-DC Boost Converter specifications, further Section (III) briefly discusses the strategies of the four MPPT methodologies, furthermore, Section (IV) presents the Simulation results and Performance Analysis of the PV System under different Dynamic Weather Conditions (DWC) finally the Conclusion in Section (V) followed by Acknowledgement and References respectively.

## II. PV System: Model and Characteristics

A typical block design/diagram illustrating the operating mechanism of a Photovoltaic (PV) System is presented in Fig. 1, comprising a Solar PV Panel, a DC-DC Converter, MPPT Control Algorithm, a PWM Control, and a Load. The solar energy illuminating the PV Panel generates PV current ($I_{pv}$) and PV Voltage ($V_{pv}$) inside the PV Panel/Module. To transfer the MPP Power from the PV Panel toward the Load, the MPPT control block absorbs the required parameters ($I_{pv}$, $V_{pv}$) based on which it in turn generates the command signal to obtain the best optimal Duty Cycle (D) such that Optimal MPP output power is achieved under all types of possible dynamic weather/climatic scenarios (DWCs) discussed here.

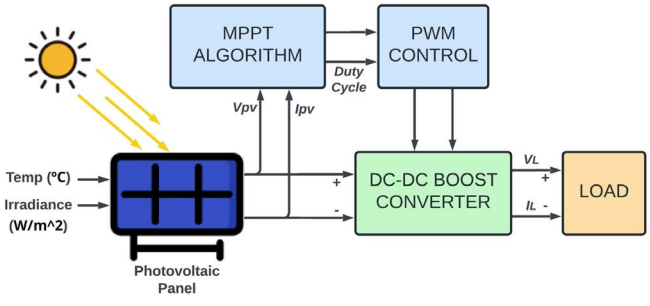



Fig. 1. General Block Design of a PV System.


978-1-6654-7258-6/22/$31.00 ©2022 IEEE

The DC-DC Boost/Step-Up Converter employed here acts as an adaptation link to ensure an efficient Maximum Power Transfer from the PV Panel to the connected DC Load. The Boost Converter implements the altered optimum value of the duty cycle adjusted by the MPPT by making the Initial impedance equal to the load impedance. As a result of which the PV System reaches the Maximum Power-Point (MPP).

### *A. Mathematical Modelling of a Single Diode PV Cell*

A PV cell is a semiconductor device, which when exposed to the Sun's Radiant Energy under high-intensity photons falling on the panel surface - creates a huge set of electron-hole pairs in the PN Junction, which further stimulates the movement of electrons and holes across the PN Junction, that in-turn generates the electric current which is known as the Photocurrent ($I_{ph}$). The Equivalent Circuit Schematic of the widely used single diode PV Cell Model constructed by the interconnection of a Current Source ($I_{ph}$), Parallel/Shunt Resistance ($R_p$), Series Resistance ($R_s$), and a Diode ($D$) has been depicted in Fig. 2. The electrical elements ($I_{ph}$, $R_p$, $R_s$, $D$) used in the schematic signifies the Photocurrent, Leakage current, the resistivity of the overall PV material, and electron-hole pair recombination, respectively. The PV cell model of the single diode has been employed here to ensure simplicity and to avoid complexity in project work and its result analysis.

The following mathematical equations validate the design for the one-diode PV Cell model depicted in Fig. 2:

Applying KCL in the PV Cell schematic (Fig. 2), the Output PV Current ($I_{pv}$) is given by the Eqn. (1)

$$I_{pv} = I_{ph} - I_d - I_{sh} \tag{1}$$

The Current ($I_d$) flowing through the Diode as leakage current is given by Eqn. (2)

$$I_d = I_o \left( e^{\frac{V_{pv} + I_{pv} R_S}{\alpha V_T}} - 1 \right) \tag{2}$$

The Current Losses ($I_{sh}$) flowing through the shunt/parallel resistance ($R_p$) is given by Eqn. (3)

$$I_{sh} = \frac{V_{pv} + I_{pv} R_S}{R_p} \tag{3}$$

The Output PV Characteristic Current ($I_{pv}$) is given and calculated by Eqn. (4)

$$I_{pv} = I_{ph} - I_o \left( e^{\frac{V_{pv} + I_{pv} R_S}{\alpha V_T}} - 1 \right) - \frac{V_{pv} + I_{pv} R_S}{R_p} \tag{4}$$

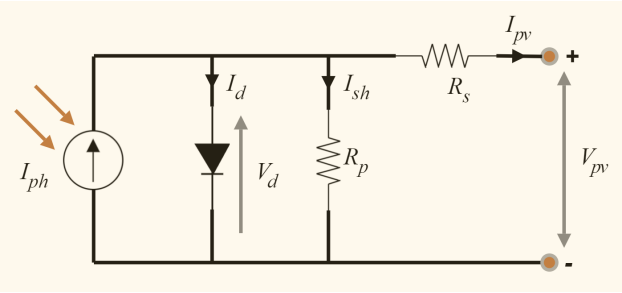



Fig. 2. Circuit Schematic of a One-Diode PV Cell.

Knowingly that,

($I_{pv}$) = PV Output Current, ($V_{pv}$) = PV Open Circuit Voltage, ($I_{ph}$) = Photocurrent, ($I_{sh}$) = Shunt Current, ($R_p$) = Parallel Resistance, ($R_s$) = Series Resistance, ($I_d$) = Diode Current, ($I_o$) = Reverse Saturation Current, ($V_T$) = Thermal Voltage, $q$ = Electronic Charge ($1.67 \times 10^{-19}$ C), ($N_s$) = Number of connected PV Cells (Series), $K$ = Boltzmann Constant ($1.3805 \times 10^{-23}$ J/K), $T$ = Temperature (STC).

### *B. Electrical Parameters of the PV Module/Panel*

The Photovoltaic (PV) System used here embodies a Vikram Solar Eldora (VSPB.60.250.03) 250 Watts Photovoltaic module by Vikram Solar Ltd. constituted by 60 PV cells mounted in series and parallel combination with each other. Moreover, the employed PV module in the PV System (Fig. 3) is connected in parallel with the bypass diode which will bypass the PV module in order to avoid the heating effect caused due to the hotspot formation, which is a phenomenon under the effect of Partial shading Irradiation Condition that can damage the PV module permanently. Various electrical characteristics & parameters of the Vikram Solar PV module employed here are summarized in Table I.

TABLE I.
ELECTRICAL PARAMETERS OF THE SOLAR 250 W PV MODULE

| Serial No. | Vikram Solar Eldora (VSPB.60.250.03) | | |
|---|---|---|---|
| | ***Parameters/Characteristics*** | ***Abbreviation*** | ***Ratings*** |
| 1. | Maximum Power | $P_{max}$ *(Watts)* | 250.14 |
| 2. | Voltage at Maximum Power | $V_{max}$ *(Volts)* | 30.58 |
| 3. | Current at Maximum Power | $I_{max}$ *(Amps.)* | 8.18 |
| 4. | Open Circuit Voltage | $V_{oc}$ *(Volts)* | 37.45 |
| 5. | Temperature Coeff. for $V_{oc}$ | $\alpha$ *(%/°C)* | -0.33 |
| 6. | Short Circuit Current | $I_{sc}$ *(Amps.)* | 8.7 |
| 7. | Temperature Coeff. for $I_{sc}$ | $\beta$ *(%/°C)* | 0.0624 |
| 8. | Number of Cells per Module | $N_{cell}$ *(Units)* | 60 |

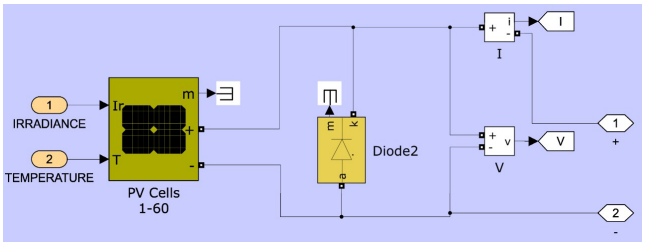



Fig. 3. PV Module (VSPB.60.250.03) Connected to the Bypass Diode.

### *C. DC-DC Boost (Step-Up) Converter for the PV System*

A DC-DC Boost (Step-Up) Converter with a Switching Frequency ($F_s$) of around 10 kHz, comprising two Capacitors C1 & C2 each of 100 µF capacitance, an Inductor L of 3mH inductance, one switching device (IGBT/MOSFET), and a Diode has been realized as a juncture between the Photovoltaic (PV) System and the DC Load, to boost the incoming voltage & implement the optimally adjusted Duty Cycle (D) by the MPPT control Algorithm, which results in successful MPP transfer from Solar PV Panels to the Load ($R_L$). The DC-DC Converter circuit is presented in Fig. 4.

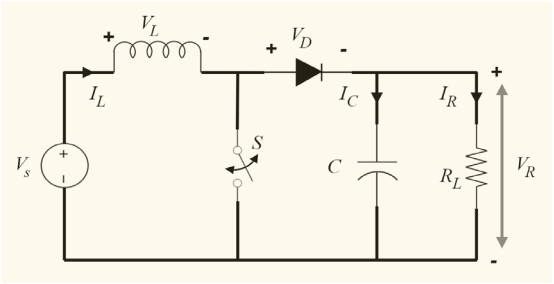



Fig. 4. Equivalent Schematic Circuit of DC-DC Step-UP Converter.

### *D. I-V & P-V Characteristics of the PV System*

Current vs Voltage (I-V) and Power vs Voltage (P-V) characteristics plots for the Vikram Solar Eldora (VSPB.60.250.03) Photovoltaic (PV) Module under Standard Weather Conditions i.e, under fixed Temperature (25ºC) and fixed irradiation (1000 W/m$^2$) are depicted in Fig. 5. The ideal MPP Power achieved under standard conditions is 250 Watts.

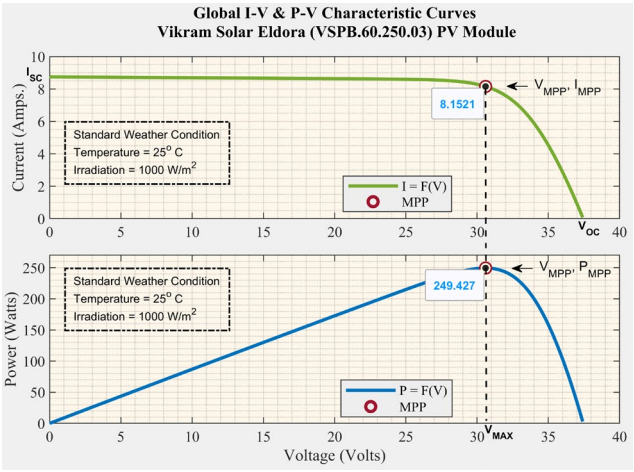



Fig. 5. I-V & P-V Characteristics of the Solar PV Module under STC.

## III. The Four Meta-Heuristic MPPT Algorithms

Understanding the extreme importance of a Maximum Power-Point Tracking (MPPT) algorithm in determining and extracting the Global MPP (GMPP) Power from the Photovoltaic (PV) Module in order to make the PV System perform at its best potential, the subsequent sections (A, B, C, and D) present a clear and concise explanation of the working strategy and implementation approach of all the four Meta-Heuristic MPPT techniques/strategies namely CSA, FPA, GWOA, and PSOA to undergo the comparative study of the Vikram Solar PV module/panel under SWC and DWC. [5]

### *A. Cuckoo Search Algorithm (CSA)*

The CSA is a Meta-Heuristic Swarm Intelligence Algorithm originally proposed by Xin-she Yang and Suash Deb in the year 2009 [1], which mimics the nature of Cuckoo bird species, these birds are parasitic in nature as they lay their eggs in the nests built by host bird species in order to let their broods be unwittingly raised by the host parents, which gradually expands the cuckoo generation fertility rate and life expectancy. The odds of the host birds discovering the alien eggs are quite low, as, over time, the cuckoos have evolved the ability to lay eggs that are identical to certain host bird species. The probability of the cuckoo eggs being discovered and tossed away by the host bird is defined as: $P_a \in (0,1)$.

This algorithm is nothing but a simulation of the Levy flight random walk search done by the Cuckoo bird in search of host nests. This algorithmic process begins by initially defining the population for the host nest, and then calculating their respective fitness values. Then, using Levy Flight, new solutions for each host nest are developed, using which the cuckoo keeps on updating its aerial location until the Optimal Global Best point is reached. This approach continuously searches for better solutions and replaces the less optimal ones with better ones to eventually achieve the optimal point.

Cuckoo Search Algorithm is implemented using Eqn. (5)

$$x_i^{t+1} = x_i^t + \alpha \oplus Lévy\ (\lambda) \tag{5}$$

The Levy's distribution (λ) can be calculated by Eqn. (6)

$$Levy \sim u = t^{-\lambda} \tag{6}$$

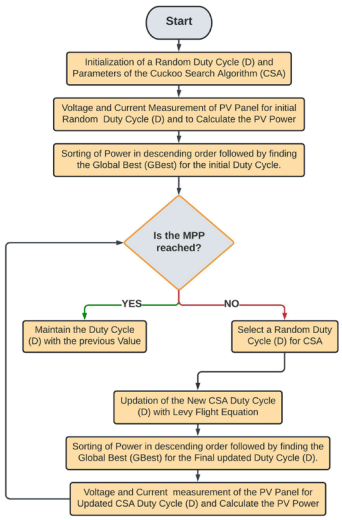



Fig. 6. Flowchart for Cuckoo Search Algorithm (CSA) based MPPT

### *B. Flower Pollination Algorithm (FPA)*

FPA is a Nature-Inspired Meta-Heuristic Algorithm which was originally devised in 2012 by Xin-She Yang. [2] This technique is inspired by the Pollination (Abiotic & Biotic) experience of the blossoming plants. Only roughly 10 % of the blooms perform Self/Abiotic pollination, and the remaining undergo Cross/Biotic Pollination. For Biotic-pollination, numerous pollinators like bees, moths, beetles, bats, insects, etc. are required. The pollinators help the plant in the transfer of the pollens from the anther to the stigma of the flower. Now, to reach the optimal point, a random flower population N is initialized, subsequently, each flower's Fitness is evaluated in the search space, using the fitness function to figure out the best pollen, which is considered the solution. Then, a switch probability is employed to decide between the biotic and abiotic pollination. In the first rule: for global optimum pollination, biotic pollination comes into play, in which the Lévy's Flight (8) is used to carry the pollen by pollinators. The mathematical model for global search in FPA is represented by Eqn. (7). [6]

$$x_i^{t+1} = x_i^t + L(g^* x_i^t) \tag{7}$$

$$\frac{\lambda\Gamma(\lambda)\ sin\ (\pi\lambda/2)}{\pi}\ \frac{1}{s^{1+\lambda}} \tag{8}$$

Similarly, In the second rule: for Local optimal Pollination, the self-pollination comes into play, we can perform the Local pollination in FPA using Eqn. (9) which mimics the flowering constancy in the limited surroundings.

$$x_i^{t+1} = x_i^t + \varepsilon(x_j^t - x_k^t) \quad (9)$$

Where, $x_i^{t+1}$ = New Solution, $x_i^t$ = Old Solution, $g^*$ = Best Solution, $L$ = Levy's Distribution, $x_j^t$ & $x_k^t$ = Pollens.

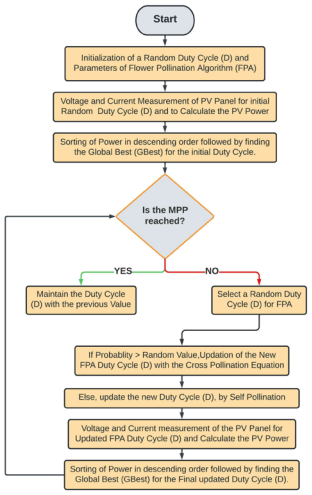



Fig. 7. Flowchart for Flower Pollination Algorithm (FPA) based MPPT

## C. *Grey Wolf Optimization Algorithm (GWOA)*

The GWOA is a high-level problem-independent swarm intelligence technique that is inspired by the social hierarchy and hunting techniques used by grey wolves. The GWOA methodology was originally fabricated by Mirjaliali Mohammad and Lewis in the year 2014. [3]

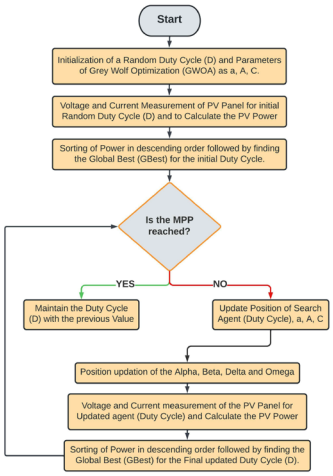



Fig. 8. Flowchart for Grey Wolf Optimization (GWOA) based MPPT

Grey wolves are considered the most successful animals on earth because of their excellent bonding, unity, and loyalty among pack members. Which makes them the survivors of the hardest scenarios. There are generally four ranks in a pack of grey wolves, i.e., Alpha (α), Beta (β), Delta (δ), and Omega (ω). The GWOA mimics and incorporates the leadership and hunting mechanisms of the grey wolves. The approaches used by grey wolves in hunting are (i) Searching, (ii) Tracking & Chasing, (iii) Pursuing & Encircling, and finally (iv) Attacking the Prey. The GWOA's mathematical modeling is validated by Eqn. (10) and (11), which represent grey wolf encircling behavior and that is used to update the grey wolf positions following the prey positions in GWOA.

$$\vec{A} = | \vec{C}.\overrightarrow{X_p} - \vec{X}(t) | \quad (10)$$

$$\vec{X}(t+1) = | \overrightarrow{X_p}(t) - \vec{A}.\vec{D} | \quad (11)$$

Where, $t$ = Current Iteration, $\overrightarrow{X_p}$ = Prey Position, $\vec{X}$ = Grey wolf Position, $\vec{A}, \vec{C}$ = Coefficient vectors.

## D. *Particle Swarm Optimization Algorithm (PSOA)*

The PSO Algorithm (PSOA) is a Meta-Heuristic methodology under the category of swarm intelligence (SI). Furthermore, the PSOA technique is a population-based probabilistic time-dependent search algorithm, that is motivated by the social demeanor of a moving flock of birds in a crowd. The PSOA computational strategy was first established by Kennedy and Eberhart in the year 1995. [4] The PSOA methodology works on the population of particles, where every particle is itself a point of a solution. The implementation of this algorithm requires only a few parameters, making it straightforward and simple to solve real-world problems. Starting with a randomly initialized population of candidate particles moving in a randomly chosen direction, each particle goes through the searching space and remembers its best previous position (PBest) and of its swarm neighbors (i.e., Global Best). Each particle searches for the optimum point by continuously updating its generations. In each iteration, every particle finds two best values: (i) PBest and (ii) GBest. The swarm particles update their positional ($x_i$) and velocity ($v_i$) values by Eqn. (12) and (13) respectively. PSOA is a simple yet strong algorithm.

New Position of the Particle = Old Position + Velocity

$$x_i^{t+1} = x_i^t + v_i^t * t \quad (12–13)$$

$$v_i^{k+1} = wv_i^k + c_1 r_1 (PBest_i^t - x_i^t) + c_2 r_2 (GBest_i^t - x_i^t)$$

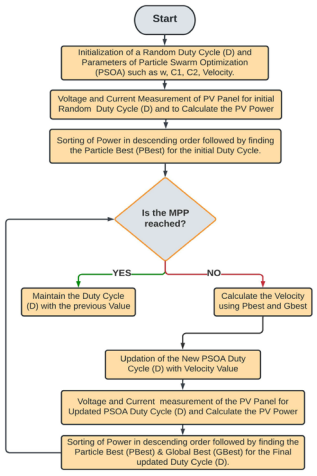



Fig. 9. Flowchart for Particle Swarm Optimization (PSOA) based MPPT

Where, $PBest$ = Particle Best, $GBest$ = Global Best, $w$ = Inertia Weight, $c_1, c_2$ = Constants, $r_1, r_2$ = random value.

## IV. MATLAB SIMULATIONS: RESULT & ANALYSIS

In this simulation, we have considered three different Weather/Climatic conditions for the Vikram Solar PV Panel/Module to operate under, in order to assess the output efficiency and record the behavior of the four Meta-Heuristic MPPT algorithms loaded to track the MPP power out of the PV System. The three weather conditions simulated here are: (A) Standard Weather Conditions (SWC) having Fixed Temperature = 25º C and Irradiation = 1000 W/m$^2$ further subdivided into – (i) Without MPPT Control and Case (ii) With MPPT Control, (B) DWC: Fixed Standard Temperature = 25º C & Variable Irradiation, and (C) DWC: Fixed Irradiation = 1000 W/m$^2$ & Variable Temperature values. The Matlab/Simulink Model for PV Simulation under Standard and Dynamic Weather Conditions is presented in Fig 10. [7]

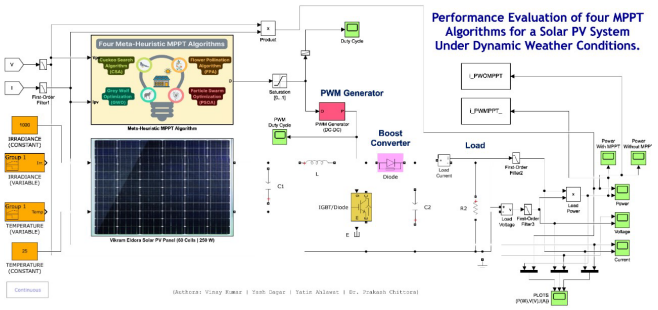



Fig. 10. Matlab/Simulink Model for PV Simulations under SWC & DWC

### A. PV System under Standard Weather/Climatic Conditions

The Photovoltaic (PV) System response/behavior under Standard Weather Conditions (SWC) having fixed temperature and fixed illumination (Temp = 25º C & Irradiation = 1000 W/m$^2$) (i) Without and (ii) With the implementation of MPPT Controls (CSA, FPA, GWOA, PSOA) is depicted in Fig. 11.

It is clearly visible that even under steady-state Standard Weather Conditions (SWC) all four MPPT Algorithms perform differently & better than the PV system operating without MPPT. In order to understand the significance of the MPPT algorithms, MPP Tracking efficiency may be calculated using Eqn. 14 for both (i) & (ii).

$$\eta_{mppt} = \frac{\int_0^t P_{mppt}(t)\,.\,dt}{\int_0^t P_{max}(t)\,.\,dt} \text{ x } 100 \qquad (14)$$

The comparative analogy under STC is not practical but provides us with good insights to have a good understanding of the mechanism and behavior of the MPPT methodologies. In the simulation process, we have considered the simulation time equal to 2 seconds. Resultant MPP Tracking efficiencies ($\eta_{mppt}$ [%]) calculated by Eqn. 14 for the above (i) without & (ii) with MPPT cases is as follows: Without MPPT (222.67 W, 89.06 %), CSA (228.39 W, 91.35 %), FPA (241.47 W, 96.58 %), GWOA (243.01 W, 97.2 %), PSOA (243.01 W, 97.2 %).

With the use of MPPT, the PV output power has increased by 2.5 %, 8.4 %, 9.1 and 9.1 % respectively. According to the MPP tracking efficiencies, it implies that the MPPT performance is in the order: PSOA > GWOA > FPA > CSA, clearly PSOA outperforms all the other MPPT techniques. This analysis shows the cruciality of the MPPT-based PV systems over PV Systems without MPPT.

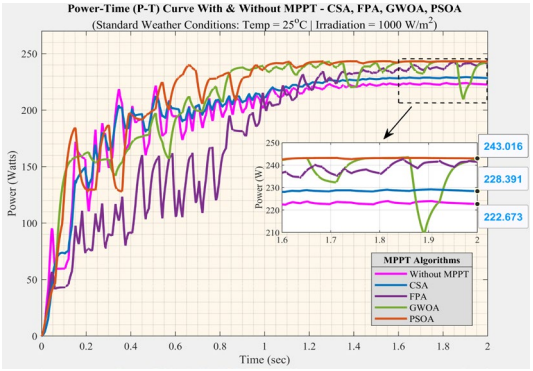
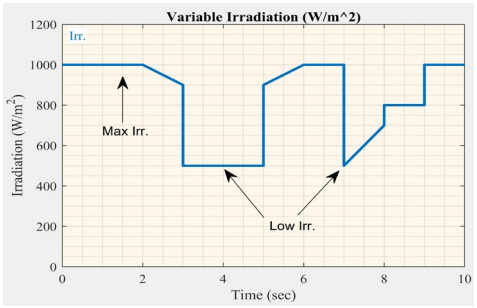


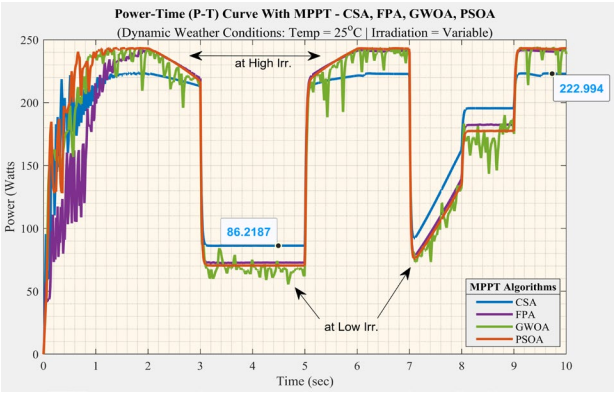

Fig. 11. Optimal MPP Power under Standard Weather Conditions (SWC)

### B. PV System under Dynamic Weather Conditions (DWC): Fixed Temperature (ºC) & Variable Irradiation (W/m$^2$)

In this first case of DWC), we have considered the PV System to be more close to the real/practical scenario, where the Temperature (ºC) is maintained at 25 ºC and solar illumination (W/m$^2$) to be variable as a linear function of time as depicted in Fig. 12 (i). The Output MPP tracking response of the PV system with MPPT is depicted in Fig. 12 (ii).




(i)




(ii)

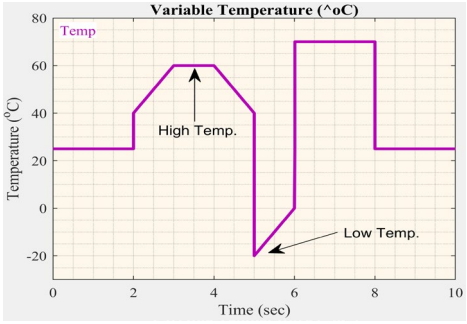
Fig. 12. (i) Input and (ii) Output under Fixed Temp. & Variable Irradiance

In the foregoing analysis, we have perceived from Fig. 12 (ii) that the P-T curves displayed for four MPPT approaches under the DWC are proportionate to the corresponding dynamic input illumination on the PV panel. Therefore the PV system extracts high MPP power during high irradiation and low MPP power under low irradiation. Moreover, under (a) Lower illumination condition at time 3-5 seconds the Cuckoo Search Algorithm (CSA) outperforms all the other MPPTs but under (b) High Illumination condition at 5-7 seconds the CSA underperforms all the other four algorithms. additionally, the FPA and PSOA techniques show relatively identical performance in extracting the MPP.

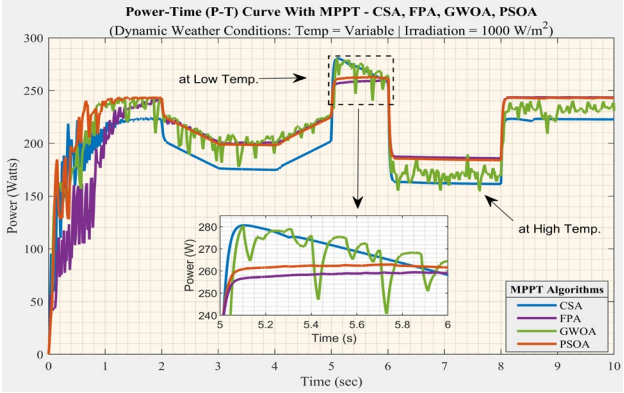
### C. *PV System under Dynamic Weather Conditions (DWC): Fixed Irradiation (W/m$^2$) & Variable Temperature (ºC)*

In the second case of DWC, we have considered the PV System to be under a real-time scenario where the solar Irradiance illumination (W/m$^2$) is fixed at 1000 W/m$^2$, and the Temperature (ºC) is varied abruptly as a linear function of time as displayed in Fig. 14 (i). because Temperature is an environmental factor/parameter which varies very slowly in real-time practical settings. The corresponding Dynamic output response of the PV system is depicted in Fig. 13 (ii).

For the PV system to be under the DWC with variable Temperature and fixed Irradiation of 1000 W/m$^2$, we observe that the P-T curves in Fig. 13 (ii) represent an inversely proportional relationship between the temperature variation and the Output Power, as in Fig. 13 (i) at time 2-5 seconds the Temperature gradually increases and then slowly decreases followed by a sudden fall at the 5th second, corresponding to which, in Fig. 13 (ii) a similar but inverse behavior is seen for the Output MPP Power. Moreover, at the Low temperature of -20 ºC, the CSA has outperformed with 280 W while at a High Temperature of 70ºC the CSA has underperformed in contrast to the FPA, GWOA, and PSOA.




(i)




(ii)

Fig. 13. (i) Input and (ii) Output under Fixed Irradiance. & Variable Temp.

## V. Conclusion

In this paper, we have performed a comparative study and analysis on the individual response of the Four Meta-Heuristic MPPT Algorithms, namely CSA, FPA, GWOA, and PSOA for a DC/DC Step-Up Converter based Vikram Solar Eldora (250 W) PV System. In the first part (A) of the simulation, it was observed that the implementation of the MPPT algorithm made the PV system attain up to 9 % of an additional power than what was earlier being achieved under no MPPT scenario. Therefore an MPPT is very crucial for a PV system to perform at its best potential under full capacity. Further, the simulation results achieved in the subsequent parts (B) and (C) under the DWCs were observed and the following insights were drawn: The CSA was found to be efficiently tracking MPP under lower temperatures than in higher temperatures, that too better than the PSOA in the latter low-temperature condition. The GWOA achieved the MPP earlier than even the most efficient MPPT i.e. PSOA, but was unable to acquire the stability in the transient response, rather constantly kept on fluctuating around the optimal point. In terms of the MPP convergence speed of algorithms, the PSOA and FPA were found to be faster than CSA and GWOA. Considering the stability of the transient response - out of the four MPPT algorithms, the PSOA, FPA, and CSA are observed to be stable with minimal fluctuations, while on the other hand GWOA undergoes numerous fluctuations and is unable to attain the final stability. Considering the complexity of their implementation, they have an order based on the number of steps/parameters involved as follows: FPA > GWOA > PSOA > CSA. This comparative analogy of these four Meta-Heuristic MPPT algorithms is meant to aid & help the current researchers in the field of Solar PV energy generation in studying and producing efficient, clean, and long-lasting sustainable solar energy for the future of humankind and life on the earth. [8]

## Acknowledgment

The authors of the paper would like to acknowledge and extend their humble & heartfelt thanks with gratitude towards their alma-mater, Delhi Technological University (Formerly DCE), Delhi, and Prof. (Dr.) Prakash Chittora, EED, DTU for their esteemed guidance and constant support in the Project.

## References

[1] X. Yang and Suash Deb, "Cuckoo Search via Levy flights," in *2009 World Congress on Nature & Biologically Inspired Computing (NaBIC)*, 2009, pp. 210–214. doi: 10.1109/NABIC.2009.5393690.

[2] X.-S. Yang, "Flower Pollination Algorithm for Global Optimization," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7445 LNCS, 2012, pp. 240–249. doi: 10.1007/978-3-642-32894-7_27.

[3] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, Mar. 2014, doi: 10.1016/j.advengsoft.2013.12.007.

[4] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, 1995, pp. 39–43. doi: 10.1109/MHS.1995.494215.

[5] M. S. Wasim, M. Amjad, S. Habib, M. A. Abbasi, A. R. Bhatti, and S. M. Muyeen, "A critical review and performance comparisons of swarm-based optimization algorithms in maximum power point tracking of photovoltaic systems under partial shading conditions," *Energy Reports*, vol. 8, pp. 4871–4898, Nov. 2022, doi: 10.1016/j.egyr.2022.03.175.

[6] J. Prasanth Ram and N. Rajasekar, "A Novel Flower Pollination Based Global Maximum Power Point Method for Solar Maximum Power Point Tracking," *IEEE Trans. Power Electron.*, vol. 32, no. 11, pp. 8486–8499, Nov. 2017, doi: 10.1109/TPEL.2016.2645449.

[7] M. Naseem *et al.*, "Assessment of Meta-Heuristic and Classical Methods for GMPPT of PV System," *Trans. Electr. Electron. Mater.*, vol. 22, no. 3, pp. 217–234, Jun. 2021, doi: 10.1007/s42341-021-00306-3.

[8] W. Anjum, A. R. Husain, J. Abdul Aziz, S. M. Fasih ur Rehman, M. P. Bakht, and H. Alqaraghuli, "A Robust Dynamic Control Strategy for Standalone PV System under Variable Load and Environmental Conditions," *Sustainability*, vol. 14, no. 8, p. 4601, Apr. 2022, doi: 10.3390/su14084601.

View publication stats

*2022 2nd International Conference on Intelligent Technologies (CONIT)*
*Karnataka, India. June 24-26, 2022*

# Prediction of COVID 19 Using Chest X-Ray Images through CNN optimised using Genetic Algorithm

Dr. Aruna Bhat
*Computer Science and Engineering*
*Delhi Technological University*
New Delhi, India
aruna.bhat@dtu.ac.in

Utkarsh Bhardwaj
*Computer Science and Engineering*
*Delhi Technological University*
New Delhi, India
utkarshbhardwaj_2k18co379@dtu.ac.in

Muskaan Singla
*Computer Science and Engineering*
*Delhi Technological University*
New Delhi, India
muskaansingla_2k18co216@dtu.ac.in

Krati Garg
*Computer Science and Engineering*
*Delhi Technological University*
New Delhi, India
kratigarg_2k18co177@dtu.ac.in

2022 International Conference on Intelligent Technologies (CONIT) | 978-1-6654-8407-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CONIT55038.2022.9848316

***Abstract* - It's been over two years since the novel Coronavirus first appeared and with the constantly evolving new variants found around the world, the havoc of n-Coronavirus seems to be unstoppable. With more than 264 million people being affected by the n-coronavirus as of 2nd December 2021 and 5.2 million deaths around the world, there is a dire need to increase the COVID-19 testing to stop the virus from spreading further. With COVID - 19 devastating the economic situation of various countries across the globe, it has become necessary to come up with a fast, efficient, and inexpensive way to test the presence of the n-Coronavirus in people. However, the methods currently being used to test COVID 19 are rather very expensive and unavailable to a large section of society. One of the most feasible solutions to this problem is through radiological detection i.e., with Chest X - ray images. Contrary to the prevalent testing methods, Chest X - ray scans are much lesser in cost and are readily available. One major problem that arises is that COVID and pneumonia have very similar X-RAY results, so having a binary classification (COVID and NOT COVID) isn't enough. In this paper, we have put forward a model based on Convolutional NN for detection of Pneumonia, COVID - 19, and Normal patients using X - ray photos of Chest. We achieved an AUC score of 90% in our results while classifying the X-Ray Images. Besides Accuracy, we have also made the ROC Curve, confusion matrix, and classification report for our model. To keep our model lightweight, we have used a Genetic Algorithm to get the best hyperparameters possible for the model.**

***Keywords—COVID 19, Coronavirus, Convolutional Neural Networks, ROC, SARS CoV 2, Callbacks, PCR, Genetic Algorithm***

## I. INTRODUCTION

The world stopped with the advent of the novel coronavirus. The lack and inadequacy of technology in healthcare was clearly evident and the fact that millions of people died due to the outburst of the pandemic came as a nightmare. In countries like India, where there is an economic gap and people die everyday due to poor financial conditions, it becomes necessary to develop mechanisms by which they can get an early detection of various diseases so that they can take the required action on time. [1] Such a system is possible if we leverage the results generated by the neural networks. Image analysis can be done using deep neural networks to classify the scans into different types of diseases based on their target body part.

In the year 2020, the COVID-19 epidemic was the most serious human disaster. In December 2019, early instances of COVID-19 were discovered in Wuhan District, China.[2, 3] The tests that are currently being conducted to test COVID 19 are the Polymerized Chain Reaction (PCR) Testing and serologic testing.

The PCR tests are relatively accurate and can detect the presence of the novel coronavirus with precise accuracy. However, it is important to take note that PCR tests are very time-consuming and complex. Considering these factors, an approach which can be used for detection of COVID - 19 in people can be radiological detection, in which X - ray images of Chest can be used to predict whether a person has SARS-Cov-2 or not. Hence, if we can come up with a method which uses X - ray images for detecting COVID - 19, it will not only be accurate but also inexpensive and readily available.

## II. RELATED WORKS

Ravneet Punia et al. [1] used various ResNet architectures for the three-way classification stated above. Instead of randomly selecting the learning rate for their models, they used a technique inspired by ADADELTA. While training some 128 images they plotted the linear value of the loss function and the learning rate and selected the learning rate with the minimum corresponding value of the loss function. They achieved the best accuracy with ResNet-50 (73%).

Rachna Sethi, Monika Mehrotra and Dhaarna Sethi in their paper [5] used and compared various model architectures leveraging transfer learning to do a binary classification on COVID - 19 and healthy classes. According to doctors, keeping track of the oxygen levels in the body could indicate the presence or absence of COVID - 19.

Mohamed Abdel-Basset et al. [6] used an amalgam of Marine Predators Algorithm (MPA) and Improved Marine Predators Algorithm (IMPA) to solve the challenges connected with picture segmentation using several thresholds. The multi - thresholding problems were overcome by adapting the IMPA using RDR. They were able to improve the binary classification models using this nature - inspired model.

S. Rajaraman et al. [7] in their paper performed the classification of 3 classes – COVID, bacterial pneumonia, and Normal. They used 8 pre-trained ImageNet models along with a custom convolutional neural network which extracted the characteristic features for the 3 different classes. To minimise



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:57:57 UTC from IEEE Xplore. Restrictions apply.

complexity and enhance the memory efficiency of the models, they pruned the best-performing models repeatedly. To further increase classification performance, they pooled the best-performing reduced models' predictions using various ensemble strategies.

The authors [8] have used a Genetic Deep Learning CNN to classify the radiology samples. They developed a genetic algorithm based on a gene map using the DCNN model. The model is trained for about 100 epochs, leading to the GD CNN having a high computational complexity as well as a high spatial complexity. They were able to successfully demonstrate the application of neuroevolution to medical images.

Michael J. Horry [9] has tried to detect the covid disease using three types of medical photos namely the ultrasound, X-ray, and CT scan. They have performed a comparative study to pick up the best CNN model which they have further optimised to overcome the challenges offered by the input dataset. With the in-depth analysis and study of the problem statement at hand, they have concluded that the ultrasound images provide the best results.

In [10], the authors talk about the severe impacts of covid and how the chest X-Rays can be confused with that of Pneumonia. They have tried to build a classifier using 278 images. Their classifier classifies the X-Rays into 3 categories which are-covid, pneumonia and normal. The authors have modified the images using various data augmentation techniques to avoid overfitting of the data, like the random rotation of the image, horizontal flip, random zooming of the image, etc. They have used the Resnet50 architecture and have trained only the last six layers of the ResNet model, keeping the remaining layers as it is.

In [11], the authors have performed a comparison study of deep learning models based on the results that they deliver for the data. They have also used ImageNet as their pretrained model to cover for the inadequacy of the dataset. They have experimented with the number of frozen layers in the pre-trained CNNs for the process of transfer learning. The authors performed a comparison between various pretrained models and concluded that with the mentioned architecture, MobileNet and VGG16 models performed the best for multiclass classification whereas VGG16 gave the best results for binary classification for COVID19.

In [12], Ahsan MM et al. have used a mixed dataset for their predictions. The dataset consists of 2 types of images :- 2191 X - ray photos, also 400 CT scans of the chest. Both of these contain COVID - 19 and Non - COVID - 19 photos. Here people experiencing symptoms of pneumonia are referred as Non-COVID-19 patients. Six different pre-trained ConvNet models have been used by them. They have used LIME for analysis of the overall prediction. The best result is achieved by MobileNet V2.

## III. PROPOSED METHODOLOGY

We train and deploy various models for the detection of the virus and also try to distinguish between COVID 19, Pneumonia and disease free people based on the image scans that are given as the input to the final model. The objective is to implement and optimise a model which can efficiently find the presence of coronavirus inside a human being The goal is to also enhance the algorithm so that it is capable of differentiating between coronavirus and Pneumonia because both of the mentioned diseases have similar chest X-ray images.

The motivation behind this paper was that not all places have sufficient medical facilities available and with the increasing number of COVID 19 cases, healthcare facilities are diminishing and are inaccessible to people in the rural areas. Hence, a solution based on artificial Intelligence can cater the needs of people easily because of its accessibility and accuracy. A great doctor might outperform an AI algorithm, but an AI algorithm performs more consistently and better than an average doctor. We identified 2 problems with the existing models –

a) Some of them didn't consider that COVID and pneumonia have very similar radiology results.

b) Some of them leveraged transfer learning which made the model quite big in size and not easy to deploy without using paid services.

The model proposed by us classifies the X - ray images of Chest into three categories i.e., Pneumonia Positive, COVID Positive, and Normal(without any disease). To deal with the second issue identified above, we use Genetic Algorithm to improve our model accuracy instead of leveraging Transfer learning and pretrained weights. The aim is also to motivate more research in this field as AI is the future of medical analysis.

The model can be deployed in an application which can be used in hospitals for identifying whether the patient has COVID 19 or Pneumonia or is not suffering from either of the 2 diseases. Since our model uses a simple CNN and not commonly used transfer learning models, the model in itself is light and can be deployed for use in the real world.

***Convolutional Neural Network (CNN) -*** ConvNet or CNN's are deep learning based approaches and they take images coded in the form of numerical data as their input. The images are then passed through multiple layers of the neural network to perform the required operation on them. In comparison to other classification techniques, a Convolutional Neural Network requires substantially less pre-processing.[5] The advantage of using this approach is that the ConvNets are capable of learning the required filters on their own once they are subjected to enough training layers.

The convolutional neural networks were initially designed to replicate the human brain and hence their learning process is also similar to how human beings learn. The human nervous system contains small neurons which are capable of responding to any actions that take place near their surface area. Multiple similar neurons can be brought together to create a bigger visual field. CNN uses a technique called Convolution.

***Genetic Algorithm -*** Genetic algorithms are search based algorithms which are used for optimization, selecting the best solutions from a pool of different solutions.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:57:57 UTC from IEEE Xplore. Restrictions apply.

Genetic algorithms use various biology-drawn functions like selection, mutation and crossover. It calculates the fitness value for each individual for evaluation. Only the fittest individuals are then allowed to mutate and reproduce for the production of better and fitter genes.

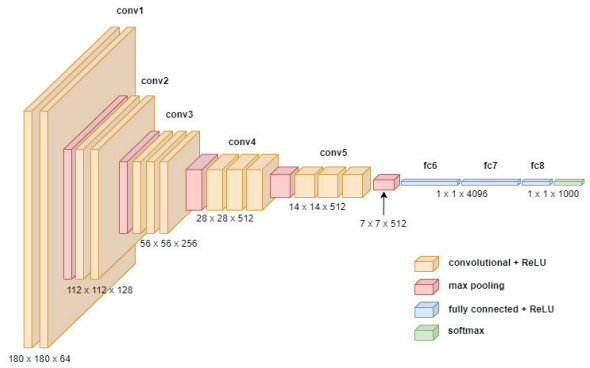



Fig. 1. CNN architecture of our model.

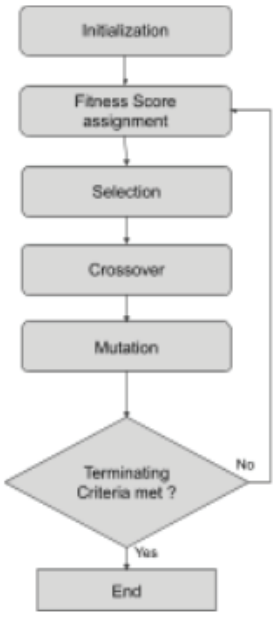



Fig. 2. Work flow of genetic algorithm.

## IV. THE DATASET

The COVID - 19 Radiography Dataset from Kaggle is the dataset we've chosen for our research. This is a huge dataset that includes X - ray pictures of the chest for virus positive patients, images of viral pneumonia and normal. [14] Out of the total images, we selected 1300 COVID-19 positive photos, 1300 images of healthy people, and 1300 images of pneumonia affected people as a subset of the entire dataset.

The test set includes 100 photos of COVID, viral pneumonia and normal with the remaining images serving as validation and training. The validation set is made up of 20 percent of the remaining data.

For the Genetic Algorithm analysis, a total of 300 images (100 for each class) have been used. Since Genetic Algorithm involves the making of several models in the form of different genes in different generations, the number of images was kept low assuming that they would be a good representation of the entire dataset.

## V. DATA PREPROCESSING

The dataset from Kaggle was in a form such that there were 3 directories each for COVID, viral pneumonia and normal. We wrote a script using the OS library to divide the data into a training and test set which is in an acceptable format for Image Data Generators.

The number of channels in the images were reduced from 3 to 1 using the OpenCV library. This was done based on the hypothesis that it would reduce the number of trainable parameters significantly without affecting the model results as the training images are radiology scans which are mostly (if not completely) in grayscale. The images were also normalised to ease computation and to provide good results without much deviation.

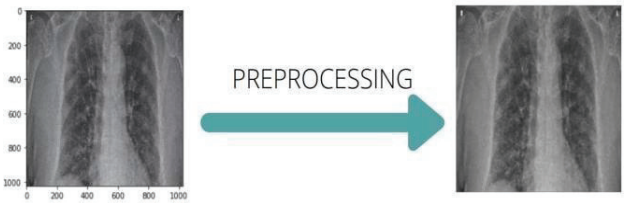



Fig. 3. Sample image after preprocessing.

We used 'Callbacks' for 'Early Stopping' while training the model. Callbacks enable us to stop the training when a particular threshold is met or is not met. The callback we used was 'Early Stopping' which monitored the 'Validation Accuracy' and whenever it didn't exceed by at least 0.1% in every 10 epochs, the training would terminate automatically, and the best weights would be restored. This not only saves us the trouble of choosing an appropriate number of epochs (in our case, we simply set it to 100) but also prevents Overfitting.

## VI. GENETIC ALGORITHM FOR BEST HYPERPARAMETERS

After preprocessing the images, we define a model architecture in a function with a gene as a hyperparameter. A gene is a permutation of a predefined pool of hyperparameters.

TABLE I. THE POOL OF HYPERPARAMETERS TAKEN

| | | | |
|---|---|---|---|
| Kernel Size | (2, 2) | (3, 3) | (5, 5) |
| Filters | 32 | 64 | 128 |
| Learning Rate | 0.001 | 0.005 | 0.01 |
| Activation | tanh | relu | sigmoid |
| Dropout | 0.1 | 0.2 | 0.3 |

TABLE II. TECHNIQUES USED IN GENETIC ALGORITHM

| | |
|---|---|
| Population Size | 12 |
| Evaluation Criteria | Accuracy of the trained neural network |
| Selection Criteria | Elitism |
| Crossover | Single-point crossover with Tournament Selection |
| Mutation | Random resetting |



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:57:57 UTC from IEEE Xplore. Restrictions apply.

## VII. RESULTS OF GENETIC ALGORITHM

The Genetic algorithm ran for 10 generations after which the 'Termination Condition' was triggered. The 'Termination Condition' for our Genetic Algorithm was that the algorithm should terminate once the highest Accuracy achieved by a gene in any generation was over 95%.

TABLE III. FINAL RESULTANT HYPERPARAMETERS

| | |
|---|---|
| Kernel Size | (2, 2) |
| Filters | 32 |
| Dropout | 0.2 |
| Activation | Relu |
| Learning Rate | 0.001 |

The accuracy achieved by this gene could be seen in Figure 4.

```
Generation 10: 0.9679487347602844

Early stopping
```

Fig. 4. Accuracy achieved by the best gene.

## VIII. TRAINING THE MODEL

***The Network Architecture*** – The architecture was defined using the hyperparameters obtained from the Genetic Algorithm. The input to the model is of dimension - (none, 180, 180, 1). All the convolutional layers have 64 filters each of size (3,3) with the activation function set as 'relu' (Rectified Linear Unit). The size of pool for the pooling layers is (2*2). Dropout of 0.2 (or 20%) is introduced to reduce the chance of overfitting. There are 2 dense layers at the end with 32 and 32 nodes each followed by the output layer consisting of 3 nodes and a softmax activation function. There are about 3 million trainable parameters. The training was done with the help of data generators. The ratio of the data in the training and validation set was 80:20.

We used callbacks to save us the trouble of hard coding the number of epochs or use. A callback is an object that may be used to accomplish various tasks at various stages throughout the training process. Callbacks are a way to tune the "no. of epochs" hyper parameter. We used a custom Early Stopping Callback and set the 'val_accuracy' up to be monitored with a patience of 10 epochs and a min_delta value of 1%.

We also set the restore_best_weights parameter to true. So, we were able to set the number of epochs to 100 and not worry about it. In simple words this means that the model would keep on training until every 10th epoch had a val_accuracy increase of at least 0.001. The moment this condition becomes false, the training will stop and the weights of the epoch with the best val_accuracy will be restored.

The model trained for 49 epochs before the Early Stopping Callback stopped the training. The validation accuracy was 84 percent and the training accuracy was 89 percent.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 178, 178, 64)      640

 conv2d_1 (Conv2D)           (None, 176, 176, 64)      36928

 max_pooling2d (MaxPooling2D  (None, 88, 88, 64)       0
 )

 dropout (Dropout)           (None, 88, 88, 64)        0

 conv2d_2 (Conv2D)           (None, 86, 86, 64)        36928

 conv2d_3 (Conv2D)           (None, 84, 84, 64)        36928

 max_pooling2d_1 (MaxPooling  (None, 42, 42, 64)       0
 2D)

 dropout_1 (Dropout)         (None, 42, 42, 64)        0

 flatten (Flatten)           (None, 112896)            0

 dense (Dense)               (None, 32)                3612704

 dense_1 (Dense)             (None, 32)                1056

 dropout_2 (Dropout)         (None, 32)                0

 dense_2 (Dense)             (None, 3)                 99

=================================================================
Total params: 3,725,283
Trainable params: 3,725,283
Non-trainable params: 0
_________________________________________________________________
```

Fig. 5. Architecture of CNN

## IX. RESULTS AND ANALYSIS

To analyse how our model is performing. we run it on the test set as well as the entire dataset as well. For this analysis, various measures like precision, accuracy, recall, f1- score as well as ROC curve have been used to check the performance of our model.

Formulas for some of these measures are given below:-

$$Accuracy = \frac{tn + tp}{tp + tn + fp + fn}$$

$$F1 - Score = 2 * \frac{precision * recall}{precision + recall}$$

***AUC-ROC*** - AUC identifies the extent or amount of distinction, whereas ROC is a probability curve. It shows how effectively the model differs between different classes. The ROC curve is drawn with TPR on the y-axis and FPR on the x-axis.

$$TPR = \frac{tp}{tp + tn}$$

$$FPR = \frac{fp}{tn + fp}$$

The model's training history was stored for visualisations. Figures 6 and 7 demonstrate the loss and accuracy curves for the training and validation data sets throughout the epochs.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:57:57 UTC from IEEE Xplore. Restrictions apply.

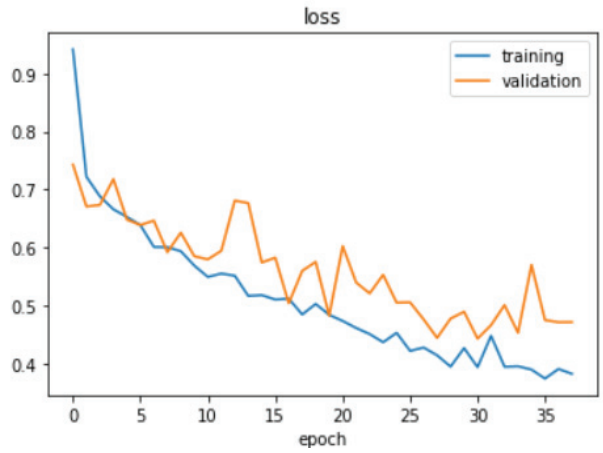



Fig. 6. Loss curve

We can see in the loss curve that both the training loss and the validation loss decrease upto a point and then begin increasing again (Figure 6). While in the accuracy curve that both the training accuracy and the validation accuracy increase upto a point and then begin increasing again (Figure 7). This could have led to overfitting but since we had a safeguard in the form of 'Early Stopping Callback', the training was paused and the weights from the best model were loaded.

Fig. 7. Accuracy curve

The accuracy of the model was 89 percent on the complete training set (training data + validation data).

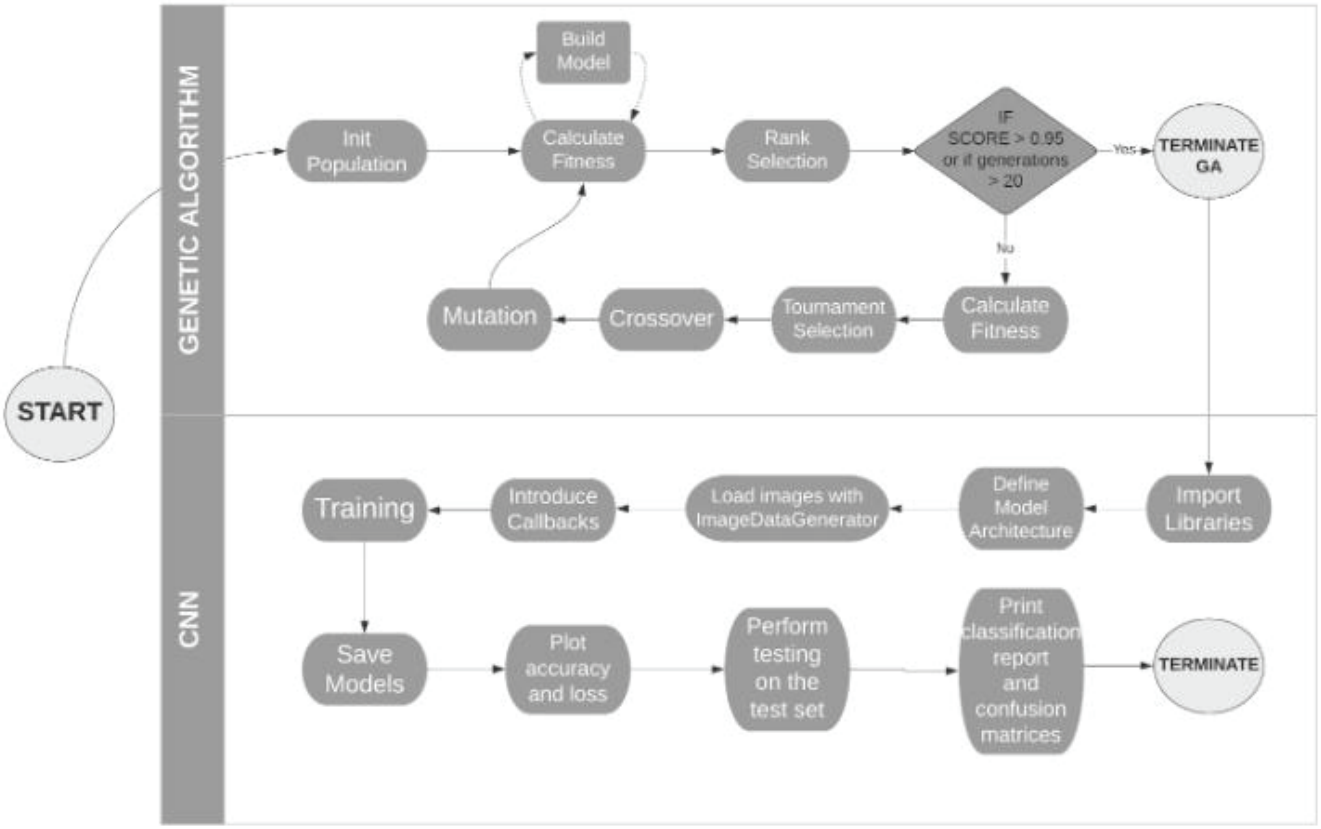



Fig. 8. Flowchart of algorithm



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:57:57 UTC from IEEE Xplore. Restrictions apply.

The label mapping used in all the visualisations is as shown in Figure 9.

```
{'COVID': 0, 'Normal': 1, 'Viral Pneumonia': 2}
```

Fig. 9. Label mapping of the 3 classes

We used Seaborn's heatmap to plot the visual representation of the confusion matrix, displayed in Figure 10.

An interesting statistic to note here is that the model had a precision of 90% in the case of COVID. The classification report and the AUC ROC are shown in Figure 11 and 12 respectively.

The f1-score is one of the most powerful indicators of the model's effectiveness. The notion of a good model is supported by high f1-scores. The AUC - ROC was another metric we used to assess the model's effectiveness. On the training set, the AUC was - 0.90. The ROC is depicted in Figure 12.

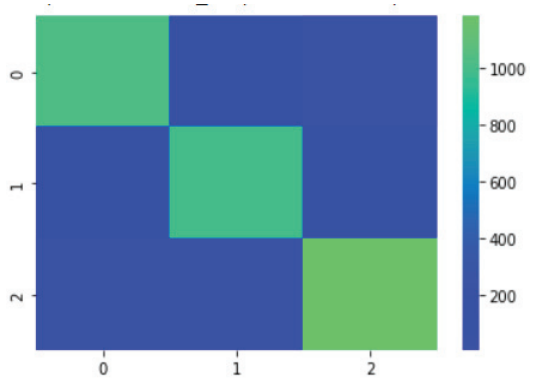



Fig. 10.Confusion matrix for testing on entire dataset

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.90 | 0.86 | 0.88 | 1200 |
| 1 | 0.86 | 0.83 | 0.84 | 1200 |
| 2 | 0.90 | 0.98 | 0.94 | 1200 |
| accuracy | | | 0.89 | 3600 |
| macro avg | 0.89 | 0.89 | 0.89 | 3600 |
| weighted avg | 0.89 | 0.89 | 0.89 | 3600 |

Fig. 11.Classification report for testing on entire dataset

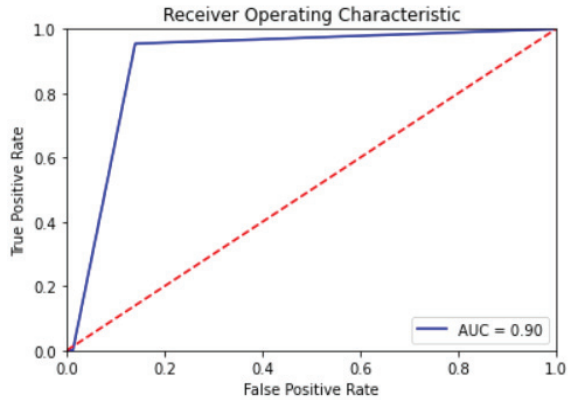



Fig. 12.ROC curve on entire set

We then ran the model on the test set (300 photos, 100 from each class) and got an accuracy of 89 % and an AUC of 0.912. The confusion matrix visualisation using seaborn, classification report and ROC for the same are shown in Figure 13, 14 and 15 respectively.

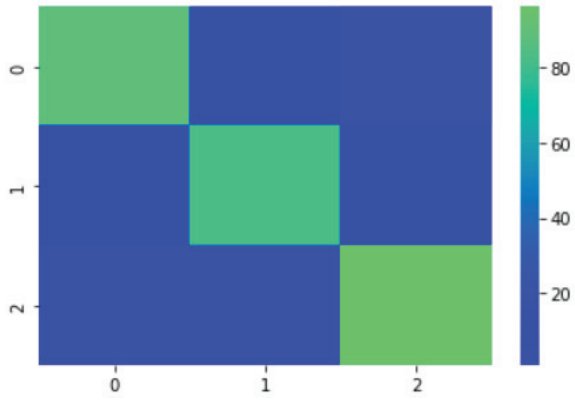



Fig. 13.Confusion matrix of test set

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.89 | 0.89 | 0.89 | 100 |
| 1 | 0.87 | 0.83 | 0.85 | 100 |
| 2 | 0.91 | 0.96 | 0.94 | 100 |
| accuracy | | | 0.89 | 300 |
| macro avg | 0.89 | 0.89 | 0.89 | 300 |
| weighted avg | 0.89 | 0.89 | 0.89 | 300 |

Fig. 14.Classification report of test set

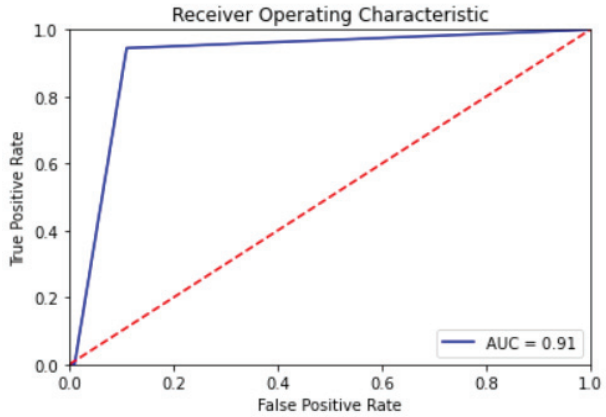



Fig. 15.ROC curve on test set

We see a tight kink in both the ROC curves and the ROC - AUC score too is high. This further establishes the potency of the model.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:57:57 UTC from IEEE Xplore. Restrictions apply.

TABLE IV. COMPARISON OF OUR RESULTS WITH PREVIOUS WORKS

| S.No | Method | Accuracy | Precision | Recall | F1-Score | Auc-ROC |
|---|---|---|---|---|---|---|
| 1 | Resnet50 [5] | 0.82 | 0.95 | 0.24 | 0.38 | – |
| 2 | Densenet 121 [8] | 0.91 | 0.87 | 0.87 | 0.87 | – |
| 3 | VGG 16 [8] | 0.88 | 0.82 | 0.92 | 0.87 | – |
| 4 | Ensemble learning with convolutional neural network [17] | 0.91 | – | 0.95 | 0.91 | 0.91 |
| 5 | Proposed approach | 0.92 | 0.90 | 0.86 | 0.88 | 0.90 |

## X. CONCLUSION

At a time when even rapid testing is failing to meet the needs of people, it is very important to come up with faster and structured solutions for detection of COVID 19. The model proposed by us distinguishes and detects people with COVID - 19 and Pneumonia.

Our model seeks to detect patients infected with COVID 19 more accurately and efficiently, with an 90% accuracy rate. Our model, which is based on CNNs, uses X - ray pictures of Chest to detect COVID - 19 cases which are readily available even in areas which do not have many medical facilities. On the complete training data, our model had a 0.90 AUC - ROC Curve score, and on the test data, it had a 0.91 AUC - ROC Curve score.

We hope that this promotes more research into this field. Our end goal is to completely intertwine medical image analysis and artificial intelligence. This could largely lighten the burden on the medical systems of the world which is mostly understaffed all the time and could expedite the treatment and recovery of patients across the world.

## XI. LIMITATIONS

We identified a few limitations to the study. The dataset was not as large as we'd like it to be and though we did apply image augmentation techniques, it is no substitute for the absence of raw data.

Experiments can be done with different levels of mutations. We could run another genetic algorithm over the current code to figure out the degree of mutation that would give the best results. This double genetic algorithm-based Convolutional Neural Network would be rather computation intensive and has thus been left as future work.

## REFERENCES


[1] Punia, R., Kumar, L., Mujahid, M., & Rohilla, R. (2020, June). Computer Vision and Radiology for COVID-19 Detection. In 2020 International Conference for Emerging Technology (INCET) (pp. 1-5). IEEE.

[2] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao, H., Guo, L., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin Q., Wang, J., and Cao, B. "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," The Lancet, Vol. 395(10223): pp. 497506, 2020.

[3] Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, X., Sun, Z., and Xia, L. "Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases," Radiology, Vol. 200642: pp. 1-23, https://doi.org/10.1148/radiol.2020200642, 2020.

[4] P. Mann, S. Jain, S. Mittal and A. Bhat, "Generation of COVID-19 Chest CT Scan Images using Generative Adversarial Networks," 2021 International Conference on Intelligent Technologies (CONIT), 2021, pp. 1-5, doi: 10.1109/CONIT51480.2021.9498272.

[5] Sethi, R., Mehrotra, M., & Sethi, D. (2020, July). Deep Learning based Diagnosis Recommendation for COVID-19 using Chest X-Rays Images. In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 1-4). IEEE.

[6] M. Abdel-Basset, R. Mohamed, M. Elhoseny, R. K. Chakrabortty and M. Ryan, "A Hybrid COVID-19 Detection Model Using an Improved Marine Predators Algorithm and a Ranking-Based Diversity Reduction Strategy," in IEEE Access, vol. 8, pp. 79521-79540, 2020, doi: 10.1109/ACCESS.2020.2990893.

[7] S. Rajaraman, J. Siegelman, P. O. Alderson, L. S. Folio, L. R. Folio and S. K. Antani, "Iteratively Pruned Deep Learning Ensembles for COVID-19 Detection in Chest X-Rays," in IEEE Access, vol. 8, pp. 115041-115050, 2020, doi: 10.1109/ACCESS.2020.3003810.

[8] R. G. Babukarthik, V. A. K. Adiga, G. Sambasivam, D. Chandramohan and J. Amudhavel, "Prediction of COVID-19 Using Genetic Deep Learning Convolutional Neural Network (GDCNN)," in IEEE Access, vol. 8, pp. 177647-177666, 2020, doi: 10.1109/ACCESS.2020.3025164.

[9] M. J. Horry et al., "COVID-19 Detection Through Transfer Learning Using Multimodal Imaging Data," in IEEE Access, vol. 8, pp. 149808-149824, 2020, doi: 10.1109/ACCESS.2020.3016780

[10] Bukhari, Syed usama & Khalid, Syed & Syed, Asmara & SHAH, SYED. (2020). The diagnostic evaluation of Convolutional Neural Network (CNN) for the assessment of chest X-ray of patients infected with COVID-19. 10.1101/2020.03.26.20044610.

[11] Taresh MM, Zhu N, Ali TAA, Hameed AS, Mutar ML. Transfer Learning to Detect COVID-19 Automatically from X-Ray Images Using Convolutional Neural Networks. Int J Biomed Imaging. 2021 May 15;2021:8828404. doi: 10.1155/2021/8828404. PMID: 34194484; PMCID: PMC8203406.

[12] Ahsan MM, Nazim R, Siddique Z, Huebner P. Detection of COVID-19 Patients from CT Scan and Chest X-ray Data Using Modified MobileNetV2 and LIME. Healthcare. 2021; 9(9):1099. https://doi.org/10.3390/healthcare9091099

[13] R. Oullette, M. Browne and K. Hirasawa, "Genetic algorithm optimization of a convolutional neural network for autonomous crack detection," Proceedings of the 2004 Congress on Evolutionary Computation (IEEE Cat. No.04TH8753), 2004, pp. 516-521 Vol.1, doi: 10.1109/CEC.2004.1330900.

[14] Sachin and A. Bhat, "A Survey on COVID-19 Detection from X-ray Images using Deep Learning," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9580164.

[15] Garg A., Aggarwal K., Saxena M., Bhat A. (2021) Classifying Medical Histology Images Using Computationally Efficient CNNs Through Distilling Knowledge. In: Tavares J.M.R.S., Chakrabarti S., Bhattacharya A., Ghatak S. (eds) Emerging Technologies in Data Mining and Information Security. Lecture Notes in Networks and Systems, vol 164. Springer, Singapore. https://doi.org/10.1007/978-981-15-9774-9_66

[16] Sakib, S., Tazrin, T., Fouda, M. M., Fadlullah, Z. M., & Guizani, M. (2020). DL-CRC: Deep Learning-Based Chest Radiograph




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:57:57 UTC from IEEE Xplore. Restrictions apply.

Classification for COVID-19 Detection: A Novel Approach. IEEE Access, 8, 171575-171589.

[17] Das, A.K., Ghosh, S., Thunder, S. et al. Automatic COVID-19 detection from X-ray images using ensemble learning with convolutional neural network. Pattern Anal Applic 24, 1111–1124 (2021). https://doi.org/10.1007/s10044-021-00970-4

[18] Goyal H., Attri D., Aggarwal G., Bhatt A. (2022) Predictive Model for COVID-19 Using Deep Learning. In: Ranganathan G., Bestak R., Palanisamy R., Rocha Á. (eds) Pervasive Computing and Social Networking. Lecture Notes in Networks and Systems, vol 317. Springer, Singapore. https://doi.org/10.1007/978-981-16-5640-8_43

[19] Dr. Aruna Bhat1, Garvit Arora2, Ashwin Joshi3, Gaurab Bhushan Singh, "Analysis of clinical texts for prediction of COVID-19 using Bag of Words and Artificial Neural Networks", Turkish Journal of Computer and Mathematics Education, Vol.12 No.12 (2021),1104-1113



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:57:57 UTC from IEEE Xplore. Restrictions apply.

# Prediction of Cryptocurrency prices using Transformers and Long Short term Neural Networks

1st Ayush Tanwar
*Department of Computer Science and Engineering*
*Delhi Technological University*
New Delhi, India
ayushtw@gmail.com

2nd Vinod Kumar
*Department of Computer Science and Engineering*
*Delhi Technological University*
New Delhi, India
vinod_k@dtu.ac.in

2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP) | 978-1-6654-7258-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICICCSP53532.2022.9862436

***Abstract*—Cryptocurrencies have taken over the whole financial world with their high risk and high reward quality. Cryptocurrency markets are one of the most complex and volatile markets in the world. In this paper, we make an attempt to predict financial time series for popular cryptocurrencies like Ethereum, Binance coin and Bitcoin, using deep learning and multifractal detrended fluctuation analysis techniques. After testing for non-linearity in the financial time series of popular cryptocurrencies, it was revealed that cryptocurrency time series manifest latent relations, short term, and long term memory. The method uses Transformers and Long-short term Neural networks(LSTM) to forecast the prices of various cryptocurrencies. Although using LSTM along with transformers leads to longer computational times, but the predictive accuracy is better as compared to traditional regression neural networks and k-NN forecasting models. Experimentation of the proposed model reveals that the deep learning model is profoundly efficient in predicting the intrinsic dynamics of cryptocurrency financial time series.**

***Index Terms*—Cryptocurrency, Deep Learning, Multifractal Detrended Fluctuation Analysis, Transformers**

## I. Introduction

Price prediction is a significant step for portfolio management and hedging. Numerous works have focused on this topic for the stock markets, but very few works have converged on this issue for Cryptocurrency markets. The ongoing regularization of these markets by governments, the volatility and the risks involved in cryptocurrencies are as significant as ever.

In this approach, first the prices data from binance.com [9], the largest cryptocurrency exchange in terms of trading volumes is obtained. Then two models - "training" and "test" are created using RapidMiner [8]. We then apply multifractal detrended fluctuation analysis(MFDFA) on proposed data in the MFDFA phase, as the data is weakly dependent on non-stationarity assumptions and noisy data.

We then use the processed data in the Transformer phase and LSTM phase parallelly to reduce computation times. LSTM neural networks system is used because it keep adjacent temporal information in a spontaneous manner and also control long-term information. Transformers are used because of their exceptional performance, as they combine self-attention mechanism, positional encoding, and parallelization. Then we combine the outputs from LSTM phase and Transformers phase in proposed hybrid combiner to obtain our final price prediction.

In section 2, we explain some previous cryptocurrency forecasting approaches. In section 3, we describe proposed approach in detail. In section 4, we analyze and conclude the proposed research.

978-1-6654-7258-6/22/$31.00 ©2022 IEEE

## II. Related work

The concept of cryptocurrencies is relatively new, and the first cryptocurrency was created in 2009 by Satoshi Nakamoto. The exponential increase in the prices of Bitcoin and other Cryptocurrencies in the 2010s led to the need for their price forecasting.

[1] used long-term memory to assess the market efficiency of cryptocurrencies. The volatility of cryptocurrencies was evaluated by [2], they assigned a few of the most popular cryptocurrencies the highest risk ratings. [4] was one of the first to use LSTM for cryptocurrency forecasting.

Sun et al. [3] used gradient boosting and decision trees for price forecasting of cryptomarkets. They used different kinds of key economic indicators to create a robust prediction model. Silva et al. [5] used variational mode decomposition (VMD) and stacking-ensemble learning to create a multi-step ahead bitcoin forecasting model. Cheng et al. [6] used detrended fluctuation analysis to investigate the price movement and MFDFA for robustness test. They found a strong momentum effect in Bitcoin and Ethereum markets. In the next section of this research, we explain proposed approach in detail.

## III. Proposed Approach

A basic data flow of proposed approach is explained in Fig 1. We first mine the training data from binance.com using rapidMiner,the training data consists of 180 days historical data of various cryptocurrencies. We then use this mined data in MFDFA phase.

### 3.1 Multifractal Detrended Fluctuation Analysis(MFDFA)

The MFDFA is based on DFA, Multifractal detrended fluctuation analysis (MFDFA) method can find higher-dimensional fractal and multifractal features lurked in time series.This method is used as an effective tool for measuring

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:54:54 UTC from IEEE Xplore. Restrictions apply.

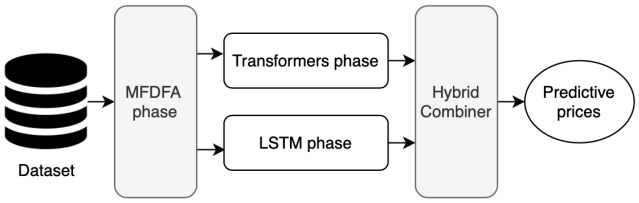



Fig. 1. Graphical Representation of Proposed Cryptocurrency forecasting approach

whether multifractal features exist in time-series data such as financial time series. We calculate variance in prices of cryptocurrencies at different dates using MFDFA algorithm [6], the variance for financial time series is calculated using the equation 1.

$$F^2(s,t) = \frac{1}{s}\sum_{i=1}^{s} z_t^2 \tag{1}$$

where s represents length of segment, $z_t^i$ represents the de-trended time series and can be calculated by the following:,

$$z_t(i) = v_t(i) - p_t(i), (1 < i < s) \tag{2}$$

where $v_t(i)$ and $p_t(i)$ represent the segment time series and trend time series at each segment t, respectively.

### 3.2 Long-short term memory deep learning neural networks

The LSTM neural networks present with a sturdy extension of the RNN (recurrent neural network) topology in regards of nonlinear modeling and, more importantly, prediction. In this respect, LSTM neural networks models not only retain adjacent temporal information in a spontaneous manner, but also control long-term-memory information. In simpler words, the LSTM can save prior information, which can significantly help improve its capability to inherent nonlinear patterns and acquire signal sequences.

LSTM uses the concept of controlling gate. Dependent on the inputs, the LSTM cell can retain or forget any cell state. The cell is maintained by three gates: input, output and forget. The input type gate defines how much current information should be regarded as input to generate the current state, while the forget type gate selects how much information can be stored from the last prior state. Lastly, the output type gate selects and filters the information that can be truly treated as essential and produces the output, which would be a prediction in our setting. Let's represent the input to every cells as $\nu_t$ and the previous time- step output as $g_{t-1}$. Then, the forget gate $f_t$ computes the input for the cell state $C_{t-1}$ using a sigmoid function, given by (3)

$$f_t = \sigma(W_f, [g_{t-1}, \nu_t] + b_f) \tag{3}$$

The input gate $i_t$ calculates the values to be updated to $C_t$ as follows:

$$i_t = \sigma(W_i, [g_{t-1}, \nu_t] + b_i) \tag{4}$$

Similarly, the output gate $o_t$ controls the output values:

$$o_t = \sigma(W_o, [g_{t-1}, \nu_t] + b_o) \tag{5}$$

Lastly, the output value of the LSTM cell is provided by:

$$g_t = o_t \odot C_t \, , where: \; C_t = f_i \odot C_{t-1} \oplus i_t \odot \bar{C}_{t-1} \tag{6}$$

with $C_t$ expressing the output of the nonlinear hyperbolic tangent (tanh) function. In this research, historical sequences are processed as the inputs of the LSTM to derive latent information, whereas the forecasting digital currency price is regarded as the targeted output.

### 3.3 Transformers phase

In the transformer phase, we use Multi-Transformers proposed by [7]. This sub-section presents the Multi-Transformer layers along with the volatility model based on them. We use Multi-Transformer layers introduced in our research to improve the accuracy and stability by using bagging in the attention mechanism. Rather than using the approach on all the data that are input into the model, the proposed approach uses bagging only in the attention mechanism of the layered architecture.Bagging requires large computational power which is one of the principal limitations of using this technique. The multi-Transformer structure is made of T Multi-Head attention units, one for each random sample of the input data. The mean of the various units is calculated in order to obtain the final attention matrix. Thus, the AMH (Average Multi-Head ) mechanism present in Multi-Transformer can be defined as follows:

$$AMH(Q,K,V) = \frac{\sum_{i=1}^{t} Concat(head_{1,t}), ...., (head_{h,t})W_t^O}{T} \tag{7}$$

where $W_i$ is the matrix of a parameter $head_{i,t}$ given by

$$head_{i,t} = Attention(Q_t W_Q^{i,t}, K_t W_K^{i,t}, V_t W_V^{i,t}) \tag{8}$$

and attention is calculated by

$$Attention(Q,K,V) = (\frac{QK^T}{\sqrt{d_k}})V \tag{9}$$

Where Q, V and K are input matrices and $d_k$ the number of input variables held into consideration in the dot-product attention mechanism.We take mean of AMH of all the parameters ($\overline{AMH}$) and send it as input to the hybrid combiner phase.

### 3.4 Hybrid Combiner phase

In this phase, we receive inputs from LSTM phase and Transformers phase. We combine the two inputs received by using the equation (10) and get the final prices prediction.

$$Finalprice = \beta.C_t + (1-\beta).\overline{AMH} \tag{10}$$

Where $\beta$ is the combining constant. $\beta$ is tuned to get the best accuracy while avoiding overfitting.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:54:54 UTC from IEEE Xplore. Restrictions apply.

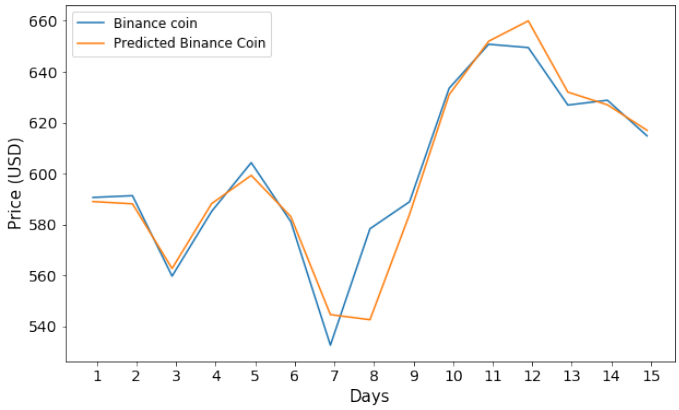



Fig. 2. 15 day actual prices and predicted prices of Binance coin

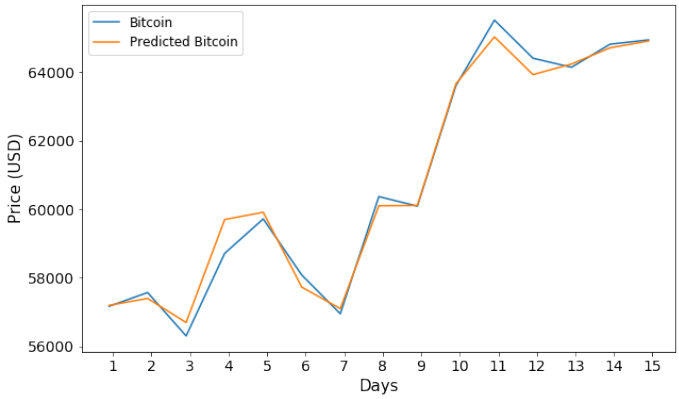



Fig. 3. 15 day actual prices and predicted prices of Bitcoin

## IV. Experimentation and Conclusion

In proposed model, we have experimented on three of the most popular cryptos: Bitcoin,Binance coin and Ethereum. We used rapidMiner [8] to mine the USD prices of these 3 cryptocurrencies from binance [9]. We use the following parameters: Open price, Low price, High price,Close price and the trading Volume. Proposed Approach first processes 180 days of data, then it forecasts the next 15 days of data using LSTM and Transformers networks.

Fig 2, 3, 4 shows 15 day actual prices and forecasted prices of the 3 cryptocurrencies. Looking at the graphs we can infer that proposed approach effectively predicts the dips and the peaks of cryptocurrency prices.

TABLE I
RMSE values of various cryptocurrency forecasting models

| Approach | RMSE | | |
|---|---|---|---|
| | Binance Coin | Bitcoin | Ethereum |
| LightGBA [3] | 6.01 $\times 10^2$ | 4.41 $\times 10^3$ | 5.45 $\times 10^3$ |
| ChoudharyKnn [10] | 6.24 $\times 10^2$ | 4.26 $\times 10^3$ | 4.31 $\times 10^3$ |
| LSTM-GRNN [4] | 4.13 $\times 10^2$ | 3.62 $\times 10^3$ | 3.14 $\times 10^3$ |
| **Proposed Approach** | 3.67$\times 10^2$ | 3.11$\times 10^3$ | 2.91 $\times 10^3$ |

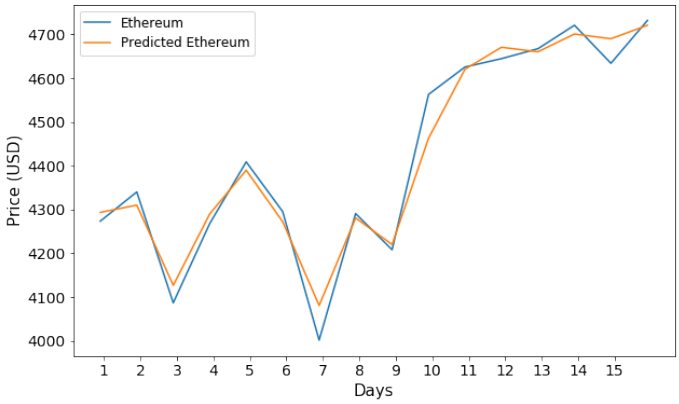



Fig. 4. 15 day actual prices and predicted prices of Ethereum

TABLE II
Average time taken (mins) per run for various cryptocurrency forecasting models

| **Approach** | **Average time (mins)** |
|---|---|
| LightGBA [3] | 193.48 |
| ChoudhKnn [10] | 720.67 |
| LSTMGRNN [4] | 410.21 |
| **Proposed model** | **580.33** |

We compare proposed model with three state of the art Cryptocurrency forecasting approaches. We perform 3 runs for each approach for each cryptocurrency and then we calculate the mean of the 3 runs for each cryptocurrency. Table I shows the various RMSE values. As seen from table I and II, LightGBA [3] which uses gradient boost, performed the fastest, but its accuracy is the lowest as it over-fits the data. It over-fits even after applying L1 and L2 regularization penalties and low learning rate. ChoudharyKnn [10] combined a variery of techniques with knn, it is more accurate than lightGBA but it took the longest time. LSTM-GRNN [4] was considerably faster than ChoudharyKnn as it used fewer machine learning techniques and it was also more accurate as LSTM networks keep adjacent temporal information in a spontaneous manner and also control long-term information which is ideal for our case. However, proposed approach is more accurate than LSTM-GRNN as we use LSTM along with Transformers, which combines self-attention mechanism, positional encoding, parallelization and long-short term memory to give us the best performance for time series prediction.

The proposed approach was more accurate than the other approaches, but it took longer processing time than other approaches. In the future we will work to optimize our approach and incorporate Natural Language Processing techniques in the time series prediction.

## References

[1] Gajardo, Gabriel, Werner D. Kristjanpoller, and Marcel Minutolo. "Does Bitcoin exhibit the same asymmetric multifractal cross-correlations with crude oil, gold and DJIA as the Euro, Great British Pound and Yen?." Chaos, Solitons & Fractals 109 (2018): 195-205.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:54:54 UTC from IEEE Xplore. Restrictions apply.

[2] Symitsi, Efthymia, and Konstantinos J. Chalvatzis. "Return, volatility and shock spillovers of Bitcoin with energy and technology companies." Economics Letters 170 (2018): 127-130.
[3] Sun, Xiaolei, Mingxi Liu, and Zeqian Sima. "A novel cryptocurrency price trend forecasting model based on LightGBM." Finance Research Letters 32 (2020): 101084.
[4] Wu, Chih-Hung, et al. "A new forecasting framework for bitcoin price with LSTM." 2018 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2018.
[5] da Silva, Ramon Gomes, et al. "Multi-step ahead bitcoin price forecasting based on vmd and ensemble learning methods." 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020.
[6] Cheng, Qing, Xinyuan Liu, and Xiaowu Zhu. "Cryptocurrency momentum effect: DFA and MF-DFA analysis." Physica A: Statistical Mechanics and its Applications 526 (2019): 120847.
[7] Ramos-Pérez, Eduardo, Pablo J. Alonso-González, and José Javier Núñez-Velázquez. "Multi-Transformer: A New Neural Network-Based Architecture for Forecasting S&P Volatility." Mathematics 9.15 (2021): 1794.
[8] Mierswa, I., & Klinkenberg, R. (2018). RapidMiner Studio (9.1) Data science, machine learning, predictive analytics. Retrieved from , https://rapidminer.com/
[9] Binance crpto exchange. https://www.binance.com/en
[10] Chowdhury, Reaz, et al. "An approach to predict and forecast the price of constituents and index of cryptocurrency using machine learning." Physica A: Statistical Mechanics and its Applications 551 (2020): 124569.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:54:54 UTC from IEEE Xplore. Restrictions apply.

# *Quad MIMO Antenna for UWB Applications*

R. Prashant,
*Department of Electronics and Communication Engineering, Delhi Technological University,*
Delhi, India, Email ID:
r.prashant_2k18ec124@dtu.ac.in

Raghav Goel,
*Department of Electronics and Communication Engineering Delhi Technological University,*
Delhi, India, Email ID:
raghavgoel_2k18ec126@dtu.ac.in

Raghav Sharma,
*Department of Electronics and Communication Engineering Delhi Technological University*,
Delhi, India, Email ID:
raghavsharma_2k18ec127@dtu.ac.in

Deepika Sipal,
*Department of Electronics and Communication Engineering, Delhi Technological University,*
Delhi, India, Email ID:
deepikasipal@dtu.ac.in

***Abstract*—A 4×4 corner etched MIMO antenna, fed with microstrip line, is presented in this paper for UWB applications. The size of the antenna is 33×30×1.6mm$^3$. The antennas are placed mutually perpendicular to each other to achieve better isolation. Four radiating patches are used for better signal quality and high data transfer rate. The operating bandwidth for the antenna lies between 3.1 GHz to 10.6 GHz. The mutual coupling value is less than -18dB over the entire operating range of frequency, which is well suited for UWB applications.**

***Keywords—Quad MIMO Antenna, UWB applications, four patch antennas, isolation, mutual decoupling***

## I. INTRODUCTION

In the recent past, with immense research and development in Wireless Communication System, Ultra-Wideband Communication with simplicity in implementation offers many advantages like higher data rate with better SNR, energy efficiency etc. According to global standards, the frequency range between 3.1GHz to 10.6GHz is widely considered as the UWB or the Ultra Wide Band Range [1]. With the popularity of the systems employing antennas having UWB frequency range, low power level and high noise immunity; the UWB antenna study has become an exciting field of research in recent times [1].

Many problems need to be solved for a feasible UWB antenna which includes impedance matching, compact antenna size, radiation stability, multipath fading etc [2]. To resolve these issues MIMO technology is combinedly used with the UWB system. MIMO uses multiple transmitters and receivers for transferring more data simultaneously. MIMO antenna improves the channel capacity of the communication system providing a good data rate and use multipath to its advantage [3].
Some hurdles associated with the design of UWB MIMO antenna include high isolation between the antenna radiators as well as reducing the size of the antenna [4].
As number of antennas are increased to improve the data rate of the system, the coupling between adjacent antennas also increases and to suppress this different type isolating circuits are used such as stub, slots, metamaterial, orthogonal polarization etc [5-10].

In this paper to overcome these challenges, design structure of antenna is chosen such that the antennas are placed perpendicular to each other; a corner etched structure is used as a radiator and for high isolation performance stub is used in ground plane side. The proposed antenna has very small size and shows good diversity performance in the entire UWB range. Section-II discuss about antenna design and section-III shows its performance analysis. The diversity parameters are presented in section-IV and finally section-V concludes the work.

## II. ANTENNA DESIGN AND SPECIFICATION

The proposed 4×4 MIMO antenna design is simulated in CST MWS and shown in the Fig.1.The proposed antenna with dimensions 33×30×1.6mm$^3$ has been designed on a FR-4 substrate with dielectric constant 4.3 with a loss tangent=0.025. The quad antenna consists of 4 radiator patch antennas, each of which are placed mutually perpendicular to the adjacent patch antenna on the same radiating plane. The design of each patch antenna on the radiator plane consists of bi-rhombic antenna patches; each bi-rhombus shaped patch is fed by microstrip feedline. The ground plane consists of a rectangular patch with triangular and rectangular slots etched onto it. A narrow rectangular stub has also been added on the other side of the antenna, that is, the ground plane which improvises the decoupling between the patches present in the antenna. The antennas are placed perpendicular to each other in order to help minimize the mutual coupling between the bi-rhombic patches. The bi- rhombus shaped patches are sized 7×7 mm$^2$ and 5.5×5.5mm$^2$ respectively and are placed one below the other on the radiating plane conjoined to each other, fed with the feedline. On the ground plane the rectangular plane is of the dimensions 10×4.3mm$^2$. Rectangular slot of size 3.1×0.4 mm$^2$ and triangular slot of base 2.78 mm and height 4.96 mm are

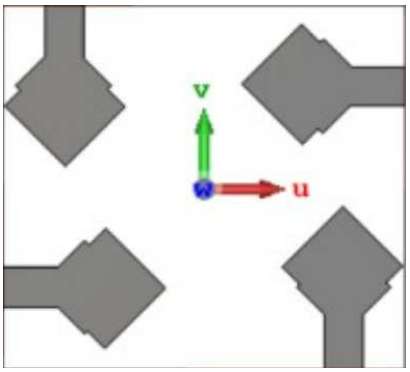
2022 IEEE Region 10 Symposium (TENSYMP) | 978-1-6654-6658-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/TENSYMP54529.2022.9864493

978-1-6654-6658-5/22/$31.00 ©2022 IEEE

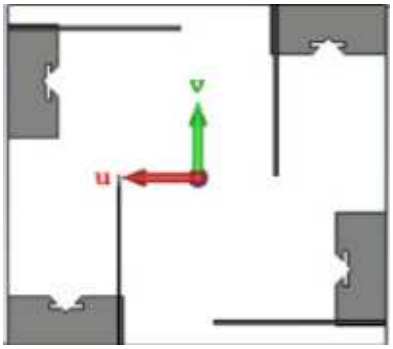


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:55:53 UTC from IEEE Xplore. Restrictions apply.

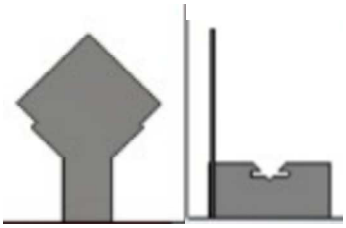
etched on the ground plane. In order to improve the decoupling effect rectangular stub of dimensions 0.25×15 $mm^2$ is added as well on the ground plane. The four patches present in the quad MIMO antenna are kept perpendicular with one another in order to have strong reception and transmission of signals.




(a)




(b)

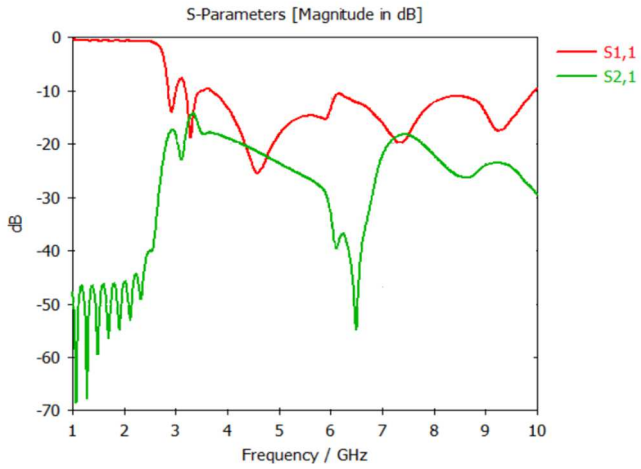


(c)

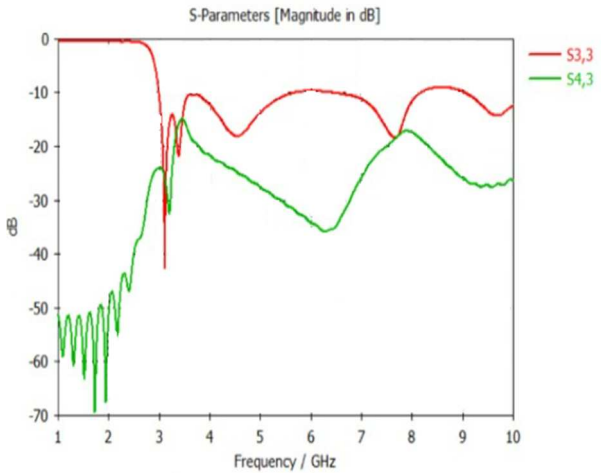
Fig:1 Antenna Design (a) Radiator Plane of Quad Antenna (b) Ground Plane of Quad Antenna (c) Single Patch Antenna Radiator and Ground Plane

## III. PERFORMANCE ANALYSIS

The performance of the quad Ultrawide Band antenna used for MIMO applications presented above is investigated by the use of CST Software. The scattering parameter or the S-parameter for the given quad patch antenna are shown in Fig. 2. It can be deduced that the proposed quad MIMO-Ultra Wide Band antenna performs extremely well with a S11 value less than -10dB for the entire range of operation. The isolation performance of MIMO-UWB antenna is denoted by S21 is shown in Fig.2 by green line on the graph. It can be analyzed that the isolation between the four patches present in the quad antenna is minimized simply with the stub and slot. Barring two ranges ie the 2.9GHz-4.3GHz range and the 7.3GHz -7.9GHz range the antenna is highly optimized and hence achieves a good mutual decoupling which lies below -20dB for the UltraWideBand range. The proposed MIMO-quad UWB antenna has a good isolation than most of the other antennas.




(a)

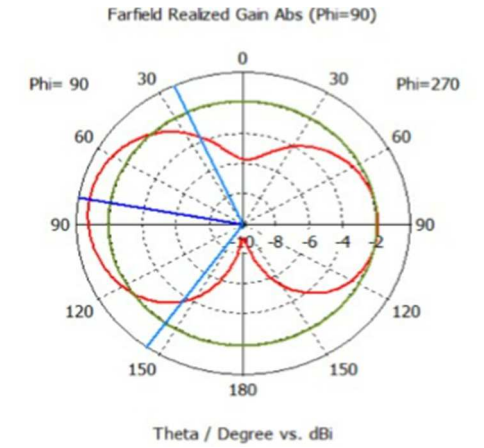



(b)

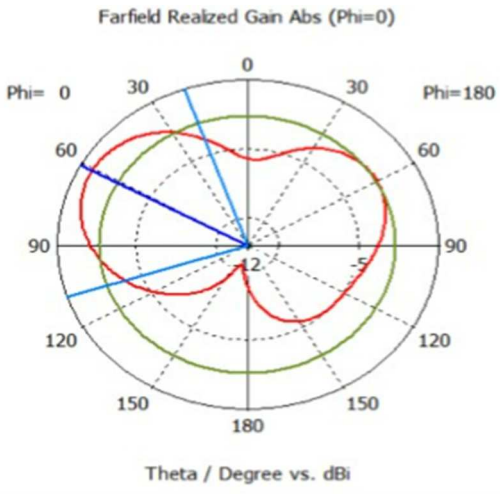
Fig.2 S-parameter of the proposed quad MIMO antenna. (a) S11 and S21 parameter of the quad MIMO-UWB antenna (b) S33 and S43 parameters of the quad MIMO-UWB antenna

2D-Farfield results of proposed quad MIMO antenna can be seen in Fig 3(a) and Fig 3 (b). In Fig 3(a), the farfield is investigated at frequency=5.5GHz and phi=90. It can be seen that the farfield pattern in following figure is in dumbbell-shaped. In Fig 3(b), the farfield is shown at frequency =5.5GHz and phi=0. In fig 4. It can be observed the 3D pattern of farfield results of the MIMO-UWB antenna. The 3D pattern will help more accurately in visualizing the farfield pattern.


Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:55:53 UTC from IEEE Xplore. Restrictions apply.




(a)




(b)

Fig 3. 2D-Farfields results of the proposed MIMO quad antenna design (a) Farfield of the MIMO quad antenna at frequency=5.5 GHz and phi=90 (b) Farfield of the MIMO quad antenna at frequency=5.5GHz and phi=0

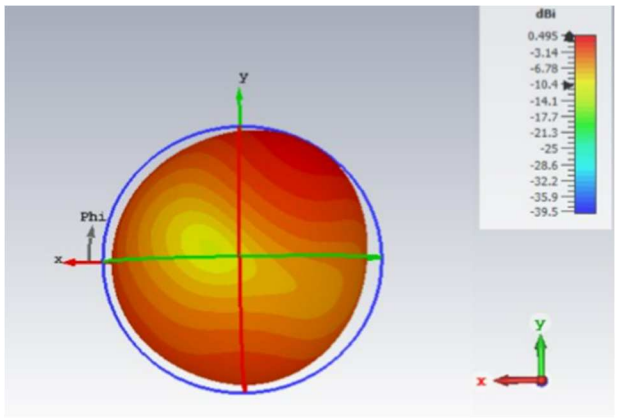



Fig 4. 3D-Farfields results of the proposed quad UWB antenna design at 5.5GHz

## IV. DIVERSITY PARAMETER

Diversity performance of the MIMO antenna is analyzed by considering its diversity gain (DG) and envelope correlation coefficient (ECC). Fig 5. Shows the DG of the quad MIMO antenna in UWB range and it has value more than 9dB in the entire range, which is considered good value for diversity performance. On the other hand, the ECC value of less than 0.2 is achieved over the band, which also signify that the patterns are uncorrelated, as shown in Fig.6. These diversity parameters show good MIMO performance of the quad UWB antenna.

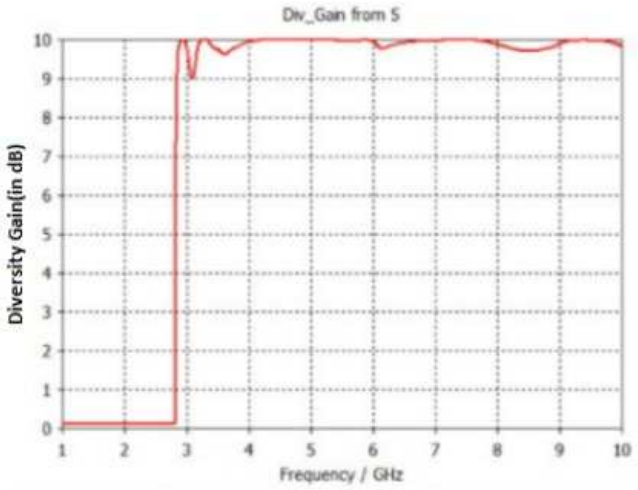



Fig.5 Diversity gain of the quad antenna

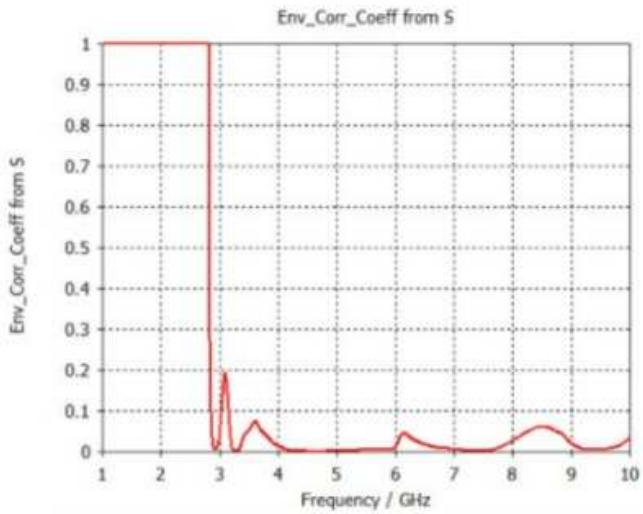



Fig.6 ECC of the quad antenna

## V. CONCLUSION

A highly optimized as well as good performance MIMO antenna, consisting of four UWB antenna, each



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:55:53 UTC from IEEE Xplore. Restrictions apply.

antenna consists of bi-rhombic antenna patches, and rhombus shaped patch is fed by microstrip line, is presented in the paper. A narrow rectangular stub is added to the ground plane which improves the decoupling effect of the antenna. The simulated antenna results show that the antenna has achieved the bandwidth of 7.2GHz ranging from 2.8GHz to 10.125GHz and the low mutual coupling of less than -18 dB throughout the UWB band. These results help us draw the conclusion that the proposed antenna has very small size and shows good performance. This quad patch MIMO Antenna can serve many revolutionary purposes in the field of communication and wireless technologies as it covers Ultra Wide Band.

## REFERENCES


[1] S. Tripathi, A. Mohan, and S. Yadav, "A compact Koch fractal UWB MIMO antenna with WLAN band-rejection," IEEE Antennas Wireless Propag. Lett., vol. 14, pp. 1565–1568, 2015.

[2] R. N. Tiwari, P. Singh, B. K. Kanaujia and P. Kumar, "UWB MIMO Antenna with Decoupling Strip for 5G Applications," 2021 International Applied Computational Electromagnetics Society Symposium (ACES), 2021, pp. 1-4,

[3] A. J. Paulraj, D. A. Gore, R. U. Nabar, and H. Bolcskei, "An overview of MIMO communications—A key to gigabit wireless," Proc. IEEE, vol. 92, no. 2, pp. 198–218, Feb. 2004.

[4] D. Sipal, M. P. Abegaonkar, and S. K. Koul, "Easily Extendable Compact Planar UWB MIMO Antenna Array," IEEE Antennas Wirel. Propag. Lett., vol. 16, pp. 2328–2331, 2017.

[5] Y. Xu, Y. Dong, S. Wen and H. Wang, "Vertically polarized quasi-Yagi MIMO antenna for 5G N78 band application", IEEE Access, vol. 9, pp. 7836-7844, Jan. 2021.

[6] L. Liu, S. W. Cheung, and T. I. Yuk, "Compact MIMO antenna for portable devices in UWB applications," IEEE Trans. Antennas Propag., vol. 61, no. 8, pp. 4257–4264, Aug. 2013.

[7] A. K. Saurabh and M. K. Meshram, "Compact sub-6 GHz 5G-multipleinput-multiple-output antenna system with enhanced isolation", Int. J. RF Microw. Comput. Aided Eng., vol. 30, no. 8, pp. e22246, Aug. 2020.

[8] H. Sakli, C. Abdelhamid, C. Essid and N. Sakli, "Metamaterial-based antenna performance enhancement for MIMO system applications", IEEE Access, vol. 9, pp. 38546-38556, March 2021.

[9] D. Sipal, M. P. Abegaonkar and S. K. Koul, "Compact printed UWB MIMO antenna with pattern diversity characteristic for portable devices," 2017 International Symposium on Antennas and Propagation (ISAP), 2017, pp. 1-2

[10] M. S. Khan et al., "Compact 4x4 UWB-MIMO antenna with WLAN band rejected operation," Electron. Lett., vol. 51, no. 14, pp. 1048–1050, 2015.




Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:55:53 UTC from IEEE Xplore. Restrictions apply.

Journal of Controlled Release 350 (2022) 698–715

Contents lists available at ScienceDirect

Journal of Controlled Release

journal homepage: www.elsevier.com/locate/jconrel

Review article

# Quantum dots: The cutting-edge nanotheranostics in brain cancer management

Pratik Chakraborty [a], Sabya Sachi Das [b,m], Abhijit Dey [c], Apala Chakraborty [d], Chiranjib Bhattacharyya [a], Ramesh Kandimalla [e,f], Biswajit Mukherjee [d], Abilash Valsala Gopalakrishnan [g], Sandeep Kumar Singh [b], Shubham Kant [h], Parma Nand [h], Shreesh Ojha [i], Pravir Kumar [j], Niraj Kumar Jha [h,**], Saurabh Kumar Jha [h,k,l,*], Saikat Dewanjee [a,**]

[a] *Advanced Pharmacognosy Research Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India*
[b] *Department of Pharmaceutical Sciences and Technology, Birla Institute of Technology, Mesra, Ranchi, Jharkhand 835215, India*
[c] *Pharmaceutics Research Laboratory, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India*
[d] *Department of Life Sciences, Presidency University, Kolkata 700073, West Bengal, India*
[e] *Department of Biochemistry, Kakatiya Medical College, Warangal 506007, Telangana, India*
[f] *Applied Biology, CSIR-Indian Institute of Technology, Uppal Road, Hyderabad 500007, India*
[g] *Department of Biomedical Sciences, School of Bio Sciences and Technology, Vellore Institute of Technology (VIT), Vellore, Tamil Nadu 632014, India*
[h] *Department of Biotechnology, School of Engineering & Technology (SET), Sharda University, UP 201310, India*
[i] *Department of Pharmacology and Therapeutics, College of Medicine and Health Sciences, United Arab Emirates University, Al Ain, P.O. Box 15551, Abu Dhabi, United Arab Emirates*
[j] *Molecular Neuroscience and Functional Genomics Laboratory, Delhi Technological University (Formerly Delhi College of Engineering), Delhi, India*
[k] *Department of Biotechnology Engineering and Food Technology, Chandigarh University, Mohali 140413, India*
[l] *Department of Biotechnology, School of Applied & Life Sciences (SALS), Uttaranchal University, Dehradun 248007, India*
[m] *School of Pharmaceutical and Population Health Informatics, DIT University, Dehradun 248009, Uttarakhand, India*

ARTICLE INFO

*Keywords:*
Bio-imaging
Blood-brain barrier
Brain cancer
Nanotheranostics
Photodynamic therapy
Quantum dots

ABSTRACT

Quantum dots (QDs) are semiconductor nanocrystals possessing unique optoelectrical properties in that they can emit light energy of specific tunable wavelengths when excited by photons. They are gaining attention nowadays owing to their all-around ability to allow high-quality bio-imaging along with targeted drug delivery. The most lethal central nervous system (CNS) disorders are brain cancers or malignant brain tumors. CNS is guarded by the blood-brain barrier which poses a selective blockade toward drug delivery into the brain. QDs have displayed strong potential to deliver therapeutic agents into the brain successfully. Their bio-imaging capability due to photoluminescence and specific targeting ability through the attachment of ligand biomolecules make them preferable clinical tools for coming times. Biocompatible QDs are emerging as nanotheranostic tools to identify/diagnose and selectively kill cancer cells. The current review focuses on QDs and associated nanoformulations as potential futuristic clinical aids in the continuous battle against brain cancer.

*Abbreviations:* ABCC, ATP-binding cassette C; Akt, Protein kinase B; BBB, Blood-brain barrier; BBTB, Blood brain tumor barrier; BCRP, Breast cancer resistant protein; bEND3, Brain-derived endothelial cells 3; CAT1, Cationic amino acid transporter; ChT, Choline transporter; CNS, Central nervous system; EGF, Endothelial growth factor; EGFR, Epidermal growth factor receptor; EPR, Enhanced permeation and retention; FRET, Forster resonance energy transfer; GLUT1, Glucose transporter; HER2, Human epidermal growth factor receptor 2; IL, Interleukin; LAT, L-type amino acid transporter; LDL, Low density lipoprotein; MAPK, Mitogen-activated protein kinase; MCT1, Monocarboxylate lactate transporter; MMP, Matrix metalloproteinase; mTOR, Mammalian target of rapamycin; NF1, Neurofibromin 1; NK cells, Natural killer cells; P-gp, P-glycoprotein; PI3K, Phosphoinositide 3-kinase; PIP3, Phosphatidylinositol (3,4,5)-triphosphate; PTEN, Phosphatase and tensin homolog deleted on chromosome 10; PTS6, Peptide transport system 6; QD, Quantum dot; QY, Quantum yield; RGD, Arginylglycylaspartic acid; ROS, Reactive oxygen species; SGLTs, Sodium-coupled glucose transporters; SiRNA-QD, Small interfering RNA-quantum dot; SPION, Supeparamagnetic iron oxide nanoparticles; TAT, Trans-activator of transcription; VEGF, Vascular endothelial growth factor; VEGFR, Vascular endothelial growth factor receptor.

* Corresponding author at: Department of Biotechnology, School of Engineering & Technology, Sharda University, Room # 311, Block-01, Plot No. 32-34, Knowledge Park III, Greater Noida, Uttar Pradesh 201310, India.
** Corresponding authors.
*E-mail addresses:* nirajkumarjha2011@gmail.com (N.K. Jha), saurabh.jha@sharda.ac.in (S.K. Jha), saikat.dewanjee@jadavpuruniversity.in (S. Dewanjee).

https://doi.org/10.1016/j.jconrel.2022.08.047
Received 18 May 2022; Received in revised form 22 August 2022; Accepted 23 August 2022
Available online 7 September 2022
0168-3659/© 2022 Elsevier B.V. All rights reserved.

## 1. Introduction

Quantum dots (QDs) are nano-sized crystals with unique optical and electrical properties, initially discovered in the 1980s by Alexis Ekimov and Louis E. Brus [1]. The energy states are quantized because the dimension range of QDs is lower than the Bohr radius, and this quantized energy level eventually gives rise to the most essential attribute of QDs, photoluminescence [2]. These nanodots can be formed from semiconductor materials including zinc, cadmium, and other metals, as well as organic compounds. The energy excites the electrons in semiconductor nanocrystals, and as they return to the ground state, they radiate energy. The photoluminescence properties of QDs comprise special characteristics including broad absorption and narrow emission spectra, along with convincing photostability [3]. These small nanocrystals (diameter 2–10 nm) absorb white light and after a few nanoseconds emit specific colours depending on the material-specific bandgap, extinction coefficient, and size [3]. QDs have shown promise in a variety of fields, ranging from quantum computing to cellular imaging. QDs have shown promise in nanobiotechnology as biological probes for cellular imaging and labelling due to their fluorescent capabilities [4,5]. Another intriguing application of QDs is their capacity to combine with antibodies, peptides, or other small molecules with high specificity and sensitivity for cancer therapy. Thus, along with molecular imaging, a QD-based probe can simultaneously deliver anticancer drugs and peptides to the targeted cancer cells. QDs are also capable of staining several tumor biomarkers at a time, which helps with an understanding of tempo-spatial molecular correlation during cancer progression in tumor microenvironment [6].

One of the most deadly diseases of the central nervous system (CNS) is brain cancer [7]. Brain cancer patients only live for roughly five years after being diagnosed [8]. Many diagnostic and therapeutic agents have failed to address brain cancers due to their inability to reach beyond the blood-brain barrier (BBB). QDs, among some other nanotheranostic systems, display potential promise to carry the agents across the barriers into CNS. The electronic and photoluminescence properties of ultrasmall QDs, thus have shown optimistic promise regarding specific targeting to the brain, displaying clinical potential to successfully address the problem of crossing the BBB, which largely limits most of the nano-scale delivery systems in this aspect. One of the most promising applications of QDs that has been documented thus far is brain targeting, which takes advantage of these nanocrystals' capacity to traverse the BBB. Because of this, QDs are a promising vehicle for delivering medicines to specific brain locations [9,10]. This review attempts to better understand and structure of applications of QDs in this specific frontier. The current article specifically covers detailed aspects of QDs for delivery into the brain beyond the barriers.

## 2. QDs: At a glance

QDs are tiny, spherical semiconductor nanodots having unique optoelectrical properties. During the initial days, QDs were capped with organic layers to form co-ordinate bonds with metal-core and stabilize the surface to prevent aggregation. Later, to make QDs compatible with aqueous systems, some hydrophilic groups were attached without disturbing the photoluminescence property. Clinically relevant QDs are broadly classified into three groups, such as semiconductor QDs, metal-doped QDs, and carbon-based QDs [11]. Photoluminescence of a semiconductor QDs is expressed when electrons within a semiconductor atom get back to the lower energy ground state from the excited state by emitting the absorbed energy in the form of light. The energy band gap decreases with an increase in QD diameter accompanied by the emission of longer wavelength (lesser energy) light [12]. Doping has been an effective approach to modifying the optical, electrical, and magnetic properties of QDs. Transition metal ions, such as $Mn^{2+}$, $Co^{2+}$, $Ni^{2+}$, $Ag^{+}$, $Cr^{3+}$, etc. have been doped into semiconductor nuclei to produce metal-doped QDs. By incorporating metals within the semiconductor core, some unique and long-lasting phosphorescence is observed [13]. The larger the energy band gap between the valence band and conductance band of the core semiconductor material(s), the more the scope to incorporate dopants into it [14]. The longer-lasting phosphorescence does not depend on fluorescent background for detectability; hence it is more suitable for imaging applications [14]. Carbon-based QDs have drawn much attention nowadays because of their great aqueous stability, low photobleaching, and high biocompatibility, all of which signal lesser toxicity [15,16]. Carbon-based and graphene-based QDs change their fluorescence emission peaks with changes in excitation wavelengths. Various reasons have been hypothesized behind this unique character, such as intrinsic band emission, triple ground state emission, electron-photon coupling, self-trapped excitations due to small organic molecules, etc.; however, the scientific community is yet to agree on a unanimous explanation [17,18]. Carbon-based QDs often require surface passivation to preserve their optoelectric properties [19]. Besides these three main classes, another type of QD is obtained by capping the surface of nanodots with certain polymer shells or copolymers. These are highly useful to carry out cellular imaging. Surface modification with organosilanes prevents the decomposition of QDs in aqueous media [20]. Freshly synthesized QDs usually present outstanding photoluminescent properties but they tend to aggregate and precipitate in contact with water. Thus, it becomes necessary to modify the surface hydrophilic moiety oriented toward an aqueous medium to achieve better colloidal stability by QDs in an aqueous medium [21]. Certain moieties can be conjugated with the QD surface to make it more compatible with biological systems. Biological uses of QDs require them to be modified into biocompatible probes by controlling reactivity and toxicity. Successful strategies for the biofunctionalization of QDs include hydrophilization, solubilization, and bioconjugation [10,22,23]. Bioconjugation can be achieved by the direct union, ligand binding, encapsulation, biotin-streptavidin binding, barnase-barster interaction, histidine-nickel nitrilotriacetic acid interaction, antibody-based interactions, and some bioorthogonal reactions [17,24–26]. QDs are often adjusted for emission peaks at the desired wavelength. The size of the dot and the energy gap between the excited state and ground state play vital roles in designing QD for specific emission peaks [20].

### 2.1. Photodynamic therapy

The photo-induced generation of reactive oxygen species (ROS) by QDs has received considerable interest in recent times in view of their potential use in the medicinal fields, especially in cancer diagnosis and treatment. Photodynamic therapy entails light-induced activation of a photosensitizer that kickstarts the process of formation of ROS finally leading to cell destruction. When energy, in the form of a photon, hits an electron within a semiconductor atom, the electron gets excited and moves to a higher energy band and electron-hole pair is generated. To come back to the ground state, the electron dissipates energy. The energy, thus released in the form of photons corresponds to the energy bandgap between the conductance band and valence band of the semiconductor (Fig. 1) [27,28]. Understandably, the wavelength of the emission is also controlled by this energy bandgap influencing the colour of emitted light. Smaller dots display a larger energy gap and hence smaller wavelength light is produced whereas larger dots correlate with a smaller energy gap and in turn produce light of a larger wavelength [29]. The core-shell configuration of QDs boosts the photoluminescence quantum yield (QY) of the core emission and shields the core from photobleaching as this configuration can limit the excitation and emission to the core [20]. QDs can act as energy donors via the Forster resonance energy transfer (FRET) mechanism to generate ROS [30]. Briefly, the energy delivered to QD from the external light source is transferred to an analyte and the specific emission of the analyte is been measured. If the distance between QD and the analyte exceeds a certain limit (Förster radius), the transfer of energy from QDs to the analyte becomes impossible leading to no analyte-specific emission [31].

Emerging evidence revealed the potential of QDs in nucleus-targeted photodynamic therapy. Streptavidin-functionalized CdSe/ZnS QDs conjugated with biotinylated plasmid DNA can cause DNA strand break upon photoactivation [32]. Several QDs have been shown potential to trigger ROS production by acting as electron (energy) donors to nearby oxygen molecules upon photoactivation in both normal and cancer cells including nerve cells [33–36]. In addition, QDs have been proven to be useful to distinguish between tumor and healthy cells and serve in cancer diagnosis [37]. Their enhanced stability, tunable emission spectra, the brightness of the produced image, and resistance to photobleaching make them superior candidates in delineation between tumor and healthy cells.

### 2.2. Biocompatibility

QDs have come up with a plausible promise for biological applications. However, for their successful use, the native QDs have to be tailored for the specific environment and biological applications [10]. In general, for biomedical applications, QDs should be small in size for deeper and better penetration. In addition, they should be efficiently excreted, bright and stable for sensitive and reliable detection, and exhibit minimal signs of toxicity [38,39]. Careful attention has to be paid to the nano-specific problems like very high surface reactivity due to large surface area and surface-to-volume ratio. The pharmacodynamic behaviour of QDs specifically their absorption, distribution, metabolism, and excretion are extremely important for drug delivery purposes. Pharmacological parameters may alter due to the quantum energy interactions. At the cellular level, QDs are accumulated either by receptor-mediated endocytosis or by means of opening ion channels through light-stimulated ionic interactions (Fig. 2). Targeted QDs can be accumulated in targeted tissue regions easily. Receptor-mediated targeting with QDs is gaining continuous interest for improved therapeutic benefits with low non-specific adverse effects [40–43]. Once inside the cell, following lysosomal degradation, QD releases free drugs into the cytosol. The core materials do not undergo any significant metabolism in the body. The biological half-life of QDs largely depends on surface modifiers [37]. The surface protectants (dopants/passivation agents/coatings/surface capping agents) tend to undergo photochemical and oxidative degradation, which results in exposing the core material. Graphene QDs are biodegradable and usually produce non-toxic degradation products [44]. QDs may interact with plasma proteins for excretion and eventually be taken up by the reticuloendothelial system [45–47]. QDs less than 5 nm in size are easily removed through kidneys [48]. Because of the metal elements present within the core, possibilities of metal-associated toxicity e.g. nephrotoxicity by Cd is a matter of concern [49]. QDs can persist for several months in the liver, spleen, and lymphatic system [50]. Hence, it is sensible to calculate the dose of QDs keeping in mind certain interrelated pharmacological and toxicological parameters including the rate and extent of tissue retention [51]. Clinical applications of QDs are somewhat hindered due to biocompatibility issues. QDs, in some cases, tend to generate ROS which might lead to non-specific organ toxicity [52–54]. Graphene-based QDs have exhibited oxidative stress-reducing properties implicating a way to do away with QD-induced non-specific organ damage [55,56]. Residual organic solvents may also impart some toxic effects. Since QDs display limited elimination in a biological system, the problem of accumulation of semiconductor metals and associated toxic effects becomes a major concern. However, owing to the high fluorescence of QDs, deposition in biological systems is easily measurable. In this issue, carbon-based QDs are gaining increasing interest to bypass the accumulation-associated

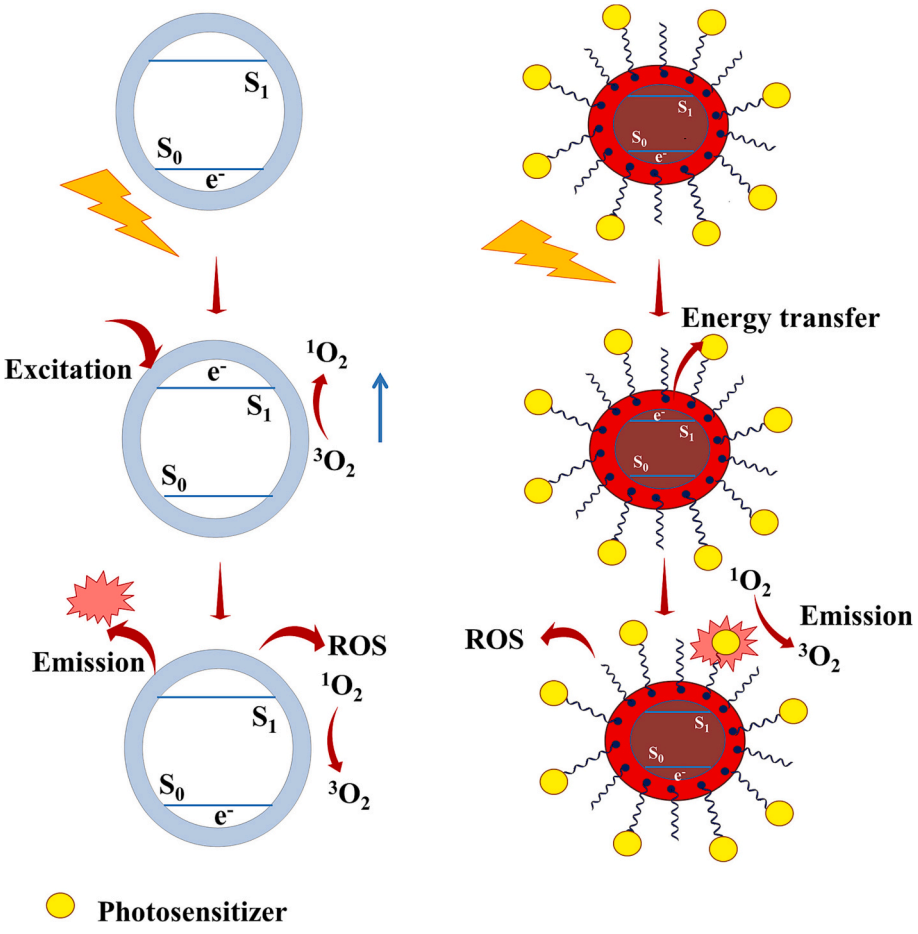



**Fig. 1.** Mechanism of photon-based excitation and fluorescent emission in QDs.

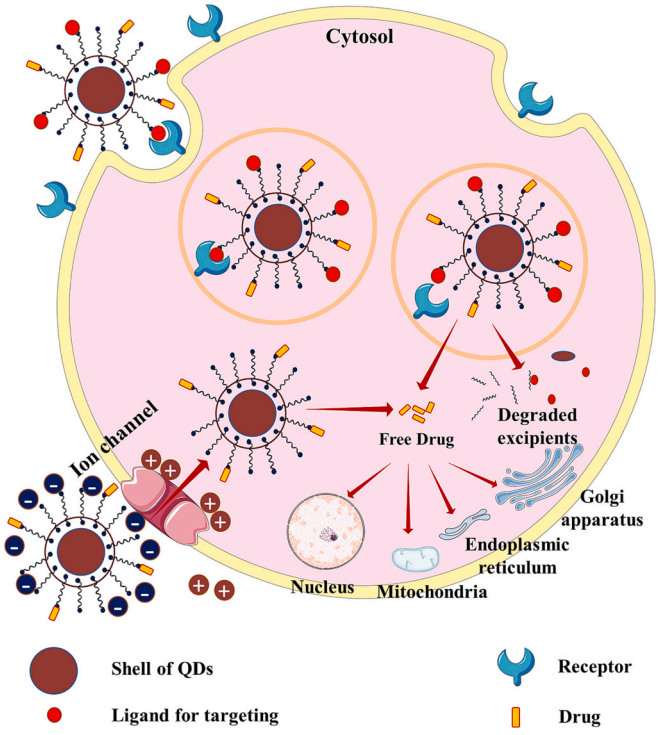



**Fig. 2.** Scheme of events during cellular accumulation of QDs and intracellular fate of the same.

problems [57].

In order to overcome biocompatibility concerns, researchers have come up with certain solution-oriented strategies. Successful attempts are being reported to eliminate organic solvents by taking refuge in sustainable green synthesis methods [58]. Newer developments like phosphorus-based QDs or organic QDs may highlight new facets of low-toxicity therapy [59–61]. Solubilization and bioconjugation steps serve to make QDs more biocompatible than they are just after synthesis, forming a QD-biomolecule complex capable of taking part in biological interactions [62]. Conjugation-based targeting can reduce the chances of non-specific toxicities to normal cells. Metal-based capping agents e.g. ZnS and CdS can be helpful to protect QDs from photo-oxidation and also improve fluorescence properties [63]. Zn and Cd act in the stabilization of membranes [64]. However, measures like the functionalization of QDs to improve water solubility and reduce toxicity often result in a larger size and lower yield than raw QDs [65]. The large surface-to-volume ratio of water-soluble, biocompatible QDs enables the conjugation of various biomolecules. The size, surface charge, capping material, functional groups, reactivity, and mechanical stability play vital roles in determining the biocompatibility aspects of QDs [66].

## 3. Brain cancer: A brief overview

Brain tumors are heterogeneous primary or metastatic neoplastic development in the central nervous system (CNS) that begins in neuroglia or precursor cells and are characterized by a poor prognosis and a low patient survival rate. In brain cancer, the neoplastic growth originates from systemic neoplasms and spreads further in the interior regions of brain parenchymal cells during the later stages of malignancy [7]. The development of primary brain tumors occurs within a few months, while secondary brain tumors arise from the lower-grade primary tumors. These brain tumors can be categorized as supratentorial and infratentorial based on their location [67]. Tumors can also be classified based on the cell types that are involved, such as astrocytoma, oligodendroglioma and oligoastrocytoma [68]. Astrocytoma, meningioma, and oligodendroglioma are supratentorial brain tumors; while, hemangioblastoma is infratentorial brain tumors.

Brain tumor patients usually survive for roughly five years after disease onset. The occurrence of glioblastoma is nearly doubled in the last two decades. Globally, around 3 per 100,000 persons develop some form of brain tumor yearly with males dominating the list. Factors including overexposure to ionizing radiation, ageing population, increasing pollution, stress etc. can be attributed to the continuous hike in the number of brain cancer patients during recent years. There is no known method of preventing cancer. Treatment usually involves surgery followed by chemotherapy and radiation therapy. Chances of relapse are very high in a non-fungible organ like the brain.

Emerging evidence revealed the molecular insights of malignant brain tumor pathogenesis. Most primary glioblastomas show epidermal growth factor receptor (*EGFR*) amplification and their activation [69]. *EGFR* variant III encourages cell proliferation and suppresses apoptosis by suppressing p27 in a PI3K/Akt/mTOR-dependent manner. PI3K/Akt/mTOR signaling plays a pivotal role in tumor progression via triggering angiogenesis, epithelial-mesenchymal transition (EMT), cell migration and cell invasion, and inhibiting apoptosis [70,71]. Phosphatase and tensin homolog (*PTEN*) is tumor suppressor gene that downregulates cell proliferation by constituting an inhibitory network on PI3K/Akt signaling via inducing enzymatic activity of phosphatidylinositol-3,4,5-trisphosphate (PIP3) [72]. The overexpression of the folate receptor, which is responsible for folic acid transport, is another significant marker of glioma. Folic acid is essential for DNA methylation of nucleotide synthesis during cancer cell division and proliferation. In glioblastoma cells, overexpression of folate receptors seeks to compensate for the demand for nucleotides and nutrients [73]. Mesenchymal glioblastoma is characterized by mutations of *NF1* and *PTEN*, followed by activation of mitogen-activated protein kinase (MAPK) and PI3K pathways [74]. The p53 pathway is responsible for the initiation of secondary glioblastoma. Loss of heterozygosity on chromosome 19q and chromosome 10, and overexpression of the *EGFR* gene are almost exclusively observed in glioblastoma. Tumor microenvironment also plays a vital role in the progression of the disease. Glioma stem cells are oncogenic and can self-propagate into the tumor mass. Mutations in transcription-regulating proteins, changes in chromosome copy numbers, and chromosomal translocations account for the development of brain cancers [70,75]. As a result of disrupted cell signaling, matrix metalloproteinases (MMPs) get upregulated. Upregulated MMPs facilitate glioblastoma cells to spread through neighbouring tissues using invadopodia by virtue of degradation of the extracellular matrix [76]. Widespread genotypic heterogeneity and high tumor interstitial pressure make the tumor cells relatively resistant to many therapy regimens [77].

## 4. Insights into the barriers to effective brain cancer management

Despite the fact that some anti-neoplastic drugs have shown promise against certain brain tumor cells, the success of *in vivo* brain targeting is still limited due to some critical factors. The transport of active pharmaceutical agents across the BBB is one of the major technical challenges faced by researchers to date. It is known that endothelial cells covering the cerebral capillaries along with some proteins chiefly comprising occludin, claudin and junctional adhesion protein, astrocyte end-feet and pericytes make up the BBB, forming a selectively permeable diffusion barrier. The presence of efflux transporters such as P-glycoprotein (P-gp), peptide transport system 6 (PTS6) and breast cancer-resistant protein (BCRP) limits the entry of most of the drug molecules into CNS [78]. P-gp efflux transporters ensure poor uptake of lipophilic drugs to the brain and also efflux out cationic substances from CNS [79]. Multi-drug resistant proteins belonging to the ATP-binding cassette C subfamily (ABCC) are responsible for the outflow of anionic substances from the brain [80]. Selective influx transporters present within the BBB include L-type amino acid transporter (LAT1), glucose transporter

(GLUT1), monocarboxylate lactate transporter (MCT1), cationic amino acid transporter 1 (CAT1), choline transporter (ChT), and sodium-coupled glucose transporters (SGLTs) [81,82]. Transferrin receptors, present in the BBB might facilitate the transfer of transferrin-conjugated cargos across the BBB [83]. The tight junction proteins, such as MMPs and claudin 4 play a crucial role in the migration of specific components through the BBB [84]. In addition, the BBB restricts anything having a molecular weight of more than 500 Da. The charged molecules can enter the brain quickly; bigger polar molecules may be preferred for receptor-mediated transcytosis due to appropriate ligand binding. However, the majority of anticancer drugs are hydrophobic, with large molecular weights and are substrates for enzymes present in the BBB [85]. Hence, they fail to translocate through the BBB. Again, enzymes present within the BBB, such as aminopeptidases and endopeptidases may convert the drug/entire cargo into an ineffective counterpart. Moreover, the drug may bind to a nontransporter protein, rendering the drug static and ineffective. In brief, the poor prognosis of glioblastoma is a combined result of limited access through drug delivery cargos and surgical procedures, multi-drug resistance and the critical ability of residual malignant cells to grow into a new tumor [86].

The angiogenic vasculatures originating from tumors constitute a modified barrier called the blood brain tumor barrier (BBTB). In comparison to the BBB, endothelial tight connections are compromised in the BBTB, making it more permeable [87]. Molecules as large as 12 nm can cross the BBTB [70]. Thus, BBTB may help in the accumulation of certain nanocarriers into tumor cells. Interestingly, it also boosts the enhanced permeation and retention (EPR) effect [87]. Tumor-mediated changes to form the BBTB from the BBB may vary with tumor type, stage, location, and size. The difference between the BBB and BBTB becomes more prominent in higher grades of glioma with increased diameters and surface area of local microvasculature [88]. Interestingly, these structural deviations do not necessarily correspond to the degree of the infiltration of the normal brain by tumor cells [89]. An additional layer of complexity is added by the regional heterogeneity of immune cell populations that influence BBTB in gliomas [88]. Vasculatures within BBTB facilitate the diversion of nutrients and oxygen to feed tumor cells [90]. During cancer progression, the integrity of BBTB gets compromised leading to the EPR effect [91].

The tumor microenvironment is a critical factor in cancer chemotherapy. Tumor progression is largely dependent on interactions between tumor cells and their immediate microenvironment. Tumor cells tend to control extracellular matrices in the surrounding microenvironment to promote chemoresistance, tumor cell proliferation, and differentiation [92]. Tumor-associated macrophages lack cytotoxic activity which actually encourages cancer growth during cancer-related inflammation. Tumor-associated macrophages also aid in protecting cancer cells in necrotic regions by secreting interleukin 10 (IL-10), VEGF, and nitric oxide synthase [93]. Tumor cells, by virtue of secreting death ligands doped with exosomes, are able to deactivate T cells and natural killer (NK) cells [94]. In case of injury and inflammation, the tumor microenvironment preferentially recruits tumor-associated cells like tumor-associated macrophages, tumor-infiltrating lymphocytes, tumor-associated fibroblasts which lack cytotoxicity over normal T cells. Glioma stem-like cells are another crucial barrier in glioblastoma treatment [95].

To effectively treat brain cancer, one must first accurately locate and characterize the tumor using imaging techniques, then select a delivery system, optimize circulation time, ensure passage within the CNS with specific targeting ability, strategize access beyond the tumor's periphery, and confirm intracellular uptake. Once inside the cell, early lysosomal destruction must be avoided so that the active pharmaceutical ingredients can act on their targets to induce apoptosis.

### 4.1. Conventional approaches versus QDs

Very few conventional drug delivery systems can successfully penetrate the BBB to place the drug at the target site. Understandably, the current treatment of glioblastoma largely depends on surgical procedures and radiotherapy where chemotherapy plays only an adjuvant role [70]. The plethora of unwanted toxic effects of anticancer drugs including but not limited to damage to DNA and otherwise healthy tissues sometimes seem to overshadow their therapeutic benefits. Approaches, such as formulations with prodrug therapies (antibody-directed enzyme prodrugs therapy, gene-directed antibody prodrug therapy and gene prodrug activation therapy), have shown promise in drug delivery across the BBB [96,97]. Modifying tight junction permeability with hyperosmotic drugs and boosting intracellular $Ca^{2+}$ concentration may offer some hope, although these are both non-specific and dangerous approaches [98]. The main challenge for the delivery of chemotherapeutic agents to the target site(s) of CNS lies in reaching the medication across protective barriers and affecting the cancer cells selectively, ensuring minimum cytotoxicity to healthy cells. Among chemotherapeutic agents, dalargin became the first drug that has been successfully delivered to CNS using nanotechnology [96]. Current diagnostic and therapeutic strategies in clinical use have been detailed in Table 1.

The development of different nanoscale carriers has shown promise in delivering cancer therapeutics and tumor imaging agents to CNS [114]. Nanocarriers offer advantages over conventional cargo in that they can improve tissue distribution of chemotherapeutic agents and preferential accumulation within the brain with proper structural and surface modifications [115]. Oligoglycerolipid nanosuspension and cell-penetrating small peptide-grafted nanoparticles have shown promise in CNS-targeted drug delivery through manipulations of endothelial lining [116,117]. Nanoformulation tailored to target insulin receptor, low-density lipoprotein (LDL) receptor, transferrin receptor, neonatal Fc receptor, and leptin receptor could be an effective approach to receptor-mediated drug transport to CNS [117,118]. Clathrin and caveolin-mediated endocytosis offer effective transport of nanomaterials inside the CNS in an energy-dependent manner [119]. In addition, tailored nanoparticles have shown potential in cell-mediated transport (Trojan horse mechanism) of drugs via macrophages and leukocytes that translocate to CNS after engulfing the drugs [120]. In various preclinical studies, surface-modified polymeric nanoparticles showed promise in improving CNS targeting of different chemotherapy drugs [121,122].

The emerging demand for less toxic and more efficacious personalized therapy can be answered by QDs as this particular type of nanocrystals are capable to track the changes in a tumor microenvironment, and anti-tumor therapy at the same time [123]. QDs are unique and superior to any other conventional nanotheronostics as multiple imaging labels for different imaging modalities can be attached simultaneously to a single nanocrystal providing significant signal amplification [6]. Further, more than one targeting ligand can be attached to the surface of QDs to enhance the binding specificity as well as affinity. QDs are able to integrate some specific biomarkers which are able to cross the BBB. Moreover, distinctive surface and structural properties (such as uniform and tunable size, as well as flexible drug loading) have enabled the use of QDs in targeted drug delivery [124]. All QDs generate broad absorption bands, indicating that a single excitation source can efficiently excite QDs with different emission wavelengths [125]. QDs are more photostable owing to their intense fluorescence and higher resistance than commercial imaging probes making them suitable for tracking cellular events over a longer period of time [13]. QDs, owing to their extremely small sizes, are able to move through small blood capillaries and gain access to cancer cells in the target tissue [126]. The process of passive diffusion allows QDs (which might be carrying therapeutic and/or diagnostic agents) to pass through the BBB and enter CNS [127]. If a diagnostic molecule is attached to QD, it is also possible to sense the level of curative effect within the brain tumor [128,129]. Transferrin-conjugated quantum rods were observed to successfully transmigrate through the BBB *in vitro* in a concentration and time-dependent manner [130]. Carbon-based QDs have shown the potential

**Table 1**
Current clinical tools for management of brain cancer.

| Sl No. | Mode of imaging/ treatments | Clinical implications | Limitations | References |
|---|---|---|---|---|
| 1 | Magnetic resonance imaging | First method of choice; fast; non-invasive; no radioactivity; high-resolution imaging. | Repeated exposure to high magnetic fields may be damaging; agents used for contrast may cause renal toxicity. | [99,100] |
| 2 | Positron emission tomography | Radiotracer technique; often used to complement computed tomography. | Radioactive hazards. | [101] |
| 3 | Computed tomographic scan | Fast; noninvasive; can produce three-dimensional images. | Ionizing radiation is used; sometimes another imaging technique is required to obtain whole anatomical information. | [102] |
| 4 | Ultrasonography | Ultrasound is used to produce biological images analysing reflections; no radioactivity is involved; may also aid in the disruption of the BBB and BBTB. | Image clarity and resolution do not always meet the requirements. | [103] |
| 5 | Surgery | Increases lifespan of patients. | Difficult to locate tumor boundary for optimal resection; complete removal not always possible. | [104] |
| 6 | Chemotherapy | Cytotoxic and cytostatic agents kill tumor cells, frequently used to kill residual cells after surgical resection. | Difficult to deliver into CNS; drug resistance; non-specific, long-lasting toxicity to healthy cells also. | [105,106] |
| 7 | Radiotherapy | Ionizing radiation kills tumor cells. | Affects peripheral cells only, stem-like cells can again proliferate into the tumor; unwanted effects on neighbouring tissues; radio-resistance. | [107,108] |
| 8 | Gene therapy | Inhibition of oncogene expression using complementary genetic codes. | Difficult to distribute delivery vector into CNS. | [109] |
| 9 | Immunotherapy | Boosts immune system with peptides, monoclonal antibodies etc. to cope up with immune suppression imposed by tumor cells. | Unwanted immunological reactions often overcome clinical benefits. | [77] |
| 10 | Alternating electric field | Optune produces electric field to interact with cellular components required for tumor cell growth. | Can not suffice alone, beneficial as an adjuvant treatment. | [110] |
| 11 | Targeted therapy | Precise targeting of certain molecular events during tumor growth using drugs. | Every tumor does not contain an available target for approved drugs. | [111] |
| 12 | Hyperthermia | Heat-induced apoptosis. | Unwanted effects on surrounding tissues and organs. | [112] |
| 13 | Sonodynamic therapy | Ultrasonic wave open tight junctions to aid tissue penetration, and activate sonosensitizers near the tumor to kill cancer cells. | Not sufficient alone, shows synergistic effect in combination with other modes of treatment. | [113] |

to cross the BBB for successful drug delivery to CNS mainly through passive diffusion, and through glucose transporters owing to suitable surface modification to facilitate the same [131,132].

### 4.2. Receptor interactions and specific targeting by QDs

Human epidermal growth factor receptor 2 (HER2), in particular, has been revealed to be a potential tumor-targeting strategy [133–135]. The trackable delivery of EGFR-targeted QDs into tumor tissue opens up new possibilities for imaging and treating various malignancies. Over-expression of EGFRs in brain tumor cells has been utilized for specific targeting of glioma cells with bioconjugated QDs [136–139]. Dudu and colleagues provided the first quantitative evidence regarding the association between EGF concentration gradient and chemotaxis of medulloblastoma-derived cells [140]. QDs coupled with anti-HER2 antibodies have shown exciting potential in terms of active tumor targeting [134]. The attachment of Trastuzumab, a monoclonal antibody specific for HER2/EGFR2, to QDs was a successful attempt to target mouse breast cancer, and it may open up a new option for improved glioma cell targeting efficiency [134,141,142]. EGFR-targeted QDs have demonstrated overwhelming promise in the field of non-invasive bioimaging for preoperative delineation between tumor cells and healthy cells [143–145]. The value of fluorescent probe-labelled EGFR-targeted QDs in brain tumor imaging has been demonstrated elsewhere [146,147]. Vibin and peers experimentally proved the supremacy of EGFR monoclonal antibody-conjugated QDs over unconjugated QDs in the bioimaging of solid tumors [148]. Due to reduced photobleaching and efficient imaging, as well as selective peptide-based targeting, HER2-conjugated QDs have demonstrated encouraging results in identifying metastasizing tumor cells [149]. Combining all, it would be worth mentioning that suitably engineered QDs to target EGFRs can be used in the effective targeting of glioma cells.

Overexpressed folate receptors have emerged as possible targets for gaining control of solid tumors [150]. Folate receptor-targeted QD-liposome conjugates have been developed to capitalize on overexpressed folate receptors in various types of malignancies and have shown to have higher cellular internalization than non-conjugated QDs [151,152]. Emerging experimental results hint toward folate conjugation of QDs as an effective strategy to deliver therapeutic moieties into tumor cells displaying overexpression of folate receptors [153–155]. Over-expression of folate receptors is also characteristic of glioblastoma, hence similar benefits can be expected in the case of brain cancer also [150,156]. In addition, folic acid-conjugated QDs have shown potential in the diagnosis of folate receptor-expressing malignancies [157]. The probe function via folate receptor-mediated endocytosis reveals bio-imaging potential in the case of brain tumors as well [62].

Transferrin receptors, which are essential components of the BBB, can be targeted to drop delivery cargo to the CNS. It has been revealed that transferrin receptors can indulge in an endocytosis mechanism to transport QDs effectively [158]. Conjugation of QDs with peptide T7 has been evidenced to enhance intracellular uptake of the QD-aptamer complex [159]. Overexpression of transferrin receptors on cell surfaces of glioblastoma multiforme may aid in delivering chemotherapeutic agents in a targeted manner [160]. Caveolin 1-specific targeting with QDs to reach beyond BBB is also displaying promises [161,162].

The use of anti-VEGF monoclonal antibody-linked QDs to track solid tumors is gaining attention as a tumor prognostic method. Breast cancer has been successfully detected using QDs in combination with a humanized anti-VEGF 165 monoclonal antibody [163]. Since VEGF overexpression is also found in grade IV brain tumors, this strategy could provide a new way of active targeting and detecting brain cancers [164].

QD fluorescent markers have been designed for endocytic tracking of tumor cells [165]. QD fluorescent marker-based tracking of endocytic pathways might lead to an easier process for bioimaging through the bioconjugation of QDs with tumor-specific antibodies. Understanding and tracking endocytic pathways with QDs could lead to improved diagnosis and treatment of malignancies including gliomas [166]. A proper understanding of neuronal transporters in detail would empower researchers to better understand the functions of the brain. QDs have allowed a big leap toward understanding the intricate details of neuronal transporters by imaging single dopamine transporter with an unparalleled precision of 20 nm [167].

## 5. Applications of QDs in brain cancers

Nanocarriers, including QDs, have demonstrated their potential in improving the biodistribution of therapeutic moieties to the specific target, reducing toxicity owing to the non-specific distribution of drugs [168]. The major goals in brain cancer management are early diagnosis and targeted delivery of therapeutic agent(s) without developing significant toxicity. Principle obstruction for drug delivery to CNS is the persistence of BBB which permeates very selective materials from blood to the brain and vice-versa [169]. QDs are attractive agents to deliver therapeutic agents through the barriers of the brain. Moreover, QDs comprising clusters of 10 to 1000 atoms are becoming increasingly popular with time for diagnostic applications. Targeting ligands, such as peptides, antibodies, nucleic acids, etc., can be conjugated to the surface of QDs to gain affinity and targeted delivery to the cancer cells. It is also possible to sense the level of curative effect within the brain tumor by attaching a diagnostic molecule. Table 2 enlists a collection of QD-based approaches regarding brain tumors to reach beyond protective barriers.

### 5.1. Diagnosis and bioimaging

The difficulty in the early detection of brain tumors is one of the key reasons for their high mortality rate. QDs cope with futuristic promises regarding non-invasive imaging of brain tumors. Proper photosensitization along with fluorescence tuning and ligand binding make QDs very promising tools for tumor detection (Fig. 3). Unique optoelectrical properties demonstrated by QDs make them attractive tools for bioimaging applications. Back in 2004, scientists succeeded to obtain subcellular images of mouse brains for optical biopsy-mediated tumor diagnosis utilizing QD-labelled serum [170]. The inclusion of the gradient index lenses enables capturing images of blood vessels in the brain of intact animals with minimum invasion up to almost 800 μm deeper in the brain. Although this strategy could be useful for imaging overall deep brain tissue, for specific tumor imaging, QDs require tailoring with suitable tumor-specific ligands. QDs bring about supremacy in bio-imaging owing to broad and tunable excitation and emission, high QY, and good photostability [206]. Emissions in the near-infrared region are desirable for high-resolution imaging as the infrared spectrum is subjected to lower absorption and scattering in biological tissues [207]. Bioimaging through near-infrared photoluminescence within the brain depicted promising results utilizing QD-dendrimer conjugate [182]. In some cases, QDs sneak into tumor cells by getting phagocytosed by macrophage cells (Trojan horse mechanism) and emit fluorescent signals [171,172]. Qdot ITK Amino(PEG) QDs achieved accurate detection of glioblastoma in rats co-localizing with macrophages in glioma cells without significant toxicity. These inflammatory cell-targeting ODs have demonstrated the ability to saturate the phagocytic capacity of the reticuloendothelial system at nanomolar intravenous doses and are capable to outline tumor borders by getting phagocytosed by the tumor-infiltrating macrophages. Lengthening the PEG chain could improve the systemic half-life of the QDs [171]. Since receptor-mediated targeting typically fulfils the dual aims of passage across the BBB and better uptake into malignant cells overexpressing the same kind of receptors, receptor-specific targeting appears to be preferable to the macrophage-dependent method [189,208]. Folate receptor alpha or EGFR variant III could serve as potential targets to recognize brain cancer cells for fluorescence imaging. Overexpression of EGFRs has been successfully employed for the active targeting of brain tumors by convection-enhanced delivery of EGFR-targeted immunoliposomes conjugated to QDs [177]. Although real-time, non-invasive monitoring of fluorescence is advantageous, delivery of QD-immunoliposomes by intracranial xenografts might limit rampant clinical application. Anti-EGFR variant III antibody-bioconjugated near-infrared QDs (Qd800) improved imaging of mouse orthotropic glioblastoma cells; the antibody, through its Fc region, can specifically target EGFRs on the surface of cancer cells in the brain corresponding to high resistance and invasiveness of the same [178,209]. Interestingly, further conjugation with cysteine residues additionally improved cellular uptake of Qd800 into U87 cells through folate receptor recognition, and certain transmembrane peptides also exhibit affinity toward cysteine groups [178]. However, utilizing EGFRs for specific targeting is a risky option as point mutations appearing from crosstalk between wild-type EGFR and EGFR variant III might lead to resistance instead of target specificity [210].

Non-invasive imaging of glioma cells took a big leap when glycine precursor-based carbon QDs could be uptaken by C6 glioma cells in an energy-dependant manner taking advantage of a faster rate of metabolism of tumor cells followed by a higher rate of ATP-dependant endocytosis than non-affected cells. In addition to energy-dependent pathways, passive targeting can also be used to gain entry into tumor cells. Nitrogen-doped carbon QDs coated with 1-methyl-2-pyrrolidinone was able to enter glioma cells via passive targeting and aid in fluorescence imaging of the same [183]. Additionally, these QDs harmonize great water solubility with a lack of inherent cytotoxicity, a highly desirable feature, opening up the possibility of clinical optimism. Overexpression of specific proteins on glioma cells can be targeted for real-time imaging and dynamic tracking purposes. In this aspect, tenascin C, an extracellular matrix protein responsible for the invasion of glioblastoma into surrounding tissues can be targeted. A single-stranded DNA aptamer (GBI-10) conjugated QD-immunoliposomes effectively produced high-quality images by targeting tenascin C on human glioblastoma cells [173]. GBI-10-tailored QDs exhibited key features including stability, uniformity, monodispersity, and strong fluorescence characteristics *in vitro*. However, the diagnostic promise of QD-labelled aptamer is yet to be fully explored *in vivo*. By virtue of functionalization and narrow emission bands, QDs can produce sharp images of multiple biomarker-associated cells simultaneously [37]. Chemically functionalized graphene QDs with amino, carboxyl, or amide groups showed promising potential for establishing themselves as biocompatible alternatives for the bioimaging of C6 glioma cells with minimal toxicity [181]. The findings of Yuan and colleagues [181] effectively answered earlier doubts raised by Marcovic and coworkers [175] regarding the potential photodynamic cytotoxicities of non-functionalized graphene QDs and established them as biocompatible brain tumor diagnostic option. Carbon QDs functionalized with maleimide-PEG-aminosuccinimide succinate conjugated with RGERPPR peptide preferentially distribute themselves within glioma tissue through and clear images were obtained using fluorescence techniques [190]. The inclusion of RGERPPR peptide within the multifunctional carbon QDs facilitates penetration through tumor vasculature to target glioma cells via neuropilin 1 receptors abundant on tumor blood vessels and glioma cells and thereby suppressing the inadequacy of the EPR effect. The target specificity achieved by Gao and colleagues [190] might open up a novel theranostic approach to brain tumor therapy in the future. Biofunctionalized with L-cysteine or poly L-arginine and ligand-capped AgInS2 QDs were internalized by glioma cells paving a new avenue toward bio-imaging and bio-labeling with fluorescent QDs [193,211]. Arginine-rich peptides are well-known for their cell-penetrating abilities, cysteine, on the other hand, is preferentially taken up by neoplastic glioblastoma cells owing to the abundance of specific cell membrane receptors [212,213]. These nanohybrids proved to be excellent alternatives for bioimaging as they carry the rare combination of target specificity with no cytotoxicity.

Interestingly, too many functionalities for target specificity, seemingly affect the optical properties of QDs somewhat robbing them of their primary diagnosis purpose [193]. Green synthesis is gaining prominence in the preparation of QDs, like many other fields. Experimentally, green synthesized QDs acted overwhelmingly well regarding bioimaging of brain tumors displaying novel futuristic promise [198]. Green synthesized QDs are currently being sought after actively to extract their clinical potential [214].

**Table 2**
QD-based approaches for brain specific targeting.

| Sl No | QD types | Bioconjugation/ attachments | Delivery cargos | Size ranges (nm) | Photoluminescence emission peak (nm) | Applications | References |
|---|---|---|---|---|---|---|---|
| 1 | QD-labelled serum | – | QD-labelled serum | – | – | *In vivo* subcellular imaging in mouse deep brain tissue for optical biopsy. | [170] |
| 2 | Intravenous QDs (Qdot) | – | Qdot ITK amino (PEG) QD | – | 705 | Qdots colocalized with macrophages in C6 gliosarcoma cells at a higher dose to produce detectable optical signals for tumor location. | [171] |
| 3 | Intravenous QDs (Qdot) | – | Qdot ITK amino (PEG) QD | – | 705 | Qdots colocalized in rat gliosarcoma cells with macrophages in a dose-dependent manner. | [172] |
| 4 | Streptavidin-coated QDs | Anti-EGFR antibody | – | – | – | Rapid visualization of EGFR overexpressing tumor cells viz. U87, SKMG-3 and frozen tissue samples of glioblastoma multiforma by antibody-conjugated QD. | [136] |
| 5 | QD-aptamer complex | – | GBI10 DNA aptamer | 28.8 ± 3.5 | Fluorescence peak intensity is more in the case of the QD-aptamer complex than QD-COOH at the same wavelength. | Tenascin C targeted DNA aptamer applied in biosensing of glioma cells. | [173] |
| 6 | Quantum rod | – | Quantum rod | Length: 14–26, Diameter: 4.0–6.5 | 608–658 | Nanocarrier to deliver agents beyond the BBB *in vitro*. | [130] |
| 7 | Streptavidin-coated PEGylated QDs | EGFR specific proteins | – | – | – | Selective labelling of tumor cells for better preoperative delineation from healthy cells. | [143] |
| 8 | CdSe/CdS/ZnS core-shell QDs | Thiol-modified RGD and HIV Tat | SiRNA-QD conjugate | 7 | 623 | Targeted delivery and tracking of siRNA molecules for the knockdown of the EGFR variant III gene in U87 cells. | [174] |
| 9 | Streptavidin conjugated QD655 | EGFR antibody | – | – | – | QD binds with the extracellular domain of EGFR to selectively label cells with EGFR overexpression. | [137] |
| 10 | Graphene QDs | – | Graphene QDs irradiated with blue light (470 nm) | Diameter: 56.6 ± 8.7 Height: 1.9 ± 0.8 | 460 | Generates ROS and singlet oxygen species to kill U251 cells by oxidative stress and autophagy induction. | [175] |
| 11 | Apomorphine loaded QDs | – | QD-loaded liposome | 142.0 ± 0.3 | – | Increased delivery of apomorphine to cancer-affected cells in brain. | [176] |
| 12 | CdSeTe/ZnS core/shell QDs | PEGylated lipids | Convection-enhanced delivery of EGF receptor-targetted QD-immunoliposome | – | Anti-EGFR QD immunoliposomes displayed higher fluorescence than carboxyl QDs and QD-liposome complexes at the same concentration. | Bioimaging underwent efficient uptake *in vitro* and displayed 1.3- to 5.0-fold higher total fluorescence in intracranial brain tumor xenografts *in vivo*. | [177] |
| 13 | Near-infrared QDs (Qd800) | Anti-EGFR variant III antibody combined with cysteine | Qd800 | – | – | *In vitro* uptake by U87 cells and improved diagnosis with better contrast *in vivo* (mouse with orthotropic glioblastoma). | [178] |
| 14 | Carbon QDs | – | – | <5 | 500 | Fluorescent, non-invasive imaging of C6 cells after endosomal uptake. | [179] |
| 15 | Carbon QDs | Angiopep-2 | Angiopep-PEG carbon nanodot | 38.5 ± 1.9 | 600–700 | Non-invasive imaging of C6 glioma cells, targeting C6 glioma cells through receptors expressed in glioma. | [180] |
| 16 | Graphene QDs | -$NH_2$<br>-COOH<br>-$CON(CH_3)_2$ | – | 7.5<br>15<br>3–10 | 500<br>440<br>500 | Bioimaging of C6 glioma cells with high biocompatibility. | [181] |
| 17 | CdTe/CdS core/ shell QDs | – | Dendrimer | 4.05 ± 0.3 | 650 | Bioimaging, through near-infrared photoluminescence. | [182] |
| 18 | | Nitrogen | | 5–15 | 420 | | [183] |

(*continued on next page*)

**Table 2** (*continued*)

| Sl No | QD types | Bioconjugation/ attachments | Delivery cargos | Size ranges (nm) | Photoluminescence emission peak (nm) | Applications | References |
|---|---|---|---|---|---|---|---|
| | Polymer-coated carbon QDs (core-shell) | | Polymer-coated nitrogen-doped carbon QDs | | | The formulation could enter glioma cells by passive targeting aiding fluorescent imaging of the same without significant cytotoxicity. | |
| 19 | Carbon QDs | – | Carbon QD-Aspartate | 2.28 ± 0.42 | 550 (when excited at 460 nm, concentration 0.4 mg/ml) | Successful imaging and targeting of C6 glioma cells in a mouse model. | [184] |
| 20 | Cd/Te QDs | Transferrin | QD-transferrin bioconjugate | 3.1 | 626 | Enhanced internalization within glioblastoma cells. | [185] |
| 21 | QD-loaded cationic liposome | – | Cationic liposome, assisted by focussed ultrasound | Bare QD: 3.05, QD loaded cationic liposome: 175.26 | – | Ultrasound aided disruption of the BBB. | [121] |
| 22 | Carbon QDs | Nitrogen | – | 2.6 | 452 | Real-time live cell imaging, traceable beyond the BBB. | [186] |
| 23 | QDs | Arginine-glycine-aspartic acid | Docetaxel-loaded, arginine-glycine-aspartic acid-conjugated QD-liposome | – | – | Targeted co-delivery of docetaxel, QD into brain for therapeutic and imaging purposes, respectively. | [187] |
| 24 | QDs | Transferrin | Docetaxel loaded, transferrin-conjugated QD-liposome | – | – | Improved brain targeting by theranostic liposome. | [83] |
| 25 | Graphene QDs | – | pH-responsive fluorescent graphene QDs | Diameter: 4, Height: 1.6, Lattice spacing: 0.21 | 480 (at pH $<$ 6.8), 440 (at pH $\geq$ 6.8). | Allows spotting of solid tumors during the initial phase of their development. | [188] |
| 26 | ZnS/CdSe QDs | Biotinylated aptamer | QD-aptamer bioconjugate | 20 | 605 ± 5 | Nanoprobe used to bind with EGFR variant III on glioma cell surface (mouse U87 cells) and guide resection surgery. | [189] |
| 27 | Carbon QDs | RGERPPR peptide | Functionalized carbon QDs | 9.0 ± 2.0 | 460 | Bioimaging of glioma tissue in mouse model. | [190] |
| 28 | Carbon QDs | – | – | 2.49 ± 0.42 | 524 | Good targeting ability on C6 glioma cells. | [191] |
| 29 | Magnetic QDs | – | The hybrid complex of liposome with superparamagnetic iron oxide nanoparticles, QDs and cilengitide | 100 ± 1.24 | Red fluorescence emission | Nanoprobe for magnetic targeting of glioma cells to guide surgical resection. | [192] |
| 30 | Fluorescent $AgInS_2$ QDs | L-cysteine and poly-L-arginine | Carboxymethyl cellulose-QD bioconjugate | 3.5 ± 0.2 | 650 | Fluorescent bioimaging of glioma cells using nanoprobe. | [193] |
| 31 | Doxorubicin loaded ZnS QDs | – | Carboxymethylcellulose-QD conjugate | 57.5 ± 0.4 | Strong emission at violet-blue region of visible spectra | Nanoprobe for bioimaging and sustained delivery of doxorubicin to glioma cells *in vitro*. | [194] |
| 32 | AgInS/ZnS core/shell QDs | Anti-VEGF receptor antibody | Chitosan conjugated semiconductor QDs | 4.4 ± 1.0 | 620 | Imaging and targeted killing of glioblastoma multiform cells *in vitro*. | [195] |
| 33 | Carbon QDs | – | Saccharide-based carbon QDs | – | – | Nanocarrier to deliver agents beyond the BBB in wild zebrafish model and rat model. | [131] |
| 34 | Core-shell chitosan-polyethylene oxide carbon QDs | Temozolomide | Core-shell chitosan-polyethylene oxide carbon QDs/carboxymethyl cellulose-polyvinyl alcohol nanofiber implant | $\leq$5 | 443 | Increased cytotoxicity of temozolomide against U251 MB cells *in vitro*. | [196] |
| 35 | Gold QDs | – | – | 4.34 | – | Decreased self renewal ability of glioma stem-like cells. | [197] |
| 36 | ZnS QDs | – | Carboxymethylcellulose-QD conjugate | 3.3–4.5 | 400–430 (at pH 3.5). | Nanoprobe for bioimaging of glioblastoma cells with better cellular uptake. | [198] |
| 37 | Gold QDs | – | Gold QDs in combination with cold atmospheric plasma | 4–5 | – | Co-treatment induced reactive oxygen/nitrogen species suppressed the growth of U87 and U373 cells. | [199] |
| 38 | Graphene QDs | Dimethyl formamide | Doxorubicin-loaded grapheme QDs | $<$10 | 420 (when excited at 320 nm), 530 (when excited at 420 nm) | Synergistic cytotoxicity with doxorubicin in U87 cells. | [200] |
| 39 | CdSe/CdS core-shell QDs with | – | – | – | – | | [167] |

(*continued on next page*)

**Table 2** (*continued*)

| Sl No | QD types | Bioconjugation/ attachments | Delivery cargos | Size ranges (nm) | Photoluminescence emission peak (nm) | Applications | References |
|---|---|---|---|---|---|---|---|
| | proper passivation | | | | | Imaging of single dopamine transporter in acute brain slices with high precision. | |
| 40 | Functionalized grapheme QDs | – | Functionalized graphene QDs | <10 | 450 (when excited at 320 nm) | Downregulation of neurosphere formation, increased membrane fluidity. | [201] |
| 41 | Graphene QDs | PEG | PEGylated biocompatible graphene QDs | – | – | Nanoparticle meadiated radio-sensitization and anti-metastatic property against U87 and T98G cells. | [202] |
| 42 | Black phosphorus QDs | $Nd^{3+}$ion | $Nd^{3+}$ co-ordinated black phosphorus QDs | 3–20 | 1050 | Impressive monitoring of glioblastoma growth along with synergistic photodynamic chemotherapy. | [203] |
| 43 | Graphene QDs | – | Graphene QDs and doxorubicin encapsulated within homotypic cancer cell membrane | 5.8 ± 2 | 520 | Efficient targeting into brain along with promising chemo-photothermal treatment of glioma. | [204] |
| 44 | Graphene QDs | Nitrogen | Nitrogen-doped grapheme QDs | 3 | 420 (when excited at 350 nm), 480 (when excited at 380 nm) | Induction of ferroptosis and inflammation mediated cell death through disruption of $Ca^{2+}$ homoeostasis in mouse hippocampus, and microglia *in vitro*. | [205] |

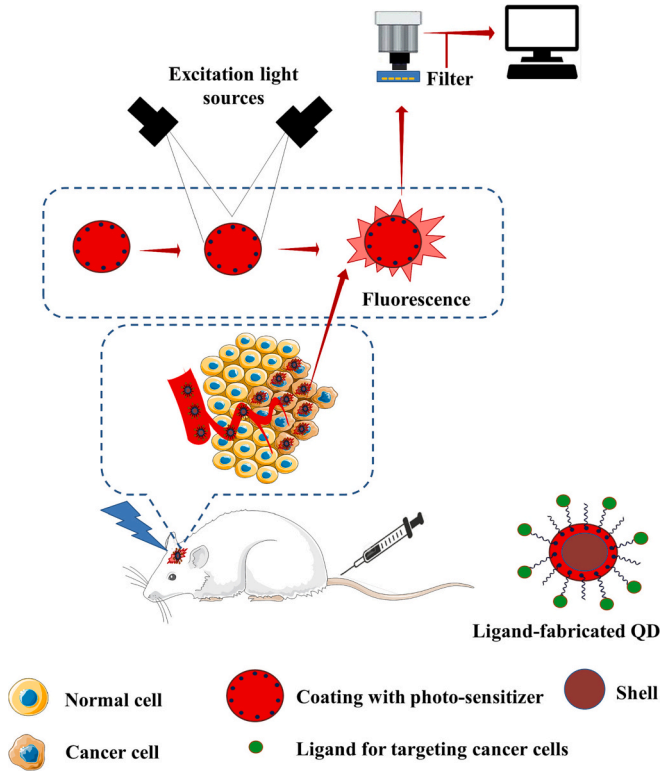



**Fig. 3.** Schematic representation of fluorescent QDs being used in brain imaging for visualization of cancer cells.

### 5.2. Therapeutic applications

QDs are appealing drug delivery methods for delivering therapeutic agents to brain cancer cells beyond brain barriers. To aid glioma-specific targeting, QDs can be bioconjugated with various functional moieties. Targeting ligands, such as various peptides, antibodies, nucleic acids, aptamers, and different small molecule ligands can be conjugated to the surface of QDs to aid affinity and targeted delivery into the brain tumor cells (Fig. 4). Distinct moieties present on the tumor cell surface are targeted to achieve target specificity.

Transferrin receptors have been regarded to be the potential target to deliver the cargo not only across the BBB but also into brain cancer cells [83]. Transferrin-conjugated docetaxel and QD-coloaded nanoliposomes improved brain-specific targeting to a significant extent. In addition to successfully adopting minimally invasive imaging and therapy, this formulation improved the drug release duration as compared to non-targeted docetaxel and QD-coloaded nanoliposomes [83]. Interestingly, both targeted and non-targeted docetaxel and QD-coloaded nanoliposomes showed greater effectiveness than the commercial formulation Docel™. Similarly, folate receptor recognition can also be subjected to active targeting in cancer theranostic [215]. In this aspect, a clear understanding of the receptor expression on normal brain cells is critical to avoid chemotherapeutic-induced toxicity, as well as the saturation of accessible receptors with respect to the metabolic needs of normal brain cells.

Graphene QDs are widely attempted nowadays for drug delivery to the brain since these biocompatible QDs can freely pass through the BBB [200]. Graphene QDs demonstrated a concentration-dependent cytotoxic effect against human glioblastoma cells without imparting cytotoxicity toward primary cortical neurons. Aiming at the antitumorigenic effect of dimethyl formamide, graphene QDs has been functionalized by conjugating with dimethyl formamide to achieve a better therapeutic effect [216]. However, the functionalization reduced the biocompatibility of graphene QDs and the functionalized formulation demonstrated cytotoxic effects on neuronal cells as well. Interestingly, graphene QDs, in combination with doxorubicin, were able to synergize the chemotherapeutic efficacy of doxorubicin against U87 cells. This synergy might be achieved by the increased membrane penetrability in a surface chemistry-dependent manner and better cellular uptake of doxorubicin by glioblastoma cells. The same research team experimentally reveals a positive association between the increased membrane penetrability and the net negative charge of graphene QDs. The higher communication between the membrane and negatively charged graphene QDs destabilizes cell-to-cell interactions, resulting in the suppression of neurosphere development. It has been hypothesized that the phase changes in the lipid bilayer of the plasma membrane of glioblastoma cells were related to the net negative surface charge of graphene QDs [201]. Graphene QDs undoubtedly show attractive promise in targeting brain

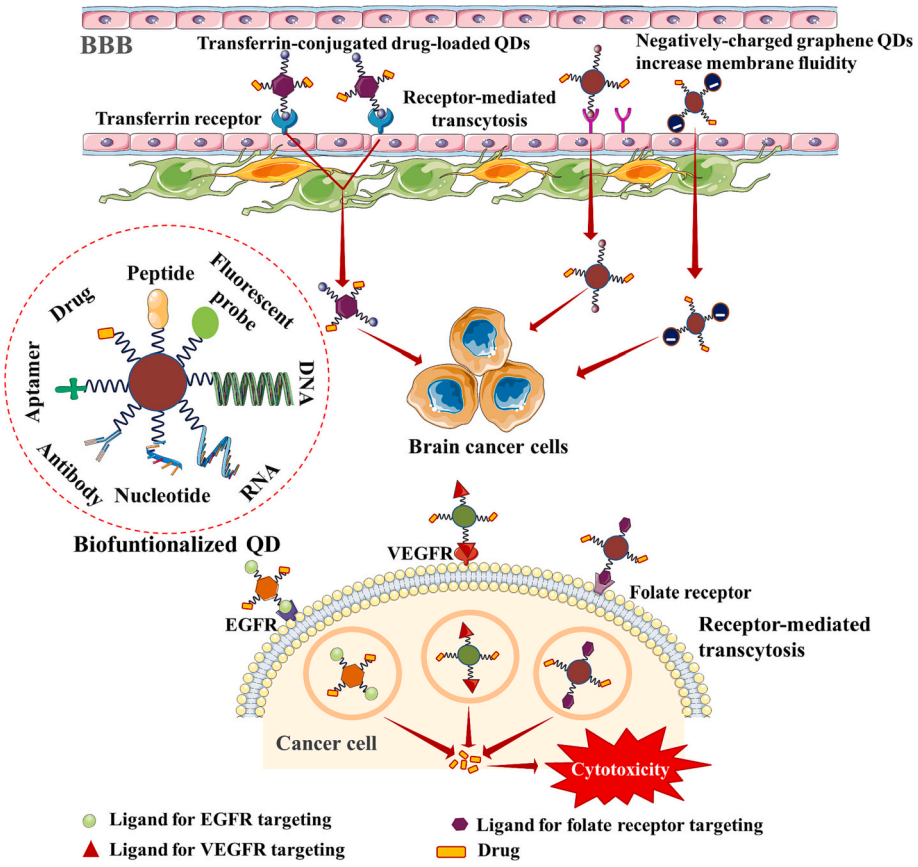



**Fig. 4.** Biofunctionalization of QDs aids drug delivery across the BBB to the target brain tumors through surface receptors.

tumors *in vitro*, but before further clinical translation, the findings must be verified through detailed *in vivo* studies. Recently, a versatile biomimetic nanostructure has been designed by encapsulating graphene QDs and doxorubicin within a homotypic cancer cell membrane for targeting glioma cells. This formulation takes advantage of homologous targeting of cancer cell membrane; however, requires an external stimulus to trigger the release of loaded drug. The coating within the cell membrane protects the drug from leakage [204]. This brain-targeted biomimetic nanoformulation grips the promise of efficient drug delivery to the target and offers an efficacious synergistic approach to precise chemo-photothermal therapy of glioma. Nitrogen-doped graphene QDs have emerged as a cost-effective and environment-friendly approach for a variety of outstanding attributes, including excellent biocompatibility, strong fluorescence, and resistance to photobleaching. Recently, nitrogen-doped graphene QDs have shown the potential to endorse multiple types of cell death, thus could aid in brain tumor treatment. Intranasal delivery of nitrogen-doped graphene QDs induced ferroptosis and inflammation-mediated cell death through disruption of $Ca^{2+}$ homeostasis and induction of ER stress in microglial cells of hippocampus in mice. A growing body of evidence reveals that ferroptosis, microglial activation, and inflammation in the hippocampus play a major role in the pathogenesis of several neurodegenerative disorders by declining neuronal function and impairing cognitive activities and thereby poses a warning toward biosafety regarding the clinical use of nitrogen-doped graphene QDs [205].

Overexpression of certain receptors on brain cancer cells can be successfully aimed to improve the target specificity of QDs [136,217]. Amplification of EGFR expression in different brain tumor cells, viz. high-grade astrocytomas, glioblastoma, etc. correlates with the progression of the disease. Interestingly, EGFR expression is absent in normal brain tissues. Thus, the conjugation of anti-EGFR antibody to QDs thus offers a unique possibility of fluorescence-guided diagnosis, intervention, or surgery of brain tumor [136]. Similarly, the VEGF receptor can also be targeted for the diagnosis and treatment of brain cancer. QDs conjugated with chitosan and a monoclonal antibody against the VEGF receptor displayed bifunctional roles uniting fluorescent bioimaging for diagnosis and a concentration-dependent cytotoxic effect on glioma cells *in vitro* [195]. The inclusion of chitosan improves the aqueous solubility of QD immunoconjugates. The selection of receptor(s) that are specifically expressed in brain cancer cells rather than normal cells is the main challenge with receptor-targeted formulations. Further research is required to address the risk-benefit ratio of biologicals in the formulation. Finally, detailed *in vivo* assays are required to substantiate the efficacy of receptor-targeted QDs in the active targeting of brain tumor within the whole system.

Encapsulation and surface modification do not interfere with the intrinsic properties of QDs at physiological levels [136]. PEGylation has been thought to lengthen circulation duration, increasing the likelihood of contact with the BBB. Upon radiation graphene QDs with a surface coated with PEG act as a prooxidant and generate ROS [218]. PEG-modified graphene QDs have been found to produce superoxide anion mediated through photoexcitation with blue light and decreasing cell viability of human glioma cells with increasing exposure time [175]. However, the generation of oxidative stress by these QDs poses questions about their potentially harmful effects on healthy cells.

A very interesting approach to fighting brain tumors is to halt the self-renewal ability of glioma-initiating cells. The multivalent gold QDs are biocompatible nanostructures that exhibit cytotoxicity to glioma spheroids at maximum concentration without imparting a cytotoxic effect on normal cells [197]. Interestingly, they demonstrated an excellent capacity to arrest the growth of cancer stem cell populations within glioma cells. The gold QDs negatively influenced the ability of glioblastoma spheroids to grow the population of glioma stem-like cells by inhibiting CD133, a marker protein of cancer stem-like cells. Moreover,

they are capable to limit EMT by abrogating catenin beta 1 (CTNNB1)/E-cadherin signaling in glioma spheroids. Thus, in brain cancer treatment, the gold QDs demonstrate a unique therapeutic attribute that abrogates the tumor recurrence by impeding the tumorigenic capacity of residual glioma stem-like cells, which are highly resistant to radiotherapy and chemotherapy. However, internalization into the cell through the plasma membrane is a potential challenge with gold QDs. Combinations of gold QDs and cold plasma nanoparticles have appeared to be a potential therapeutic approach, especially in the management of glioblastoma [199]. Cold plasma is thought to increase the permeability of the cell membrane provisionally to endorse endocytotic uptake of gold QDs. The co-treatment of gold QDs along with cold atmospheric plasma exhibited dual cytotoxicity to glioblastoma cells via activation of the extrinsic apoptotic pathways. In addition, both migration and self-renewal ability were significantly prevented by this combination therapy. The overall therapeutic effects of the co-treatment comprising gold QDs along with cold atmospheric plasma were superior to that of gold QDs alone. However, the application of a soft jet non-thermal plasma device to deliver cold plasma nanoparticles requires serious attention with respect to clinical perspectives. Though gold QDs by themselves or in conjunction with cold plasma nanoparticles exhibited therapeutic promise against brain cancer cells *in vitro*, more extensive research on clinically appropriate animal models and safety assessment must be undertaken before further clinical translation.

### 5.3. Miscellaneous

Theranostic nanomedicine is a unified nanotherapeutic platform whereby therapeutic and diagnostic functions are combined within a single system providing scopes for simultaneous detection, spatial targeting, and also tracking of the therapeutic response [129]. Advances in the synthesis of bioengineered QDs have produced multifunctional systems integrating diagnostic and therapeutic purposes including tumor targeting, bio-imaging, and drug delivery. In a recent development, neodymium (3+) ($Nd^{3+}$)-coordinated black phosphorus QDs demonstrated impressive monitoring of glioblastoma growth via intracranial second near-infrared fluorescence imaging along with synergistic photodynamic chemotherapy [203]. The ultrasmall structure and effective cargo-loading ability of $Nd^{3+}$-synchronized black phosphorus QDs enable them to cross the BBB easily to accurately monitor tumor growth by near-infrared fluorescence and suppress tumor growth by specific X-ray switching. The inclusion of this rare earth ion significantly improved the performance in fluorescence imaging and phototherapy are concerned, advancing from the recoverable near-infrared/X-ray optoelectronic swapping effects between black phosphorus QDs and $Nd^{3+}$. However, the $Nd^{3+}$-synchronized QDs may pose a risk of metal accumulation and associated adverse effects. Thus, the safety of this nanotheranostic agent is a key issue that requires proper justification before clinical application. Experimentally, QD-transferrin bioconjugation has achieved success to some extent to improve the internalization of QDs into human glioblastoma cells [185]. In addition, the recurrent glioblastoma cells more efficiently endorse transferrin receptor recycling than primary tumor-derived glioblastoma cells, which signifies the effective role of the biophotonic QDs against tumor recurrence. On the contrary, such targeting, due to its competitive nature might saturate available receptors affecting the transport of required substances for which, healthy cells might suffer, and normal brain function might be compromised [219]. A clinically comparable animal model would put some more light on the issue. Energy-driven, clathrin, and caveolin-mediated endocytotic pathways are very useful to cross the BBB [220]. Theranostic liposome-encapsulated QDs loaded with apomorphine can efficiently enter mouse brain cells through energy-dependent clathrin and caveolin-mediated endocytosis. The liposomal encapsulation not only improved fluorescence intensity in mouse brains but simultaneously enhance the BBB penetration over free QDs [176]. This drug-QD nanohybrid would offer a prospect of efficient diagnosis and treatment of brain disorders. The incorporation of QDs within liposomes significantly reduced uptake by the liver, thus allowing higher concentration in the brain. However, the synthesis of the cargo involves a complex multistep process that may be difficult to scale up. Knockdown of the *EGFR variant III* gene can result in the downregulation of PI3K-Akt signaling [221,222]. Multifunctional siRNA-QD conjugates functionalized with thiol-modified RGD and HIV-Tat peptides were used for the targeted delivery to glioblastoma cells and tracking of the efficacy of siRNA in silencing *EGFR variant III* gene via multimodal imaging probes [174]. Subsequent downregulation of PI3K-Akt signaling led to apoptosis of glioblastoma cells. RGD functionalization enables better uptake by glioblastoma cells and aids in fluorescence detection; while, HIV-Tat modification endorses effective transfection in glioblastoma cells. To achieve the effective knockdown of the target gene, a ratio of 1:10 for both QD to siRNA and siRNA to each peptide has been considered optimum. The functionalization and surface modification of QDs with specific ligands and functional biomolecules are key aspects in formulating this multifunctional siRNA-QD-based delivery system with desired theranostic effects. Furthermore, the promising *in vitro* results need to be followed up with suitable *in vivo* assays to ensure the theranostic efficacy of such multi-conjugated systems in the actual pathological milieu. A biotinylated aptamer was conjugated with multifunctional PEGylated streptavidin-linked QDs to preferentially bind with EGFR variant III on the surface of glioblastoma cells and to demonstrate strong fluorescent signals [189]. This biotin-aptamer-QD nanoprobe is a biocompatible and non-toxic formulation that is capable to cross the BBB and selectively accumulates on the tumor via EGFR variant III binding igniting possibilities of fluorescence-guided surgery to allow accurate delineation, and resection of glioblastoma cells. However, stability and structure modification of nucleic acid aptamers are key issues in designing an aptamer-based targeted formulation [223]. Conjugation of ligands to QDs can immensely improve targeting efficiency. Bioconjugation of a glioma-specific ligand, angio-pep2 to PEGylated carbon QDs can successfully target glioma cells with high specificity as compared to normal cells mediated by endosomes. PEG-modified angio-pep2-tailored carbon nanorods exhibit good signs of serum stability and biocompatibility leading to the possibility of a non-invasive tool for imaging with specific targeting [180]. However, ultrasmall particles pose a high possibility of rapid renal clearance, as described by others [224,225]. A growing body of evidence shows that superparamagnetic nanoparticles can serve as multipurpose agents to obtain magnetic resonance images of glioma via magnetic targeting because of their ability to contrast negatively with good biocompatibility. Contextually, magnetic targeting of glioma cells by conjugated QDs has displayed experimental success in guiding surgical resection via multimodal imaging [192]. Encapsulation of superparamagnetic iron oxide nanoparticles, QDs, and cilengitide within a theranostic nanoliposome not only facilitates *in vivo* dual-imaging (computed tomography and magnetic resonance imaging) during surgery but also inhibits tumor by the action of cilengitide under magnetic targeting. Cilengitide (CGT) is a cyclic arginine-glycine-aspartic acid pentapeptide antagonist that suppresses integrin receptor activation on glioblastoma and glioblastoma-associated endothelial cells and thus arrests tumor growth and angiogenesis. The inclusion of arginine-glycine-aspartic acid sequence within theranostic liposomes containing both QDs and docetaxel has also achieved success as a brain-targeted theranostic agent [187]. In normal rats, this biocompatible nanoplatform appears to be very promising for brain theranostic, but it is important to assess their theranostic potential in brain tumor-bearing animals before further clinical translation.

The aqueous solubility of QDs on many occasions needs to be enhanced through some kind of surface functionalization. Modification of QDs with biopolymers like carboxymethyl cellulose improves the hydrophilicity of QDs [194]. A QD-biopolymer-doxorubicin nanohybrid exhibit an intriguing promise against glioma for theranostic purposes including bioimaging, targeting, and chemotherapeutic drug delivery

[194]. This biocompatible supramolecular nanoconjugate acts as an innovative fluorescent nanoprobe, which can simultaneously deliver a chemotherapeutic drug to the brain cancer cells in a sustained manner. A facile and environment-friendly synthesis aid attention to this formulation for further clinical applications. Similarly, dual functionality (nanoprobe and nanocarrier) has been achieved with drug-loaded, biofunctionalized carbon QDs against human glioma cells [196]. Nitrogen-doped carbon QDs with high QY can cross the BBB in a concentration-dependent manner, paving a new way for real-time live-cell imaging and a trackable delivery mechanism into the brain [186]. Further establishing the superiority of organic QDs, biocompatible graphene QDs displayed promising radiosensitization and anti-metastatic behaviour with U87 and T98G brain cancer cells [202]. Ionizing radiation contributed significantly to the extent of attachment and migration between irradiated and non-irradiated T98G cells. The expression of aspartate β-hydroxylase can serve as a potential therapeutic target in malignant gliomas [226]. In an attempt at glioma-specific targeting, aspartate-loaded carbon QDs demonstrated efficient BBB penetration and excellent targeting of glioma tissue [184]. The successful targeting efficiency corresponded to success in noninvasive bioimaging of glioma in mice. Similarly, fluorescent carbon QDs prepared by conjugation with l-aspartic acid and D-glucose at a molar ratio of 3:7 exhibit excellent targeting of C6 glioma cells [191]. Conjugation of a chemotherapeutic drug to these aspartate-conjugated carbon QDs could corroborate a therapeutic option against brain tumors. Excellent biocompatibility, and targeting ability without complicated tailoring, lifts the interest in these carbon QDs to several folds. Though promising, the research should be continued with respect to clinically similar animal models before further clinical translation.

## 6. Future prospects

Owing to the highly selective semipermeable property of the BBB, most drugs are restricted from entering the brain. QDs have made significant progress in overcoming hurdles to brain targeting in recent years. For better functionalization, QDs can be conjugated with a variety of ligands. Current research focuses on developing more luminescent hydrophilic QDs in order to meet the demand for higher efficiency and better fluorescence with minimum untoward effects [207]. Biocompatibility, stability, and aqueous solubility can be improved by virtue of capping, solubilization, and functionalization techniques [227]. QDs present with unique optoelectrical properties compared to conventional dyes for better imaging and detection [17]. The fact that QDs can be directly tracked through intrinsic fluorescence makes them superior candidates for diagnostic studies. QDs have come up with exceptional promises to detect tumor boundaries preoperatively [228]. Even in therapeutics, QDs have shown exciting promise. Specific targeting can be achieved by virtue of functionalization. Novel futuristic findings that have been discussed earlier may open up a new horizon in brain tumor therapeutics by suppressing tumor-initiating cells with gold QDs [197,199]. Graphene QDs, in combination with doxorubicin, could lead to a better chemotherapeutic potential at a low dose, which simultaneously lower doxorubicin-associated side effects [200].

Clinical usage of semiconductor QDs is still limited due to their slow and incomplete elimination, which could leave chances of potential toxic manifestations. Bioaccumulation-associated toxicity continues to be a serious issue that limits the large-scale clinical usage of QDs. Another point of concern is regarding some of the exceptionally small size of QDs; as can be seen from Table 2, few studies discussed earlier have reported particle sizes <5 nm, which are potential subjects for rapid renal excretion with possibilities of compromise in activity profile. Site-specific targeting exhibits promise in this regard, further translational insight is awaited. Furthermore, batch-to-batch uniformity in terms of optical characteristics, as well as cost-effective scale-up of QD production, are issues that must be addressed before the rampant clinical use of QDs in the future [65]. However, carbon QDs and graphene QDs are clearly gaining popularity, probably due to the futuristic promises they make in terms of water solubility and biocompatibility when compared to metal-based QDs. Multifunctional QDs are showing promise in the detection and therapy of brain cancer. However, QD-based theranostics are still in their infancy. The experimental success of QDs is yet to be completely complemented by translational research [229]. Mechanisms of barrier penetration by QDs still need investigations for better understanding [230]. Reproducibility in drug loading capacity and release behaviour has to be specifically revealed before moving on to the clinical stage with QDs [231]. Most of the reported QDs emit fluorescence in short-wavelength, which makes it difficult to track the loaded drug/s, especially *in vivo* [232]. Long-emissive carbon dots might offer a solution to this problem. Brain tumor-specific nanotheranostics could probably gain a lot through QD-nanozyme complex systems, however, rigorous investigations are still awaited [233].

## 7. Conclusion

QDs have already established their significance in the developing disciplines of nanomedicine and nanobiotechnology due to their unique properties. They can fit in to meet the need for nanosized carrier vehicles that can safely deliver drugs to the brain through systemic microvascular beds found in blood capillaries and endothelial cells. Already they have shown futuristic promises toward cellular imaging and targeted therapeutics. The emergence of carbon and graphene-based QDs has remarkably improved the scenario regarding biocompatibility and toxicity aspects. The mechanism of photodynamics has the prospect to initiate a new era of non-invasive bio-imaging, especially in brain regions as far as clarity is concerned. Widespread clinical use of QDs is on the cards. Newer developments like organic QDs and magnetic QDs may highlight new facets of targeted therapy of brain tumors with low toxicity. QDs also hold promise in the field of personalized treatment. As a whole, current research hints toward a huge potential of QDs in diagnostics and targeted therapy of brain cancers.

## Author contributions

**Saikat Dewanjee** and **Saurabh Kumar Jha** have prepared the backbone of the manuscript. **Pratik Chakraborty, Sabya Sachi Das, Apala Chakraborty, and Chiranjib Bhattacharyya** wrote the original draft of the manuscript with **Abhijit Dey, Ramesh Kandimalla, Biswajit Mukherjee**, **Abilash Valsala Gopalakrishnan**, **Sandeep Kumar Singh, Shubham Kant, Parma Nand,** and **Niraj Kumar Jha** refined the first draft. **Saikat Dewanjee**, **Pratik Chakraborty,** and **Saurabh Kumar Jha** critically revised the manuscript for intellectually correct content. **Saikat Dewanjee** and **Pratik Chakraborty** contributed to the figures and answered the reviewer comments. All authors approved the submitted version.

## Declaration of Competing Interest

None.

## Acknowledgements

The authors are thankful to the Council of Scientific and Industrial Research, India for awarding a research project (grant number 02 (0275)/16/EMR-II dated 06.12.2016) to Saikat Dewanjee. Jadavpur University, India is acknowledged for awarding the State Government Research Fellowship to PC (Ref no: R-11/13/20 dated 22.01.2020). Saikat Dewanjee also wants to acknowledge his colleague, Dr. Ketousetuo Kuotsu, Associate Professor, Department of Pharmaceutical Technology, Jadavpur University, India to give the idea to write a review on this topic. Finally, the authors would like to express their gratitude to the editor and reviewers for their insightful comments and

criticism, which contributed to improving the scientific calibre of this review.

## References


[1] J.J. Zhu, J.J. Li, H.P. Huang, F.F. Cheng (Eds.), Quantum Dots for DNA Biosensing, Springer Nature, Switzerland, 2013.

[2] W.R. Algar, M. Massey, K. Rees, R. Higgins, K.D. Krause, G.H. Darwish, W. J. Peveler, Z. Xiao, H.Y. Tsai, R. Gupta, K. Lix, Photoluminescent nanoparticles for chemical and biological analysis and imaging, Chem. Rev. 121 (2021) 9243–9358, https://doi.org/10.1021/acs.chemrev.0c01176.

[3] A. Valizadeh, H. Mikaeili, M. Samiei, S.M. Farkhani, N. Zarghami, A. Akbarzadeh, S. Davaran, Quantum dots: synthesis, bioapplications, and toxicity, Nanoscale Res. Lett. 7 (2012) 1–14, https://doi.org/10.1186/1556-276X-7-480.

[4] N.S. Dey, M.B. Rao, Quantum dot: novel carrier for drug delivery, Int. J. Res. Pharmaceut. Biomed. Sci. 2 (2011) 448–458.

[5] V. Gubala, G. Giovannini, F. Kunc, M.P. Monopoli, C.J. Moore, Dye-doped silica nanoparticles: synthesis, surface chemistry and bioapplications, cancer, Nanotechnology 11 (2020) 1–43, https://doi.org/10.1186/s12645-019-0056-x.

[6] M. Fang, C.W. Peng, D.W. Pang, Y. Li, Quantum dots for cancer research: current status, remaining issues, and future perspectives, Cancer Biol. Med. 9 (2012) 151, https://doi.org/10.7497/j.issn.2095-3941.2012.03.001.

[7] N.L. Goerig, B. Frey, K. Korn, B. Fleckenstein, K. Überla, M.A. Schmidt, A. Dörfler, T. Engelhorn, I. Eyüpoglu, P.F. Rühle, F. Putz, Early mortality of brain cancer patients and its connection to cytomegalovirus reactivation during radiochemotherapy, Clin. Cancer Res. 26 (2010) 3259–3270, https://doi.org/10.1158/1078-0432.CCR-19-3195.

[8] M.G. Baratta, Getting to the brain, Nat. Nanotechnol. 13 (2018) 536, https://doi.org/10.1038/s41565-018-0182-3.

[9] S. Ghaderi, B. Ramesh, A.M. Seifalian, Fluorescence nanoparticles "quantum dots" as drug delivery system and their toxicity: a review, J. Drug Target. 19 (2011) 475–486, https://doi.org/10.3109/1061186X.2010.526227.

[10] A.M. Wagner, J.M. Knipe, G. Orive, N.A. Peppas, Quantum dots in biomedical applications, Acta Biomater. 94 (2019) 44–63, https://doi.org/10.1016/j.actbio.2019.05.022.

[11] M.K. Wagner, F. Li, J. Li, X.F. Li, X.C. Le, Use of quantum dots in the development of assays for cancer biomarkers, Anal. Bioanal. Chem. 397 (2010) 3213–3224, https://doi.org/10.1007/s00216-010-3847-9.

[12] H.S. Mansur, Quantum dots and nanocomposites, Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol. 2 (2010) 113–129, https://doi.org/10.1002/wnan.78.

[13] U. Badilli, F. Mollarasouli, N.K. Bakirhan, Y. Ozkan, S.A. Ozkan, S.A., Role of quantum dots in pharmaceutical and biomedical analysis, and its application in drug delivery, TrAC-Trend. Anal. Chem. 131 (2020) (2020), 116013, https://doi.org/10.1016/j.trac.2020.116013.

[14] P. Wu, X.P. Yan, Doped quantum dots for chemo/biosensing and bioimaging, Chem. Soc. Rev. 42 (2013) 5489–5521, https://doi.org/10.1039/C3CS60017C.

[15] X. Xu, R. Ray, Y. Gu, H.J. Ploehn, L. Gearheart, K. Raker, W.A. Scrivens, Electrophoretic analysis and purification of fluorescent single-walled carbon nanotube fragments, J. Am. Chem. Soc. 126 (2004) 12736–12737, https://doi.org/10.1021/ja040082h.

[16] Y. Chong, Y. Ma, H. Shen, X. Tu, X. Zhou, J. Xu, J. Dai, S. Fan, Z. Zhang, The in vitro and in vivo toxicity of graphene quantum dots, Biomaterials. 35 (2014) 5041–5048, https://doi.org/10.1016/j.biomaterials.2014.03.021.

[17] M. Díaz-González, A. de la Escosura-Muñiz, M.T. Fernandez-Argüelles, F.J. G. Alonso, J.M. Costa-Fernandez, Quantum dot bioconjugates for diagnostic applications, Top. Curr. Chem. 378 (2020) 1–44, https://doi.org/10.1007/s41061-020-0296-6.

[18] R. Mohammadi, H. Naderi-Manesh, L. Farzin, Z. Vaezi, N. Ayarri, L. Samandari, M. Shamsipur, Fluorescence sensing and imaging with carbon-based quantum dots for early diagnosis of cancer: A review, J. Pharm. Biomed. Anal. (2022), 114628, https://doi.org/10.1016/j.jpba.2022.114628.

[19] T. Yuan, T. Meng, P. He, Y. Shi, Y. Li, X. Li, L. Fan, S. Yang, Carbon quantum dots: an emerging material for optoelectronic applications, J. Mater. Chem. C 7 (2019) 6820–6835, https://doi.org/10.1039/C9TC01730E.

[20] N. Bajwa, N.K. Mehra, K. Jain, N.K. Jain, Pharmaceutical and biomedical applications of quantum dots, Artif. Cells Nanomed. Biotechnol. 44 (2016) 758–768, https://doi.org/10.3109/21691401.2015.1052468.

[21] G. De Crozals, R. Bonnet, C. Farre, C. Chaix, Nanoparticles with multiple properties for biomedical applications: A strategic guide, Nano Today 11 (2016) 435–463, https://doi.org/10.1016/j.nantod.2016.07.002.

[22] S. Mazumder, R. Dey, M.K. Mitra, S. Mukherjee, G.C. Das, Review: biofunctionalized quantum dots in biology and medicine, J. Nanomater. 2009 (2009) 1–17, https://doi.org/10.1155/2009/815734.

[23] A. Banerjee, T. Pons, N. Lequeux, B. Dubertret, Quantum dots-DNA bioconjugates: synthesis to applications, Interface Focus. 6 (2016) 20160064, https://doi.org/10.1098/rsfs.2016.0064.

[24] G. Aizik, N. Waiskopf, M. Agbaria, M. Ben-David-Naim, Y. Levi-Kalisman, A. Shahar, U. Banin, G. Golomb, Liposomes of quantum dots configured for passive and active delivery to tumor tissue, Nano Lett. 19 (2019) 5844–5852, https://doi.org/10.1021/acs.nanolett.9b01027.

[25] G. Pereira, C.A. Monteiro, G.M. Albuquerque, M.I. Pereira, M.P. Cabrera, P. E. Cabral Filho, G.A. Pereira, A. Fontesa, B.S. Santos, (Bio) conjugation strategies applied to fluorescent semiconductor quantum dots, J. Braz. Chem. Soc. 30 (2019) 2536–2561, https://doi.org/10.21577/0103-5053.20190163.

[26] A. Wadhawan, M. Chatterjee, G. Singh, Present scenario of bioconjugates in cancer therapy: A review, Int. J. Mol. Sci. 20 (2019) 5243, https://doi.org/10.3390/ijms20215243.

[27] G. Vampa, T.J. Hammond, N. Thiré, B.E. Schmidt, F. Légaré, C.R. McDonald, T. Brabec, D.D. Klug, P.B. Corkum, All-optical reconstruction of crystal band structure, Phys. Rev. Lett. 115 (2015), 193603, https://doi.org/10.1103/PhysRevLett.115.193603.

[28] M. Galliani, G. Signore, Poly (lactide-co-glycolide) nanoparticles co-loaded with chlorophyllin and quantum dots as photodynamic therapy agents, Chempluschem 84 (2019) 1653–1658, https://doi.org/10.1002/cplu.201900342.

[29] A.V. Krasheninnikov, K. Nordlund, Ion and electron irradiation-induced effects in nanostructured materials, J. Appl. Phys. 107 (2010) 3, https://doi.org/10.1063/1.3318261.

[30] M.C. Dos Santos, W.R. Algar, I.L. Medintz, N. Hildebrandt, Quantum dots for Förster resonance energy transfer (FRET), TrAC-Trend. Anal. Chem. 125 (2020), 115819, https://doi.org/10.1016/j.trac.2020.115819.

[31] O.A. Goryacheva, N.V. Beloglazova, A.M. Vostrikova, M.V. Pozharov, A. M. Sobolev, I.Y. Goryacheva, Lanthanide-to-quantum dot Förster resonance energy transfer (FRET): application for immunoassay, Talanta 164 (2017) 377–385, https://doi.org/10.1016/j.talanta.2016.11.054.

[32] A. Anas, H. Akita, H. Harashima, T. Itoh, M. Ishikawa, V. Biju, Photosensitized breakage and damage of DNA by CdSe− ZnS quantum dots, J. Phys. Chem. B 112 (2008) 10005–10011, https://doi.org/10.1021/jp8018606.

[33] J. Lovrić, H.S. Bazzi, Y. Cuie, G.R. Fortin, F.M. Winnik, D. Maysinger, Differences in subcellular distribution and toxicity of green and red emitting CdTe quantum dots, J. Mol. Med. 83 (2005) 377–385, https://doi.org/10.1007/s00109-004-0629-x.

[34] S.J. Cho, D. Maysinger, M. Jain, B. Röder, S. Hackbarth, F.M. Winnik, Long-term exposure to CdTe quantum dots causes functional impairments in live cells, Langmuir. 23 (2007) 1974–1980, https://doi.org/10.1021/la060093j.

[35] V. Biju, S. Mundayoor, R.V. Omkumar, A. Anas, M. Ishikawa, Bioconjugated quantum dots for cancer research: present status, prospects and remaining issues, Biotechnol. Adv. 28 (2010) 199–213, https://doi.org/10.1016/j.biotechadv.2009.11.007.

[36] M. Hao, R. Liu, Molecular mechanism of CAT and SOD activity change under MPA-CdTe quantum dots induced oxidative stress in the mouse primary hepatocytes, Spectrochim. Acta A Mol. Biomol. Spectrosc. 220 (2019), 117104, https://doi.org/10.1016/j.saa.2019.05.009.

[37] K.J. McHugh, L. Jing, A.M. Behrens, S. Jayawardena, W. Tang, M. Gao, R. Langer, A. Jaklenec, Biocompatible semiconductor quantum dots as cancer imaging agents, Adv. Mater. 30 (2018) 1706356, https://doi.org/10.1002/adma.201706356.

[38] C.M. Nastiti, Y. Mohammed, K.C. Telaprolu, X. Liang, J.E. Grice, M.S. Roberts, H. A. Benson, Evaluation of quantum dot skin penetration in porcine skin: effect of age and anatomical site of topical application, Skin Pharmacol. Physiol. 32 (2019) 1–10, https://doi.org/10.1159/000499435.

[39] H. Huang, J.F. Lovell, Advanced functional nanomaterials for theranostics, Adv. Funct. Mater. 27 (2017) 1603524, https://doi.org/10.1002/adfm.201603524.

[40] M.Z. Zhang, R.N. Yu, J. Chen, Z.Y. Ma, Y.D. Zhao, Targeted quantum dots fluorescence probes functionalized with aptamer and peptide for transferrin receptor on tumor cells, Nanotechnol. 23 (2012), 485104, https://doi.org/10.1088/0957-4484/23/48/485104.

[41] S. Paris-Robidas, D. Brouard, V. Emond, M. Parent, F. Calon, Internalization of targeted quantum dots by brain capillary endothelial cells in vivo, J. Cereb. Blood Flow Metab. 36 (2016) 731–742, https://doi.org/10.1177/0271678X15608201.

[42] M.X. Zhao, B.J. Zhu, The research and applications of quantum dots as nano-carriers for targeted drug delivery and cancer therapy, Nanoscale Res. Lett. 11 (2016) 1–9, https://doi.org/10.1186/s11671-016-1394-9.

[43] L. Qi, T. Pan, L. Ou, Z. Ye, C. Yu, B. Bao, Z. Wu, D. Cao, L. Dai, Biocompatible nucleus-targeted graphene quantum dots for selective killing of cancer cells via DNA damage, Commun. Biol. 4 (2021) 1–12, https://doi.org/10.1038/s42003-021-01713-1.

[44] B. Ma, C. Martín, R. Kurapati, A. Bianco, Degradation-by-design: how chemical functionalization enhances the biodegradability and safety of 2D materials, Chem. Soc. Rev. 49 (2020) 6224–6247, https://doi.org/10.1039/C9CS00822E.

[45] K. Nienhaus, H. Wang, G.U. Nienhaus, Nanoparticles for biomedical applications: exploring and exploiting molecular interactions at the nano-bio interface, Mater. Today Adv. 5 (2020), 100036, https://doi.org/10.1016/j.mtadv.2019.100036.

[46] G. Yang, S.Z.F. Phua, A.K. Bindra, Y. Zhao, Degradability and clearance of inorganic nanoparticles for biomedical applications, Adv. Mater. 31 (2019) 1805730, https://doi.org/10.1002/adma.201805730.

[47] J. Bourquin, A. Milosevic, D. Hauser, R. Lehner, F. Blank, A. Petri-Fink, B. Rothen-Rutishauser, Biodistribution, clearance, and long-term fate of clinically relevant nanomaterials, Adv. Mater. 30 (2018) 1704307, https://doi.org/10.1002/adma.201704307.

[48] H.S. Choi, W. Liu, P. Misra, E. Tanaka, J.P. Zimmer, B. Itty Ipe, M.G. Bawendi, J. V. Frangioni, Renal clearance of quantum dots, Nat. Biotechnol. 25 (2007) 1165–1170, https://doi.org/10.1038/nbt1340.

[49] V.G. Reshma, P.V. Mohanan, Quantum dots: applications and safety consequences, J. Lumin. 205 (2019) 287–298, https://doi.org/10.1016/j.jlumin.2018.09.015.

[50] K.T. Yong, W.C. Law, R. Hu, L. Ye, L. Liu, M.T. Swihart, P.N. Prasad, Nanotoxicity assessment of quantum dots: from cellular to primate studies, Chem. Soc. Rev. 42 (2013) 1236–1250, https://doi.org/10.1039/c2cs35392j.

[51] B. Gidwani, V. Sahu, S.S. Shukla, R. Pandey, V. Joshi, V.K. Jain, A. Vyas, Quantum dots: prospectives, toxicity, advances and applications, J. Drug Deliv.

Sci. Technol. 61 (2010), 102308, https://doi.org/10.3109/1061186X.2010.526227.
[52] A. Solanki, J.D. Kim, K.B. Lee, Nanotechnology for regenerative medicine: nanomaterials for stem cell imaging, Nanomed. 3 (2008) 567–578, https://doi.org/10.2217/17435889.3.4.567.
[53] K.M. Tsoi, Q. Dai, B.A. Alman, W.C. Chan, Are quantum dots toxic? Exploring the discrepancy between cell culture and animal studies, Acc. Chem. Res. 46 (2013) 662–671, https://doi.org/10.1021/ar300040z.
[54] M.S. Stan, I. Memet, C. Sima, T. Popescu, V.S. Teodorescu, A. Hermenean, A. Dinischiotu, Si/SiO2 quantum dots cause cytotoxicity in lung cells through redox homeostasis imbalance, Chem. Biol. Interact. 220 (2014) 102–115, https://doi.org/10.1016/j.cbi.2014.06.020.
[55] C. Ren, X. Hu, Q. Zhou, Graphene oxide quantum dots reduce oxidative stress and inhibit neurotoxicity in vitro and in vivo through catalase-like activity and metabolic regulation, Adv. Sci. 5 (2018) 1700595, https://doi.org/10.1002/advs.201700595.
[56] C. Zhang, X. Wang, J. Du, Z. Gu, Y. Yong, Reactive oxygen species-regulating strategies based on nanomaterials for disease treatment, Adv. Sci. 8 (2021) 2002797, https://doi.org/10.1002/advs.202002797.
[57] R. Bjorkland, D. Tobias, E.J. Petersen, Increasing evidence indicates low bioaccumulation of carbon nanotubes, Environ. Sci. Nano 4 (2017) 747–766, https://doi.org/10.1039/C6EN00389C.
[58] S. Iravani, R.S. Varma, Green synthesis, biomedical and biotechnological applications of carbon and graphene quantum dots. A review, Environ. Chem. Lett. 18 (2020) 703–727, https://doi.org/10.1007/s10311-020-00984-0.
[59] T. Guo, Y. Wu, Y. Lin, X. Xu, H. Lian, G. Huang, J.Z. Liu, X. Wu, H.H. Yang, Black phosphorus quantum dots with renal clearance property for efficient photodynamic therapy, Small 14 (2018) 1702815, https://doi.org/10.1002/smll.201702815.
[60] L.J. Desmond, A.N. Phan, G. Piergiorgio, Critical overview on the green synthesis of carbon quantum dots and their application for cancer therapy, Environ. Sci. Nano 8 (2021) 848–862, https://doi.org/10.1039/D1EN00017A.
[61] C. Hadad, J.M. González-Domínguez, S. Armelloni, D. Mattinzoli, M. Ikehata, A. Istif, A. Ostric, F. Cellesi, C.M. Alfieri, P. Messa, B. Ballesteros, Graphene quantum dots: from efficient preparation to safe renal excretion, Nano Res. 14 (2021) 674–683, https://doi.org/10.1007/s12274-020-3096-y.
[62] S.K. Tripathi, G. Kaur, R.K. Khurana, S. Kapoor, B. Singh, Quantum dots and their potential role in cancer theranostics, Crit. Rev. Ther. Drug Carrier Syst. 32 (2015) 461–502, https://doi.org/10.1615/critrevtherdrugcarriersyst.2015012360.
[63] C.D. Vo, H. Pang, I. Naasani, Nanoco Technologies Ltd, Surface-modified nanoparticles, U.S. Patent 10 (2019) 329,479.
[64] D.D.N. Marreiro, K.J.C. Cruz, J.B.S. Morais, J.B. Beserra, J.S. Severo, A.R.S. De Oliveira, Zinc and oxidative stress: current mechanisms, Antioxidants. 6 (2017) 24, https://doi.org/10.3390/antiox6020024.
[65] S. Filali, F. Pirot, P. Miossec, Biological applications and toxicity minimization of semiconductor quantum dots, Trends Biotechnol. 38 (2020) 163–177, https://doi.org/10.1016/j.tibtech.2019.07.013.
[66] R. Hardman, A toxicologic review of quantum dots: toxicity depends on physicochemical and environmental factors, Environ. Health Perspect. 114 (2006) 165–172, https://doi.org/10.1289/ehp.8284.
[67] X. Lou, Q.P. Gui, L. Sun, N.Z. Wu, J.H. Lyu, L. Ma, Comparisons of MR findings between supratentorial and infratentorial gangliogliomas, Clin. Neuroradiol. 26 (2016) 65–71, https://doi.org/10.1007/s00062-014-0333-3.
[68] B.T. Whitfield, J.T. Huse, Classification of adult-type diffuse gliomas: Impact of the World Health Organization 2021 Update, Brain Pathol. (2022), e13062, https://doi.org/10.1111/bpa.13062.
[69] Q.R. Lu, L. Qian, X. Zhou, Developmental origins and oncogenic pathways in malignant brain tumors, Wiley Interdiscip. Rev. Dev. Biol. 8 (2019), e342, https://doi.org/10.1002/wdev.342.
[70] M. Mendes, J.J. Sousa, A. Pais, C. Vitorino, Targeted theranostic nanoparticles for brain tumor treatment, Pharmaceutics. 10 (2018) 181, https://doi.org/10.3390/pharmaceutics10040181.
[71] D.D. Wu, Y.R. Gao, T. Li, D.Y. Wang, D. Lu, S.Y. Liu, Y. Hong, H.B. Ning, J.P. Liu, J. Shang, J.F. Shi, PEST-containing nuclear protein mediates the proliferation, migration, and invasion of human neuroblastoma cells through MAPK and PI3K/AKT/mTOR signaling pathways, BMC Cancer 18 (2018) 499, https://doi.org/10.1186/s12885-018-4391-9.
[72] M.J. Ramaiah, K.R. Kumar, mTOR-Rictor-EGFR axis in oncogenesis and diagnosis of glioblastoma multiforme, Mol. Biol. Rep. 48 (2021) 4813–4835, https://doi.org/10.1007/s11033-021-06462-2.
[73] M.M. Zarou, A. Vazquez, G. Vignir Helgason, Folate metabolism: a re-emerging therapeutic target in haematological cancers, Leukemia. 35 (2021) 1539–1551, https://doi.org/10.1038/s41375-021-01189-2.
[74] W. Szopa, T.A. Burley, G. Kramer-Marek, W. Kaspera, Diagnostic and therapeutic biomarkers in glioblastoma: current status and future perspectives, Biomed. Res. Int. 2017 (2017) 8013575, https://doi.org/10.1155/2017/8013575.
[75] T.C. Archer, S. Sengupta, S.L. Pomeroy, Brain cancer genomics and epigenomics, Handb. Clin. Neurol. 148 (2018) 785–797, https://doi.org/10.1016/B978-0-444-64076-5.00050-8.
[76] F. Seker-Polat, N. Pinarbasi Degirmenci, I. Solaroglu, T. Bagci-Onder, Tumor cell infiltration into the brain in glioblastoma: From mechanisms to clinical perspectives, Cancers 14 (2022) 443, https://doi.org/10.3390/cancers14020443.
[77] J.H. Sampson, M.D. Gunn, P.E. Fecci, D.M. Ashley, Brain immunology and immunotherapy in brain tumours, Nat. Rev. Cancer 20 (2020) 12–25, https://doi.org/10.1038/s41568-019-0224-7.
[78] B.S. Mahmoud, A.H. AlAmri, C. McConville, Polymeric nanoparticles for the treatment of malignant gliomas, Cancers. 12 (2020) 175, https://doi.org/10.3390/cancers12010175.
[79] A.R. Khan, X. Yang, M. Fu, G. Zhai, Recent progress of drug nanoformulations targeting to brain, J. Control. Release 291 (2018) 37–64, https://doi.org/10.1016/j.jconrel.2018.10.004.
[80] A. Elfadadny, H.M. El-Husseiny, A. Abugomaa, R.F. Ragab, E.A. Mady, M. Aboubakr, H. Samir, A.S. Mandour, A. El-Mleeh, A.H. El-Far, A. El-Aziz, Role of multidrug resistance-associated proteins in cancer therapeutics: past, present, and future perspectives, Environ. Sci. Pollut. Res. 28 (2021) 49447–49466, https://doi.org/10.1007/s11356-021-15759-5.
[81] S. Ogata, S. Ito, T. Masuda, S. Ohtsuki, Changes of blood-brain barrier and brain parenchymal protein expression levels of mice under different insulin-resistance conditions induced by high-fat diet, Pharm. Res. 36 (2019) 141, https://doi.org/10.1007/s11095-019-2674-8.
[82] C.P. Profaci, R.N. Munji, R.S. Pulido, R. Daneman, The blood–brain barrier in health and disease: important unanswered questions, J. Exp. Med. 217 (2020), e20190062, https://doi.org/10.1084/jem.2019006.
[83] R.P. Sonali, N. Singh, G. Singh, M.R. Sharma, B. Vijayakumar, S. Koch, U. Singh, D. Singh, B.L. Dash, M.S. Pandey, Muthu, Transferrin liposomes of docetaxel for brain-targeted cancer applications: formulation and brain theranostics, Drug Deliv. 23 (2016) 1261–1271, https://doi.org/10.3109/10717544.2016.1162878.
[84] S. Dhanda, R. Sandhir, Blood-brain barrier permeability is exacerbated in experimental model of hepatic encephalopathy via MMP-9 activation and downregulation of tight junction proteins, Mol. Neurobiol. 55 (2018) 3642–3659, https://doi.org/10.1007/s12035-017-0521-7.
[85] M.D. Sweeney, Z. Zhao, A. Montagne, A.R. Nelson, B.V. Zlokovic, Blood-brain barrier: from physiology to disease and back, Physiol. Rev. 99 (2019) 21–78, https://doi.org/10.1152/physrev.00050.2017.
[86] M. Loizidou, V. Sefcikova, J.O. Ekert, M. Bone, G. Samandouras, Reforming support systems of newly diagnosed brain cancer patients: a systematic review, J. Neuro-Oncol. 156 (2021) 61–71, https://doi.org/10.1007/s11060-021-03895-4.
[87] O. Van Tellingen, B. Yetkin-Arik, M.C. De Gooijer, P. Wesseling, T. Wurdinger, H. E. De Vries, Overcoming the blood–brain tumor barrier for effective glioblastoma treatment, Drug Resist. Updat. 19 (2015) 1–12, https://doi.org/10.1016/j.drup.2015.02.002.
[88] E. Belykh, K.V. Shaffer, C. Lin, V.A. Byvaltsev, M. Preul, L. Chen, Blood-brain barrier, blood-brain tumor barrier, and fluorescence-guided neurosurgical oncology: delivering optical labels to brain tumors, Front. Oncol. 10 (2020) 739, https://doi.org/10.3389/fonc.2020.00739.
[89] K. Jütten, L. Weninger, V. Mainz, S. Gauggel, F. Binkofski, M. Wiesmann, D. Merhof, H. Clusmann, C.H. Na, Dissociation of structural and functional connectomic coherence in glioma patients, Sci. Rep. 11 (2021) 1–12, https://doi.org/10.1038/s41598-021-95932-5.
[90] N. El Demerdash, J. Kedda, N. Ram, H. Brem, B. Tyler, Novel therapeutics for brain tumors: current practice and future prospects, Expert Opin. Drug Deliv. 17 (2020) 9–21, https://doi.org/10.1080/17425247.2019.1676227.
[91] S.K. Golombek, J.N. May, B. Theek, L. Appold, N. Drude, F. Kiessling, T. Lammers, Tumor targeting via EPR: strategies to enhance patient responses, Adv. Drug Deliv. Rev. 130 (2018) 17–38, https://doi.org/10.1016/j.addr.2018.07.007.
[92] D.A. Senthebane, A. Rowe, N.E. Thomford, H. Shipanga, D. Munro, M.A. Al Mazeedi, H.A. Almazyadi, K. Kallmeyer, C. Dandara, M.S. Pepper, I. Parker, The role of tumor microenvironment in chemoresistance: to survive, keep your enemies closer, Int. J. Mol. Sci. 18 (2017) 1586, https://doi.org/10.3390/ijms18071586.
[93] T. Chanmee, P. Ontong, K. Konno, N. Itano, Tumor-associated macrophages as major players in the tumor microenvironment, Cancers 6 (3) (2014) 1670–1690.
[94] T.L. Whiteside, Immune modulation of T-cell and NK (natural killer) cell activities by TEXs (tumour-derived exosomes), Biochem. Soc. Trans. 41 (2013) 245–251, https://doi.org/10.1042/BST20120265.
[95] D. Garnier, O. Renoult, M.C. Alves-Guerra, F. Paris, C. Pecqueur, Glioblastoma stem-like cells, metabolic strategy to kill a challenging target, Front. Oncol. 9 (2019) 118, https://doi.org/10.3389/fonc.2019.00118.
[96] R.K. Upadhyay, Drug delivery systems, CNS protection, and the blood brain barrier, Biomed. Res. Int. 2014 (2014), 869269, https://doi.org/10.1155/2014/869269.
[97] P. Dharun, Y. Ismail, V.P.M. Vijaya, Latentiated prodrug approach of drugs: an overview, Int. J. App. Pharm. 13 (2021), https://doi.org/10.22159/ijap.2021v13i6.42794.
[98] P. Poudel, S. Park, Recent advances in the treatment of Alzheimer's disease using nanoparticle-based drug delivery systems, Pharmaceutics. 14 (2022) 835, https://doi.org/10.3390/pharmaceutics14040835.
[99] D. Leung, X. Han, T. Mikkelsen, L.B. Nabors, Role of MRI in primary brain tumor evaluation, J. Natl. Compr. Cancer Netw. 12 (2014) 1561–1568, https://doi.org/10.6004/jnccn.2014.0156.
[100] A. Hadjipanteli, P. Doolan, E. Kyriacou, A. Constantinidou, Breast cancer brain metastasis: the potential role of MRI beyond current clinical applications, Cancer Manag. Res. 12 (2020) 9953, https://doi.org/10.2147/CMAR.S252801.
[101] H. Shooli, H. Dadgar, Y.X.J. Wáng, M.S. Vafaee, S.R. Kashuk, R. Nemati, E. Jafari, I. Nabipour, A. Gholamrezanezhad, M. Assadi, M. Larvie, An update on PET-based molecular imaging in neuro-oncology: challenges and implementation for a precision medicine approach in cancer care, Quant. Imaging. Med. Sur. 9 (2019) 1597, https://doi.org/10.21037/qims.2019.08.16.

[102] C. Özütemiz, E.C. Neil, M. Tanwar, N.T. Rubin, K. Ozturk, Z. Cayci, The role of dual-phase FDG PET/CT in the diagnosis and follow-up of brain tumors, Am. J. Roentgenol. 215 (2020) 985–996, https://doi.org/10.2214/AJR.19.22571.
[103] S.C. Hester, M. Kuriakose, C.D. Nguyen, S. Mallidi, Role of ultrasound and photoacoustic imaging in photodynamic therapy for cancer, Photochem. Photobiol. 96 (2020) 260–279, https://doi.org/10.1111/php.13217.
[104] C.M. Carapella, N. Gorgoglione, P.A. Oppido, The role of surgical resection in patients with brain metastases, Curr. Opin. Oncol. 30 (2018) 390–395, https://doi.org/10.3332/ecancer.2013.308.
[105] M.J. Van Den Bent, M.E. Hegi, R. Stupp, Recent developments in the use of chemotherapy in brain tumours, Eur. J. Cancer 42 (2006) 582–588, https://doi.org/10.1016/j.ejca.2005.06.031.
[106] M. Alemany, R. Velasco, M. Simó, J. Bruna, Late effects of cancer treatment: consequences for long-term brain cancer survivors, Neuro-Oncol. Pract. 8 (2020) 18–30, https://doi.org/10.1093/nop/npaa039.
[107] D. Steinmann, C. Schäfer, B. van Oorschot, H.J. Wypior, F. Bruns, T. Bölling, S. Sehlen, J. Hagg, A. Bayerl, H. Geinitz, M. Hipp, Effects of radiotherapy for brain metastases on quality of life (QoL), Strahlenther. Onkol. 185 (2009) 190–197, https://doi.org/10.1007/s00066-009-1904-0.
[108] R. Kotecha, V. Gondi, M.S. Ahluwalia, P.K. Brastianos, M.P. Mehta, Recent advances in managing brain metastasis, F1000Research (2018) 7, https://doi.org/10.12688/f1000research.15903.1.
[109] J.W. Kim, K.C. Pituch, A. Xiao, J.S. Young, W.K. Panek, M.E. Muroski, A. Rashidi, J.R. Kane, D. Kanojia, M.S. Lesniak, Gene/viral treatment approaches for malignant brain cancer, in: H.B. Newton (Ed.), Handbook of Brain Tumor Chemotherapy, Molecular Therapeutics, and Immunotherapy, Academic Press, Chembridge, Massachusetts, United States, 2018, pp. 211–227.
[110] E.T. Wong, E. Lok, K.D. Swanson, Alternating electric fields therapy for malignant gliomas: From bench observation to clinical reality, in: M.F. Chernov, Y. Muragaki, S. Kesari, I.E. McCutcheon (Eds.), Intracranial Gliomas Part III-Innovative Treatment Modalities, Karger Publishers, Basel, Switzerland, 2018, pp. 180–195.
[111] R. Di Lorenzo, M.S. Ahluwalia, Targeted therapy of brain metastases: latest evidence and clinical implications, Ther. Adv. Med. Oncol. 9 (2017) 781–796, https://doi.org/10.1177/1758834017736252.
[112] K. Mahmoudi, A. Bouras, D. Bozec, R. Ivkov, C. Hadjipanayis, Magnetic hyperthermia therapy for the treatment of glioblastoma: a review of the therapy's history, efficacy and application in humans, Int. J. Hyperth. 34 (2018) 1316–1328, https://doi.org/10.1080/02656736.2018.1430867.
[113] X. Wang, Y. Jia, P. Wang, Q. Liu, H. Zheng, Current status and future perspectives of sonodynamic therapy in glioma treatment, Ultrason. Sonochem. 37 (2017) 592–599, https://doi.org/10.1016/j.ultsonch.2017.02.020.
[114] G. Nabil, K. Bhise, S. Sau, M. Atef, H.A. El-Banna, A.K. Iyer, Nano-engineered delivery systems for cancer imaging and therapy: recent advances, future direction and patent evaluation, Drug Discov. Today 24 (2019) 462–491, https://doi.org/10.1016/j.drudis.2018.08.009.
[115] W. Tang, W. Fan, J. Lau, L. Deng, Z. Shen, X. Chen, Emerging blood-brain-barrier-crossing nanotechnology for brain cancer theranostics, Chem. Soc. Rev. 48 (2019) 2967–3014, https://doi.org/10.1039/c8cs00805a.
[116] P.J. Gaillard, Company profile: case study: to-BBB's G-Technology®, getting the best from drug-delivery research with industry–academia partnerships, Ther. Deliv. 2 (2011) 1391–1394, https://doi.org/10.4155/tde.11.111.
[117] C. Sun, Y. Ding, L. Zhou, D. Shi, L. Sun, T.J. Webster, Y. Shen, Noninvasive nanoparticle strategies for brain tumor targeting. Nanomedicine: Nanotechnol, Biol. Med. 13 (2017) 2605–2621, https://doi.org/10.1016/j.nano.2017.07.009.
[118] J.V. Georgieva, D. Hoekstra, I.S. Zuhorn, Smuggling drugs into the brain: an overview of ligands targeting transcytosis for drug delivery across the blood-brain barrier, Pharmaceutics. 6 (2014) 557–583, https://doi.org/10.3390/pharmaceutics6040557.
[119] M.S. de Almeida, E. Susnik, B. Drasler, P. Taladriz-Blanco, A. Petri-Fink, B. Rothen-Rutishauser, Understanding nanoparticle endocytosis to improve targeting strategies in nanomedicine, Chem. Soc. Rev. 50 (2021) 5397–5434, https://doi.org/10.1039/D0CS01127D.
[120] T.C. Sorrell, P.G. Juillard, J.T. Djordjevic, K. Kaufman-Francis, A. Dietmann, A. Milonig, V. Combes, G.E. Grau, Cryptococcal transmigration across a model brain blood-barrier: evidence of the Trojan horse mechanism and differences between Cryptococcus neoformans var. grubii strain H99 and Cryptococcus gattii strain R265, Microbes Infect. 18 (2016) 57–67, https://doi.org/10.1016/j.micinf.2015.08.017.
[121] T. Lin, P. Zhao, Y. Jiang, Y. Tang, H. Jin, Z. Pan, H. He, V.C. Yang, Y. Huang, Blood–brain-barrier-penetrating albumin nanoparticles for biomimetic drug delivery via albumin-binding protein pathways for antiglioma therapy, ACS Nano 10 (2016) 9999–10012, https://doi.org/10.1021/acsnano.6b04268.
[122] K. Fan, X. Jia, M. Zhou, K. Wang, J. Conde, J. He, J. Tian, X. Yan, Ferritin nanocarrier traverses the blood brain barrier and kills glioma, ACS Nano 12 (2018) 4105–4115, https://doi.org/10.1021/acsnano.7b06969.
[123] M.A. Alghamdi, A.N. Fallica, N. Virzì, P. Kesharwani, V. Pittalà, K. Greish, The promise of nanotechnology in personalized medicine, J. Pers. Med. 12 (2022) 673, https://doi.org/10.3390/jpm12050673.
[124] M. Pohanka, Quantum dots in the therapy: current trends and perspectives, Mini-Rev. Med. Chem. 17 (2017) 650–656, https://doi.org/10.2174/1389557517666170120153342.
[125] C. Yan, X. Du, J. Li, X. Ding, Z. Li, Y. Tang, Effect of excitation wavelength on optical performances of quantum-dot-converted light-emitting diode, Nanomater. 9 (2019) 1100, https://doi.org/10.3390/nano9081100.
[126] A.S. Thakor, S.S. Gambhir, Nanooncology: the future of cancer diagnosis and therapy, CA Cancer J. Clin. 63 (2013) 395–418, https://doi.org/10.3322/caac.21199.
[127] G. Xu, S. Mahajan, I. Roy, K.T. Yong, Theranostic quantum dots for crossing blood–brain barrier in vitro and providing therapy of HIV-associated encephalopathy, Front. Pharmacol. 4 (2013) 140, https://doi.org/10.3389/fphar.2013.00140.
[128] Y. Wang, L. Chen, Quantum dots, lighting up the research and development of nanomedicine, Nanomedicine 7 (2011) 385–402, https://doi.org/10.1016/j.nano.2010.12.006.
[129] M.K.V. Sonali, R.P. Singh, P. Agrawal, A.K. Mehata, N. Datta Maroti Pawde, R. Sonkar, M.S. Muthu, Nanotheranostics: emerging strategies for early diagnosis and therapy of brain cancer, Nanotheranostics. 2 (2018) 70, doi:1010.7150/ntno.21638.
[130] G. Xu, K.T. Yong, I. Roy, S.D. Mahajan, H. Ding, S.A. Schwartz, P.N. Prasad, Bioconjugated quantum rods as targeted probes for efficient transmigration across an in vitro blood− brain barrier, Bioconjug. Chem. 19 (2008) 1179–1185, https://doi.org/10.1021/bc700477u.
[131] E.S. Seven, Y. Zhou, Y.B. Seven, G.S. Mitchell, R.M. Leblanc, Crossing blood-brain barrier with carbon quantum dots, FASEB J. 33 (2019) (2019) 785–788, https://doi.org/10.1096/fasebj.2019.33.1_supplement.785.8.
[132] B.S. Alotaibi, M. Buabeid, N.A. Ibrahim, Z.J. Kharaba, M. Ijaz, S. Noreen, G. Murtaza, Potential of nanocarrier-based drug delivery systems for brain targeting: a current review of literature, Int. J. Nanomedicine 16 (2021) 7517, https://doi.org/10.2147/IJN.S333657.
[133] S.L. Gibbs-Strauss, K.S. Samkoe, J.A. O'Hara, S.C. Davis, P.J. Hoopes, T. Hasan, B. W. Pogue, Detecting epidermal growth factor receptor tumor activity in vivo during cetuximab therapy of murine gliomas, Acad. Radiol. 17 (2010) 7–17, https://doi.org/10.1016/j.acra.2009.07.027.
[134] D. Ag, R. Bongartz, L.E. Dogan, M. Seleci, J.G. Walter, D.O. Demirkol, F. Stahl, S. Ozcelik, S. Timur, T. Scheper, Biofunctional quantum dots as fluorescence probe for cell-specific targeting, Colloids Surf. B: Biointerfaces 114 (2014) 96–103, https://doi.org/10.1016/j.colsurfb.2013.09.033.
[135] J. Rico-Varela, T. Singh, S. McCutcheon, M. Vazquez, EGF as a new therapeutic target for medulloblastoma metastasis, Cell. Mol. Bioeng. 8 (2015) 553–565, https://doi.org/10.1007/s12195-015-0395-6.
[136] J. Wang, W.H. Yong, Y.H. Sun, P.T. Vernier, H.P. Koeffler, M.A. Gundersen, L. Marcu, Receptor-targeted quantum dots: fluorescent probes for brain tumor diagnosis, J. Biomed. Opt. 12 (2007), 044021, https://doi.org/10.1117/1.2764463.
[137] V. Dudu, V. Rotari, M. Vazquez, Targeted extracellular nanoparticles enable intracellular detection of activated epidermal growth factor receptor in living brain cancer cells, Nanomedicine. 7 (2011) 896–903, https://doi.org/10.1016/j.nano.2011.05.002.
[138] T. Hong, Y. Kai, Z. Cheng, B. Yunlong, H. Hao, In situ visual imaging and in vivo distribution of head and neck squamous cell carcinoma in mice by near-infrared fluorescent quantum dots epidermal growth factor receptor monoclonal antibody probe, Hua Xi Kou Qiang Yi Xue Za Zhi. 30 (2012) 568–573.
[139] A.L. Ribeiro de Souza, K. Marra, J. Gunn, K.S. Samkoe, S. Hull, K.D. Paulsen, B. W. Pogue, Optimizing glioma detection using an EGFR-targeted fluorescent Affibody, Photochem. Photobiol. 94 (2018) 1167–1171, https://doi.org/10.1111/php.13003.
[140] V. Dudu, R.A. Able, V. Rotari, Q. Kong, M. Vazquez, Role of epidermal growth factor-triggered PI3K/Akt signaling in the migration of medulloblastoma-derived cells, Cell. Mol. Bioeng. 5 (2012) 402–413, https://doi.org/10.1007/s12195-012-0253-8.
[141] H. Tada, H. Higuchi, T.M. Wanatabe, N. Ohuchi, In vivo real-time tracking of single quantum dots conjugated with monoclonal anti-HER2 antibody in tumors of mice, Cancer Res. 67 (2007) 1138–1144, https://doi.org/10.1158/0008-5472.CAN-06-1185.
[142] J. Fukai, K. Nishio, T. Itakura, F. Koizumi, Antitumor activity of cetuximab against malignant glioma cells overexpressing EGFR deletion mutant variant III, Cancer Sci. 99 (2008) 2062–2069, https://doi.org/10.1111/j.1349-7006.2008.00945.x.
[143] D.J. Arndt-Jovin, S.R. Kantelhardt, W. Caarls, A.H. de Vries, A. Giese, T.M. Jovin, Tumor-targeted quantum dots can help surgeons find tumor boundaries, IEEE Trans. Nanobiosci. 8 (2009) 65–71, https://doi.org/10.1109/TNB.2009.2016548.
[144] K. Yang, C. Zhao, Y.A. Cao, H. Tang, Y.L. Bai, H. Huang, C.R. Zhao, R. Chen, D. Zhao, In vivo and in situ imaging of head and neck squamous cell carcinoma using near-infrared fluorescent quantum dot probes conjugated with epidermal growth factor receptor monoclonal antibodies in mice, Oncol. Rep. 27 (2012) 1925–1931, https://doi.org/10.3892/or.2012.1705.
[145] K. Leung, QD800-anti-epidermal growth factor receptor monoclonal antibody nanoparticles, in: Molecular Imaging and Contrast Agent Database (MICAD), National Center for Biotechnology Information (US), 2013.
[146] R.W. Holt, J.L.H. Demers, K.J. Sexton, J.R. Gunn, S.C. Davis, K.S. Samkoe, B. W. Pogue, Tomography of epidermal growth factor receptor binding to fluorescent Affibody in vivo studied with magnetic resonance guided fluorescence recovery in varying orthotopic glioma sizes, J. Biomed. Opt. 20 (2015), 026001, https://doi.org/10.1117/1.JBO.20.2.026001.
[147] A.L. De, K. Souza, J. Marra, K.S. Gunn, P.J. Samkoe, J. Hoopes, K.D. Feldwisch, B. W. Paulsen, Pogue, Fluorescent affibody molecule administered in vivo at a microdose level labels EGFR expressing glioma tumor regions, Mol. Imaging Biol. 19 (2017) 41–48, https://doi.org/10.1007/s11307-016-0980-7.

[148] M. Vibin, R. Vinayakan, F.B. Fernandez, A. John, A. Abraham, A novel fluorescent quantum dot probe for the rapid diagnostic high contrast imaging of tumor in mice, Fluoresc. 27 (2017) 669–677, https://doi.org/10.1007/s10895-016-1996-8.
[149] F. Ramos-Gomes, J. Bode, A. Sukhanova, S.V. Bozrova, M. Saccomano, M. Mitkovski, J.E. Krueger, A.K. Wege, W. Stuehmer, P.S. Samokhvalov, D. Baty, Single-and two-photon imaging of human micrometastases and disseminated tumour cells with conjugates of nanobodies and quantum dots, Sci. Rep. 8 (2018) 1–12, https://doi.org/10.1038/s41598-018-22973-8.
[150] E. McCord, S. Pawar, T. Koneru, K. Tatiparti, S. Sau, A.K. Iyer, Folate receptors' expression in gliomas may possess potential nanoparticle-based drug delivery opportunities, ACS Omega 6 (2021) 4111–4118, https://doi.org/10.1038/s41598-018-22973-8.
[151] M. Geszke, M. Murias, L. Balan, G. Medjahdi, J. Korczynski, M. Moritz, J. Lulek, R. Schneider, Folic acid-conjugated core/shell ZnS: Mn/ZnS quantum dots as targeted probes for two photon fluorescence imaging of cancer cells, Acta Biomater. 7 (2011) 1327–1338, https://doi.org/10.1016/j.actbio.2010.10.012.
[152] P. Suriamoorthy, X. Zhang, G. Hao, A.G. Joly, S. Singh, M. Hossu, X. Sun, W. Chen, Folic acid-CdTe quantum dot conjugates and their applications for cancer cell targeting, Cancer Nanotechnol. 1 (2010) 19–28, https://doi.org/10.1007/s12645-010-0003-3.
[153] I.B. Bwatanglang, F. Mohammad, N.A. Yusof, J. Abdullah, M.Z. Hussein, N. B. Alitheen, N. Abu, Folic acid targeted Mn:ZnS quantum dots for theranostic applications of cancer cell imaging and therapy, Int. J. Nanomedicine 11 (2016) 413–428, https://doi.org/10.2147/IJN.S90198.
[154] S. Bansal, J. Singh, U. Kumari, I.P. Kaur, R.P. Barnwal, R. Kumar, S. Singh, G. Singh, M. Chatterjee, Development of biosurfactant-based graphene quantum dot conjugate as a novel and fluorescent theranostic tool for cancer, Int. J. Nanomedicine 14 (2019) 809–818, https://doi.org/10.2147/IJN.S188552.
[155] S. Feng, J. Pan, C. Li, Y. Zheng, Folic acid-conjugated nitrogen-doped graphene quantum dots as a fluorescent diagnostic material for MCF-7 cells, Nanotechnol. 31 (2020), 135701, https://doi.org/10.1088/1361-6528/ab5f7f.
[156] J. Guo, M. Schlich, J.F. Cryan, C.M. O'Driscoll, Targeted drug delivery via folate receptors for the treatment of brain cancer: can the promise deliver? J. Pharm. Sci. 106 (2017) (2017) 3413–3420, https://doi.org/10.1016/j.xphs.2017.08.009.
[157] G. Singh, M. Kumar, U. Soni, V. Arora, V. Bansal, D. Gupta, M. Bhat, A.K. Dinda, S. Sapra, H. Singh, Cancer cell targeting using folic acid/anti-HER2 antibody conjugated fluorescent CdSe/CdS/ZnS-MPA and CdTe-MSA quantum dots, J. Nanosci. Nanotechnol. 15 (2015) 9382–9395, https://doi.org/10.1166/jnn.2015.10767.
[158] J. Lee, Y. Choi, K. Kim, S. Hong, H.Y. Park, T. Lee, G.J. Cheon, R. Song, Characterization and cancer cell specific binding properties of anti-EGFR antibody conjugated quantum dots, Bioconjug. Chem. 21 (2010) 940–946, https://doi.org/10.1021/bc9004975.
[159] M.Z. Zhang, R.N. Yu, J. Chen, Z.Y. Ma, Y.D. Zhao, Targeted quantum dots fluorescence probes functionalized with aptamer and peptide for transferrin receptor on tumor cells, Nanotechnol. 23 (2012), 485104, https://doi.org/10.1088/0957-4484/23/48/485104.
[160] H. Choudhury, M. Pandey, P.X. Chin, Y.L. Phang, J.Y. Cheah, S.C. Ooi, K.K. Mak, M.R. Pichika, P. Kesharwani, Z. Hussain, B. Gorain, Transferrin receptors-targeting nanocarriers for efficient targeted delivery and transcytosis of drugs into the brain tumors: a review of recent advancements and emerging trends, Drug Deliv. Transl. Res. 8 (2018) 1545–1563, https://doi.org/10.1007/s13346-018-0552-2.
[161] H. Chen, J. Xue, Y. Zhang, X. Zhu, J. Gao, B. Yu, Comparison of quantum dots immunofluorescence histochemistry and conventional immunohistochemistry for the detection of caveolin-1 and PCNA in the lung cancer tissue microarray, J. Mol. Histol. 40 (2009) 261–268, https://doi.org/10.1007/s10735-009-9237-y.
[162] Ö. Şen, M. Emanet, G. Ciofani, Nanotechnology-based strategies to evaluate and counteract cancer metastasis and neoangiogenesis, Adv. Healthc. Mater. 10 (2021) 2002163, https://doi.org/10.1002/adhm.202002163.
[163] L. Pang, J. Xu, C. Shu, J. Guo, X. Ma, Y. Liu, W. Zhong, Characterization and cancer cell targeted imaging properties of human antivascular endothelial growth factor monoclonal antibody conjugated CdTe/ZnS quantum dots, Luminescence 29 (2014) 1177–1182, https://doi.org/10.1002/bio.2679.
[164] Y. Song, C. Hu, Y. Fu, H. Gao, Modulating the blood–brain tumor barrier for improving drug delivery efficiency and efficacy, View 3 (2022) 20200129, https://doi.org/10.1002/VIW.20200129.
[165] A.A. Mansur, H.S. Mansur, S.M. Carvalho, Z.I. Lobato, M.D.F. Leite, L.L. Mansur, Fluorescent ZnS quantum dots–phosphoethanolamine nanoconjugates for bioimaging live cells in cancer research, ACS Omega 3 (2018) 15679–15691, https://doi.org/10.1021/acsomega.8b02098.
[166] D. Manzanares, V. Ceña, Endocytosis: the nanoparticle and submicron nanocompounds gateway into the cell, Pharmaceutics 12 (2020) 371, https://doi.org/10.3390/pharmaceutics12040371.
[167] L.B. Thal, V.R. Mann, D. Sprinzen, J.R. McBride, K.R. Reid, I.D. Tomlinson, D. G. McMahon, B.E. Cohen, S.J. Rosenthal, Ligand-conjugated quantum dots for fast sub-diffraction protein tracking in acute brain slices, Biomater. Sci. 8 (2020) 837–845, https://doi.org/10.1039/C9BM01629E.
[168] O.A. Aladesuyi, O.S. Oluwafemi, Synthesis strategies and application of ternary quantum dots-in cancer therapy, Nano-Struct. Nano-Objects. 24 (2020), 100568, https://doi.org/10.1016/j.nanoso.2020.100568.
[169] P. Henrich-Noack, D. Nikitovic, M. Neagu, A.O. Docea, A.B. Engin, S. Gelperina, M. Shtilman, P. Mitsias, G. Tzanakakis, I. Gozes, A. Tsatsakis, The blood-brain barrier and beyond: nano-based neuropharmacology and the role of extracellular matrix, Nanomedicine 17 (2019) 359–379, https://doi.org/10.1016/j.nano.2019.01.016.
[170] M.J. Levene, D.A. Dombeck, K.A. Kasischke, R.P. Molloy, W.W. Webb, In vivo multiphoton microscopy of deep brain tissue, J. Neurophysiol. 91 (2004) 1908–1912, https://doi.org/10.1152/jn.01007.2003.
[171] H. Jackson, O. Muhammad, H. Daneshvar, J. Nelms, A. Popescu, M. A. Vogelbaum, M. Bruchez, S.A. Toms, Quantum dots are phagocytized by macrophages and colocalize with experimental gliomas, Neurosurgery. 60 (2007) 524–530, https://doi.org/10.1227/01.NEU.0000255334.95532.DD.
[172] O. Muhammad, A. Popescu, S.A. Toms, Macrophage-mediated colocalization of quantum dots in experimental glioma, Methods Mol. Biol. 374 (2007) 161–171, https://doi.org/10.1385/1-59745-369-2:161.
[173] X.C. Chen, Y.L. Deng, Y. Lin, D.W. Pang, H. Qing, F. Qu, H.Y. Xie, Quantum dot-labeled aptamer nanoprobes specifically targeting glioma cells, Nanotechnol. 19 (2008), 235105, https://doi.org/10.1088/0957-4484/19/23/235105.
[174] J. Jung, A. Solanki, K.A. Memoli, K.I. Kamei, H. Kim, M.A. Drahl, L.J. Williams, H. R. Tseng, K. Lee, Selective inhibition of human brain tumor cells through multifunctional quantum-dot-based siRNA delivery, Angew. Chem. 49 (2010) 103–107, https://doi.org/10.1002/anie.200905126.
[175] Z.M. Markovic, B.Z. Ristic, K.M. Arsikin, D.G. Klisic, L.M. Harhaji-Trajkovic, B. M. Todorovic-Markovic, D.P. Kepic, T.K. Kravic-Stevovic, S.P. Jovanovic, M. M. Milenkovic, D.D. Milivojevic, Graphene quantum dots as autophagy-inducing photodynamic agents, Biomaterials. 33 (2012) 7084–7092, https://doi.org/10.1016/j.biomaterials.2012.06.060.
[176] C.J. Wen, L.W. Zhang, S.A. Al-Suwayeh, T.C. Yen, J.Y. Fang, Theranostic liposomes loaded with quantum dots and apomorphine for brain targeting and bioimaging, Int. J. Nanomedicine 7 (2012) 1599, https://doi.org/10.2147/IJN.S29369.
[177] K.C. Weng, R. Hashizume, C.O. Noble, L.P. Serwer, D.C. Drummond, D. B. Kirpotin, A.M. Kuwabara, L.X. Chao, F.F. Chen, C.D. James, J.W. Park, Convection-enhanced delivery of targeted quantum dot-immunoliposome hybrid nanoparticles to intracranial brain tumor models, Nanomedicine. 8 (2013) 1913–1925, https://doi.org/10.2217/nnm.12.209.
[178] D. Fatehi, T.N. Baral, A. Abulrob, In vivo imaging of brain cancer using epidermal growth factor single domain antibody bioconjugated to near-infrared quantum dots, J. Nanosci. Nanotechnol. 14 (2014) 5355–5362, https://doi.org/10.1166/jnn.2014.9076.
[179] S. Ruan, J. Qian, S. Shen, J. Zhu, X. Jiang, Q. He, H. Gao, A simple one-step method to prepare fluorescent carbon dots and their potential application in non-invasive glioma imaging, Nanoscale. 6 (2014) 10040–10047, https://doi.org/10.1039/C4NR02657H.
[180] S. Ruan, J. Qian, S. Shen, J. Chen, J. Zhu, X. Jiang, Q. He, W. Yang, H. Gao, Fluorescent carbonaceous nanodots for noninvasive glioma imaging after angiopep-2 decoration, Bioconjug. Chem. 25 (2014) 2252–2259, https://doi.org/10.1021/bc500474p.
[181] X. Yuan, Z. Liu, Z. Guo, Y. Ji, M. Jin, X. Wang, Cellular distribution and cytotoxicity of graphene quantum dots with different functional groups, Nanoscale Res. Lett. 9 (2014) 108, https://doi.org/10.1186/1556-276X-9-108.
[182] Q. Bai, Z. Zhao, H. Sui, J. Chen, X. Xie, F. Wen, The preparation and application of dendrimer modified CdTe/CdS near infrared quantum dots for brain cancer cells imaging, Appl. Sci. 5 (2015) 1076–1085, https://doi.org/10.3390/app5041076.
[183] Y. Wang, Y. Meng, S. Wang, C. Li, W. Shi, J. Chen, J. Wang, R. Huang, Direct solvent-derived polymer-coated nitrogen-doped carbon nanodots with high water solubility for targeted fluorescence imaging of glioma, Small. 11 (2015) 3575–3581, https://doi.org/10.1002/smll.201403718.
[184] M. Zheng, S. Ruan, S. Liu, T. Sun, D. Qu, H. Zhao, Z. Xie, H. Gao, X. Jing, Z. Sun, Self-targeting fluorescent carbon dots for diagnosis of brain cancer cells, ACS Nano 9 (2015) 11455–11461, https://doi.org/10.1021/acsnano.5b05575.
[185] P.E. Cabral Filho, A.L. Cardoso, M.I. Pereira, A.P. Ramos, F. Hallwass, M. M. Castro, C.F. Geraldes, B.S. Santos, M.C. Pedroso de Lima, G.A. Pereira, A. Fontes, CdTe quantum dots as fluorescent probes to study transferrin receptors in glioblastoma cells, Biochim. Biophys. Acta 1860 (2016) 28–35, https://doi.org/10.1016/j.bbagen.2015.09.021.
[186] S. Lu, S. Guo, P. Xu, X. Li, Y. Zhao, W. Gu, M. Xue, Hydrothermal synthesis of nitrogen-doped carbon dots with real-time live-cell imaging and blood–brain barrier penetration capabilities, Int. J. Nanomedicine 11 (2016) 6325, https://doi.org/10.2147/IJN.S119252.
[187] R.P. Singh, G. Sharma, L. Kumari, B. Koch, S. Singh, S. Bharti, P.S. Rajinikanth, B. L. Pandey, M.S. Muthu, RGD-TPGS decorated theranostic liposomes for brain targeted delivery, Colloids Surf. B: Biointerfaces 147 (2016) 129–141, https://doi.org/10.1016/j.colsurfb.2016.07.058.
[188] Z. Fan, S. Zhou, C. Garcia, L. Fan, J. Zhou, pH-responsive fluorescent graphene quantum dots for fluorescence-guided cancer surgery and diagnosis, Nanoscale. 9 (2017) 4928–4933, https://doi.org/10.1039/C7NR00888K.
[189] J. Tang, N. Huang, X. Zhang, T. Zhou, Y. Tan, J. Pi, L. Pi, S. Cheng, H. Zheng, Y. Cheng, Aptamer-conjugated PEGylated quantum dots targeting epidermal growth factor receptor variant III for fluorescence imaging of glioma, Int. J. Nanomedicine 12 (2017) 3899–3911, https://doi.org/10.2147/IJN.S133166.
[190] L. Gao, X. Zhao, J. Wang, Y. Wang, L. Yu, H. Peng, J. Zhu, Multiple functionalized carbon quantum dots for targeting glioma and tissue imaging, Opt. Mater. 75 (2018) 764–769, https://doi.org/10.1016/j.optmat.2017.11.044.
[191] L. Qiao, T. Sun, X. Zheng, M. Zheng, Z. Xie, Exploring the optimal ratio of d-glucose/l-aspartic acid for targeting carbon dots toward brain tumor cells, Mater. Sci. Eng. C 85 (2018) 1–6, https://doi.org/10.1016/j.msec.2017.12.011.
[192] H.L. Xu, J.J. Yang, D.L. ZhuGe, M.T. Lin, Q.Y. Zhu, B.H. Jin, M.Q. Tong, B. X. Shen, J. Xiao, Y.Z. Zhao, Glioma-targeted delivery of a theranostic liposome

integrated with quantum dots, superparamagnetic iron oxide, and cilengitide for dual-imaging guiding cancer surgery, Adv. Healthc. Mater. 7 (2018), e1701130, https://doi.org/10.1002/adhm.201701130.
[193] I.C. Carvalho, A.A. Mansur, S.M. Carvalho, R.M. Florentino, H.S. Mansur, L-cysteine and poly-L-arginine grafted carboxymethyl cellulose/Ag-in-S quantum dot fluorescent nanohybrids for in vitro bioimaging of brain cancer cells, Int. J. Biol. Macromol. 133 (2019) 739–753, https://doi.org/10.1016/j.ijbiomac.2019.04.140.
[194] A.A. Mansur, A.J. Caires, S.M. Carvalho, N.S. Capanema, I.C. Carvalho, H. S. Mansur, Dual-functional supramolecular nanohybrids of quantum dot/biopolymer/chemotherapeutic drug for bioimaging and killing brain cancer cells in vitro, Colloids Surf. B: Biointerfaces 184 (2019), 110507, https://doi.org/10.1016/j.colsurfb.2019.110507.
[195] C.P. Santana, A.A. Mansur, S.M. Carvalho, A. da Silva-Cunha Jr, H.S. Mansur, Bi-functional quantum dot-polysaccharide-antibody immunoconjugates for bioimaging and killing brain cancer cells in vitro, Mater. Lett. 252 (2019) 333–337, https://doi.org/10.1016/j.matlet.2019.06.022.
[196] M. Shamsipour, A.M. Mansouri, P. Moradipour, Temozolomide conjugated carbon quantum dots embedded in core/shell nanofibers prepared by coaxial electrospinning as an implantable delivery system for cell imaging and sustained drug release, AAPS PharmSciTech 20 (2019) 1–14, https://doi.org/10.1208/s12249-019-1466-0.
[197] R. Wahab, N. Kaushik, F. Khan, N.K. Kaushik, S.J. Lee, E.H. Choi, A.A. Al-Khedhairy, Gold quantum dots impair the tumorigenic potential of glioma stem-like cells via β-catenin downregulation in vitro, Int. J. Nanomedicine 14 (2019) 1131, https://doi.org/10.2147/IJN.S195333.
[198] A.J. Caires, A.A. Mansur, I.C. Carvalho, S.M. Carvalho, H.S. Mansur, Green synthesis of ZnS quantum dot/biopolymer photoluminescent nanoprobes for bioimaging brain cancer cells, Mater. Chem. Phys. 244 (2020), 122716, https://doi.org/10.1016/j.matchemphys.2020.122716.
[199] N.K. Kaushik, N. Kaushik, R. Wahab, P. Bhartiya, N.N. Linh, F. Khan, A.A. Al-Khedhairy, E.H. Choi, Cold atmospheric plasma and gold quantum dots exert dual cytotoxicity mediated by the cell receptor-activated apoptotic pathway in glioblastoma cells, Cancers. 12 (2020) 457, https://doi.org/10.3390/cancers12020457.
[200] G. Perini, V. Palmieri, G. Ciasca, M. De Spirito, M. Papi, Unravelling the potential of graphene quantum dots in biomedicine and neuroscience, Int. J. Mol. Sci. 21 (2020) 3712, https://doi.org/10.3390/ijms21103712.
[201] G. Perini, V. Palmieri, G. Ciasca, A. Primiano, J. Gervasoni, M. De Spirito, M. Papi, Functionalized graphene quantum dots modulate malignancy of glioblastoma multiforme by downregulating neurospheres formation, C 7 (2021) 4, https://doi.org/10.3390/c7010004.
[202] C. Thiegs, Nanoparticle-mediated Assessment of ROS and Radiosensitization of Brain Cancer Cells for Improved Radiotherapy Outcomes (Doctoral dissertation), Creighton University, 2021. Creighton Theses and Dissertations, Available online at, https://dspace2.creighton.edu/xmlui/handle/10504/133740 (accessed on March 17, 2022).
[203] Z. Li, C. Zhao, Q. Fu, J. Ye, L. Su, X. Ge, L. Chen, J. Song, H. Yang, Neodymium (3+)-coordinated black phosphorus quantum dots with retrievable NIR/X-Ray optoelectronic switching effect for anti-glioblastoma, Small. 18 (2022) 2105160, https://doi.org/10.1002/smll.202105160.
[204] Y. Ren, C. Miao, L. Tang, Y. Liu, P. Ni, Y. Gong, H. Li, F. Chen, S. Feng, Homotypic cancer cell membranes camouflaged nanoparticles for targeting drug delivery and enhanced chemo-photothermal therapy of glioma, Pharmaceuticals. 15 (2022) 157, https://doi.org/10.3390/ph15020157.
[205] T. Wu, X. Wang, J. Cheng, X. Liang, Y. Li, M. Chen, L. Kong, M. Tang, Nitrogen-doped graphene quantum dots induce ferroptosis through disrupting calcium homeostasis in microglia, Part. Fibre Toxicol. 19 (2022) 22, https://doi.org/10.1186/s12989-022-00464-z.
[206] P. Pericleous, M. Gazouli, A. Lyberopoulou, S. Rizos, N. Nikiteas, E. P. Efstathopoulos, Quantum dots hold promise for early cancer imaging and detection, Int. J. Cancer 131 (2012) 519–528, https://doi.org/10.1002/ijc.27528.
[207] S. Chinnathambi, N. Shirahata, Recent advances on fluorescent biomarkers of near-infrared quantum dots for in vitro and in vivo imaging, Sci. Technol. Adv. Mater. 20 (2019) 337–355, https://doi.org/10.1080/14686996.2019.1590731.
[208] A. Cheung, H.J. Bax, D.H. Josephs, K.M. Ilieva, G. Pellizzari, J. Opzoomer, J. Bloomfield, M. Fittall, A. Grigoriadis, M. Figini, S. Canevari, Targeting folate receptor alpha for cancer treatment, Oncotarget. 7 (2016) 52553, https://doi.org/10.18632/oncotarget.9651.
[209] D.R. Emlet, P. Gupta, M. Holgado-Madruga, C.A. Del Vecchio, S.S. Mitra, S. Y. Han, G. Li, K.C. Jensen, H. Vogel, L.W. Xu, S.S. Skirboll, Targeting a glioblastoma cancer stem-cell population defined by EGF receptor variant III, Cancer Res. 74 (2014) 1238–1249, https://doi.org/10.1158/0008-5472.CAN-13-1407.
[210] A.H. Thorne, C. Zanca, F. Furnari, Epidermal growth factor receptor targeting and challenges in glioblastoma, Neuro-oncol. 18 (2016) 914–918, https://doi.org/10.1093/neuonc/nov319.
[211] C. Allolio, A. Magarkar, P. Jurkiewicz, K. Baxová, M. Javanainen, P.E. Mason, R. Šachl, M. Cebecauer, M. Hof, D. Horinek, V. Heinz, Arginine-rich cell-penetrating peptides induce membrane multilamellarity and subsequently enter via formation of a fusion pore, Proc. Natl. Acad. Sci. U. S. A. 115 (2018) 11923–11928, https://doi.org/10.1073/pnas.1811520115.
[212] M.F. Hansen, E. Greibe, S. Skovbjerg, S. Rohde, A.C. Kristensen, T.R. Jensen, C. Stentoft, K.H. Kjær, C.S. Kronborg, P.M. Martensen, Folic acid mediates activation of the pro-oncogene STAT3 via the folate receptor alpha, Cell. Signal. 27 (2015) 1356–1368, https://doi.org/10.1016/j.cellsig.2015.03.020.
[213] P. Mucha, E. Sikorska, P. Rekowski, J. Ruczyński, Interaction of arginine-rich cell-penetrating peptides with an artificial neuronal membrane, Cells 11 (2022) 1638, https://doi.org/10.3390/cells11101638.
[214] A.A.H. Abdellatif, M.A. Younis, M. Alsharidah, O.A. Rugaie, H.M. Tawfeek, Biomedical applications of quantum dots: overview, challenges, and clinical potential, Int. J. Nanomedicine 17 (2022) 1951, https://doi.org/10.2147/IJN.S357980.
[215] M.S. Muthu, S.A. Kulkarni, A. Raju, S.S. Feng, Theranostic liposomes of TPGS coating for targeted co-delivery of docetaxel and quantum dots, Biomater. 33 (2012) 3494–3501, https://doi.org/10.1016/j.biomaterials.2012.01.036.
[216] G. Perini, V. Palmieri, G. Ciasca, M. D'Ascenzo, A. Primiano, J. Gervasoni, F. De Maio, M. De Spirito, M. Papi, Enhanced chemotherapy for glioblastoma multiforme mediated by functionalized graphene quantum dots, Materials. 13 (2020) 4139, https://doi.org/10.3390/ma13184139.
[217] N. Haider, S. Fatima, M. Taha, M. Rizwanullah, J. Firdous, R. Ahmad, F. Mazhar, M.A. Khan, Nanomedicines in diagnosis and treatment of cancer: an update, Curr. Pharm. Design. 26 (2020) 1216–1231, https://doi.org/10.2174/1381612826666200318170716.
[218] I.L. Christensen, Y.P. Sun, P. Juzenas, Carbon dots as antioxidants and prooxidants, J. Biomed. Nanotechnol. 7 (2011) 667–676, https://doi.org/10.1166/jbn.2011.1334.
[219] M. Hersom, H.C. Helms, N. Pretzer, C. Goldeman, A.I. Jensen, G. Severin, M. S. Nielsen, R. Holm, B. Brodin, Transferrin receptor expression and role in transendothelial transport of transferrin in cultured brain endothelial monolayers, Mol. Cell. Neurosci. 76 (2016) 59–67, https://doi.org/10.1016/j.mcn.2016.08.009.
[220] T. Tashima, Smart strategies for therapeutic agent delivery into brain across the blood–brain barrier using receptor-mediated transcytosis, Chem. Pharm. Bull. 68 (2020) 316–325, https://doi.org/10.1248/cpb.c19-00854.
[221] X. Yue, F. Lan, W. Yang, Y. Yang, L. Han, A. Zhang, J. Liu, H. Zeng, T. Jiang, P. Pu, C. Kang, Interruption of β-catenin suppresses the EGFR pathway by blocking multiple oncogenic targets in human glioma cells, Brain Res. 1366 (2010) 27–37, https://doi.org/10.1016/j.brainres.2010.10.032.
[222] H. Zhao, Y. Wang, C. Yang, J. Zhou, L. Wang, K. Yi, Y. Li, J. Shi, C. Kang, L. Zeng, EGFR-vIII downregulated H2AZK4/7AC though the PI3K/AKT-HDAC2 axis to regulate cell cycle progression, Clin. Transl. Med. 9 (2020) 1–15, https://doi.org/10.1186/s40169-020-0260-7.
[223] L. Li, S. Xu, H. Yan, X. Li, H.S. Yazd, X. Li, T. Huang, C. Cui, J. Jiang, W. Tan, Nucleic acid aptamers for molecular diagnostics and therapeutics: advances and perspectives, Angew. Chem. Int. Ed. 60 (2021) 2221–2231, https://doi.org/10.1002/anie.202003563.
[224] H.S. Choi, W. Liu, F. Liu, K. Nasr, P. Misra, M.G. Bawendi, J.V. Frangioni, Design considerations for tumour-targeted nanoparticles, Nat. Nanotechnol. 5 (2010) 42–47, https://doi.org/10.1038/nnano.2009.314.
[225] R. Zein, W. Sharrouf, K. Selting, Physical properties of nanoparticles that result in improved cancer targeting, J. Oncol. 2020 (2020) 5194780, https://doi.org/10.1155/2020/5194780.
[226] L.M. Sturla, M. Tong, N. Hebda, J. Gao, J.M. Thomas, M. Olsen, M. Suzanne, Aspartate-β-hydroxylase (ASPH): A potential therapeutic target in human malignant gliomas, Heliyon. 2 (2016), e00203, https://doi.org/10.1016/j.heliyon.2016.e00203.
[227] N.A. Charoo, Z. Rahman, M.A. Khan, Nanoparticles for improvement in oral bioavailability, in: A.M. Grumezescu (Ed.), Nanoarchitectonics in Biomedicine, William Andrew Publishing, Elsevier, Netherlands, 2019, pp. 371–410.
[228] R.Q. Yang, K.L. Lou, P.Y. Wang, Y.Y. Gao, Y.Q. Zhang, M. Chen, W.H. Huang, G. J. Zhang, Surgical navigation for malignancies guided by near-infrared-II fluorescence imaging, Small Methods. 5 (2021) 2001066, https://doi.org/10.1002/smtd.202001066.
[229] M. Alshamrani, Broad-spectrum theranostics and biomedical application of functionalized nanomaterials, Polymers. 14 (2022) 1221, https://doi.org/10.3390/polym14061221.
[230] W. Zhang, G. Sigdel, K.J. Mintz, E.S. Seven, Y. Zhou, C. Wang, R.M. Leblanc, Carbon dots: a future blood-brain barrier penetrating nanomedicine and drug nanocarrier, Int. J. Nanomedicine 16 (2021) 5003, https://doi.org/10.2147/IJN.S318732.
[231] M. Souri, M. Soltani, F.M. Kashkooli, M.K. Shahvandi, M. Chiani, F.S. Shariati, M. R. Mehrabi, L.L. Munn, Towards principled design of cancer nanomedicine to accelerate clinical translation, Materials Today Bio 13 (2022), 100208, https://doi.org/10.1016/j.mtbio.2022.100208.
[232] W. Tao, O.C. Farokhzad, Theranostic nanomedicine in the NIR-II window: classification, fabrication, and biomedical applications, Chem. Rev. 122 (2022) 5405–5407, https://doi.org/10.1021/acs.chemrev.2c00089.
[233] Q. Jia, Z. Zhao, K. Liang, F. Nan, Y. Li, J. Wang, J. Ge, P. Wang, Recent advances and prospects of carbon dots in cancer nanotheranostics, Mater. Chem. Front. 4 (2020) 449–471, https://doi.org/10.1039/C9QM00667B.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/363093115

# Recent Advances in Understanding SARS-CoV-2 Infection and Updates on Potential Diagnostic and Therapeutics for COVID-19

**Article** · April 2022

DOI: 10.2174/2666796703666220302143102

CITATIONS
0

READS
27

**7 authors**, including:

Damini Verma
Amity University
**16** PUBLICATIONS **152** CITATIONS
SEE PROFILE

Amit K. Yadav
Jawaharlal Nehru University
**25** PUBLICATIONS **261** CITATIONS
SEE PROFILE

Navneet Chaudhary
Jawaharlal Nehru University
**9** PUBLICATIONS **76** CITATIONS
SEE PROFILE

Pramod Kumar
Sri Aurobido College, Delhi University
**15** PUBLICATIONS **136** CITATIONS
SEE PROFILE

**Some of the authors of this publication are also working on these related projects:**

Project
Electrochemical microfluidics based nanobiochips for mycotoxins detection View project

Project
Electrochemical biosensors platform based on naosized labels for advanced healthcare View project

All content following this page was uploaded by Amit K. Yadav on 07 September 2022.

The user has requested enhancement of the downloaded file.

*Send Orders for Reprints to reprints@benthamscience.net*

*Coronaviruses*, 2022, *3*, e020322201607

REVIEW ARTICLE

# Recent Advances in Understanding SARS-CoV-2 Infection and Updates on Potential Diagnostic and Therapeutics for COVID-19

Damini Verma[1,2,#], Amit K. Yadav[1,#], Navneet Chaudhary[3], Maumita D. Mukherjee[2,*], Pramod Kumar[4], Anil Kumar[5,*], Pratima R. Solanki[1,*]

[1]*Nano-Bio Laboratory, Special Centre for Nanoscience, Jawaharlal Nehru University, New Delhi-110067, India;* [2]*Amity Institute of Applied Sciences, Amity University, Noida, Uttar Pradesh-201313, India;* [3]*Department of Biotechnology, Delhi Technological University, Delhi-110042, India;* [4]*Sri Aurobindo College, Delhi University, New Delhi-110017, India;* [5]*National Institute of Immunology, New Delhi-110067, India*

**Abstract:** A more focused approach is needed to understand the SARS-CoV-2 virulence, structure, and genomics to devise more effective diagnostic and treatment interventions as this virus can evade the immune attack and causes life-threatening complications such as cytokine storm. The spread of the virus is still amplifying and causing thousands of new cases worldwide. It is essential to review current diagnostics and treatment approaches to pave the way to correct or modify our current practices to make more effective interventions against COVID-19. COVID-19 vaccine development has moved at a breakneck pace since the outbreak began, utilizing practically all possible platforms or tactics to ensure the success of vaccines. A total of 42 vaccine candidates have already entered clinical trials, including promising results from numerous vaccine candidates in phase 1 or phase 2 trials. Further, many existing drugs are being explored on broad-spectrum antiviral medications for their use in clinical recovery against COVID-19. The present review attempts to re-examine the SARS-CoV-2 structure, its viral life cycle, clinical symptoms and pathogenesis, mode of transmission, diagnostics, and treatment strategies that may be useful for resorting to more effective approaches for controlling COVID-19. Various antiviral drugs and vaccination strategies with their strengths and weaknesses are also discussed in the paper to augment our understanding of COVID-19 management.

ARTICLE HISTORY

Received: July 23, 2021
Revised: November 09, 2021
Accepted: December 13, 2021

*DOI:*
*10.2174/2666796703666220302143102*

**Keywords:** Coronavirus, COVID-19, diagnostic kits, SARS-CoV-2, therapeutic drugs, vaccines.

## 1. INTRODUCTION AND BACKGROUND

From late December 2019, the outbreak of 2019 novel coronavirus disease termed COVID-19 (by the WHO on February 11, 2020) has shown increased alarms and potential danger due to the severe acute respiratory syndrome virus 2 (SARS-CoV-2), representing a pandemic warning to public health globally [1]. The epicenter of the COVID-19 was located in the Wuhan province of China, where numerous cases of adults having severe pneumonia with unknown etiology were admitted to the nearby hospitals [2-4]. The SARS-CoV-2 was first isolated on January 7, 2020. Since then, the virus has spread globally, affected approximately all countries (Fig. **1**), and very quickly led to the announcement of a global health alert. While preparing this manuscript, the WHO reported 231,416,660 confirmed cases and 4,741,330 confirmed deaths till September 27th, 2021 [5].

Coronaviruses (CoVs) are one of the large groups of viruses, having a positive sense single-stranded (ss) RNA genome encapsulated inside the membrane that primarily targets the respiratory system of humans. Previously, the CoVs outbreak has posed a significant threat to public health, such as the severe acute respiratory syndrome (SARS-CoV) and the Middle East respiratory syndrome (MERS-CoV), as they were characterized as infectious agents for public destruction. Though, WHO has classified the CoVs into four types as alpha, beta, delta, and gamma, *i.e.*, α-CoVs, β-CoVs, δ-CoVs γ-CoVs, respectively [6], in all these types, β-CoVs have potentially caused SARS-CoV, MERS-CoV, and SARS-CoV-2, *i.e.*, the 7th member of the CoVs family that has created disease in human beings [7]. However, the current data states that SARS-CoV-2 is more contagious than SARS-CoV and MERS-CoV, which previously caused epidemics in 2002 and 2012. Also, SARS-CoV-2 lies close by to the group of SARS-CoV genome, and it has shown over

*Address correspondence to these authors at the Nano-Bio Laboratory, Special Centre for Nanoscience, Jawaharlal Nehru University, New Delhi-110067, India; Tel: 011-26704740/26704699;
E-mail: partima@mail.jnu.ac.in
Amity Institute of Applied Sciences, Amity University, Noida, Uttar Pradesh-201313, India; E-mail: mmaumita@gmail.com
National Institute of Immunology, New Delhi-110067, India;
E-mail: anilk@nii.ac.in
#*These authors contributed equally to this article.*



2666-7975/22 $65.00+.00 © 2022 Bentham Science Publishers

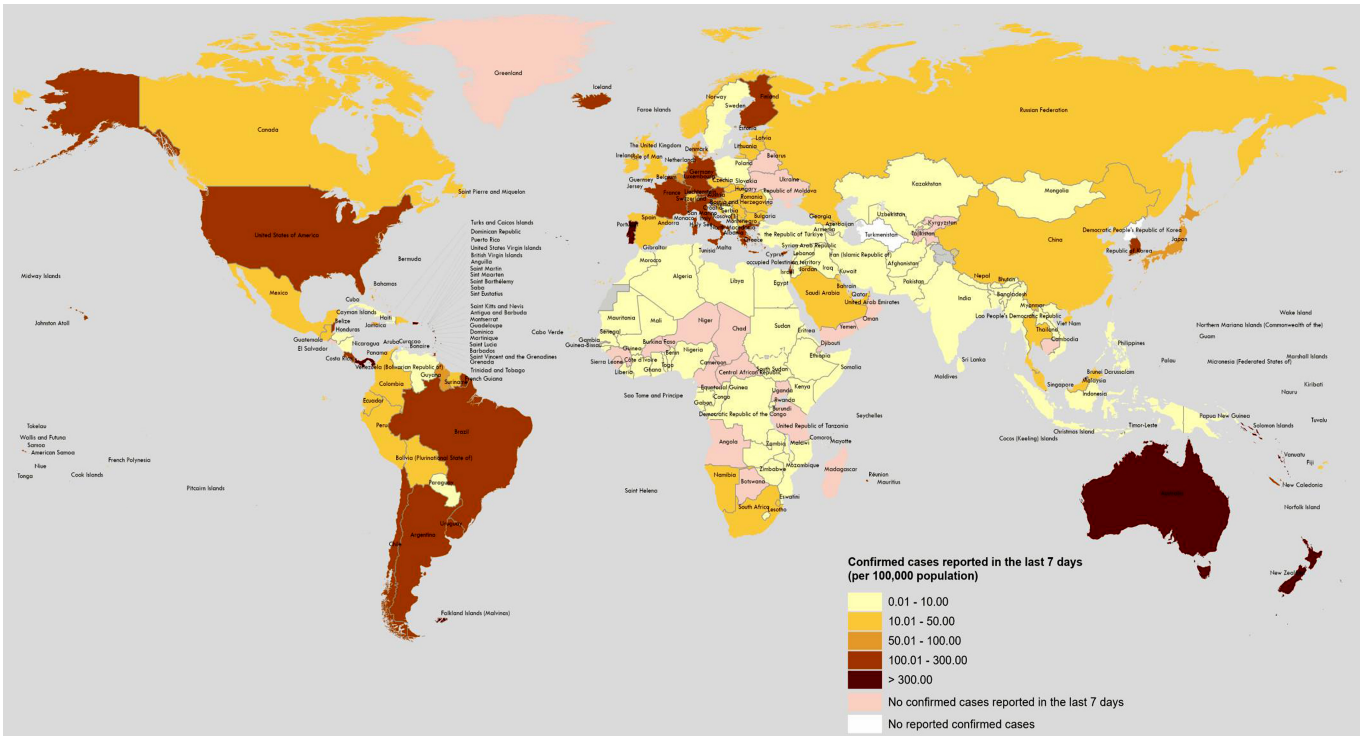



**Fig. (1).** The map depicts the global spreading of COVID-19 confirmed cases (reproduced image from WHO COVID-19 situation reports (Copyright 2020, WHO). (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

80% identical nature with the earlier reported human CoVs (SARS-like bat CoV) according to the phylogenetic tree depicted in Fig. (**2**) [8, 9]. Thus, bats are considered the critical reservoir of CoVs [10, 11].

Moreover, the real possibility of an intermediary animal may be snakes or pangolins and has been proved by Xu and his co-workers. He separated SARS-CoV-2 from the pangolin that showed potential intermediate host and had 99% high similarity with the CoVs, affecting human life. Fig. (**3**) depicts a schematic illustration of potential reservoirs and the mode of COVID-19 transmission [12-14]. But still, these results have not proven to be enough to confirm their possible COVID-19 intermediate and host.

Since this virus is transmitting rapidly across the globe, many countries struggle hard to deal with the casualties. This disease has affected people throughout the world; nonetheless, the future development of this virus is still not predicted. There is no information on any effective vaccine or treatment developed for treating COVID-19; hence, this disease is not curable until now. Moreover, there are no specific preventive measures for COVID-19, but the European Centre for Disease Prevention and Control (ECDC) and WHO have developed guidelines to break its transmission chain. Numerous characteristics of this transmitting virus in the incubation period before the appearance of symptoms, undefined features of the disease, long incubation period, transmission from asymptomatic healthy persons, and long duration of the illness have made it difficult to prevent this disease. If the transmission of COVID-19 is not controlled on time, this outbreak's ripple effect would bring potential challenges to global health systems and will have a long-lasting impact on the worldwide economy. Therefore, an essential requirement arises for controlling and eradicating alarming issues to save human life from this deadly virus.

COVID-19 has so far been managed using social distancing, face masks, and other transmission-mitigation techniques [15, 16]. These approaches minimize disease risk, but they do not provide resistance to SARS-CoV-2; thus, people are still at risk. Normal social and economic activities will be suspended until universal herd immunity is obtained. The reliance of human immunity on infection is untenable due to the advent of re-infection and the notion that just 1% fatality rate of COVID-19 will kill millions globally [17-19]. As a result, secure and reliable COVID-19 anti-viral medicines and vaccines are urgently needed. To advance the development of post-infection therapy, antiviral medications, and vaccines, the drug manufacturers, the Chinese Government, the United States Food and Drug Administration, European Medicines Agency, and WHO have partnered with industrial and academic researchers. ADP-ribose diphosphatase, S protein, RNA-dependent RNA polymerase, helicase, and papain-like protease are the most critical targets for SARS-CoV-2 [20]. This review will give a general overview of COVID-19, including genetic structure, clinical symptoms, pathogenesis, transmission route, diagnosis, and treatment of the SARS-CoV-2. In addition, we have briefly described the latest research development toward SARS-CoV-2 diagnostics, vaccines development, and potential therapeutic anti-viral drugs to summarize the current knowledge of COVID-19 disease. It may benefit readers to

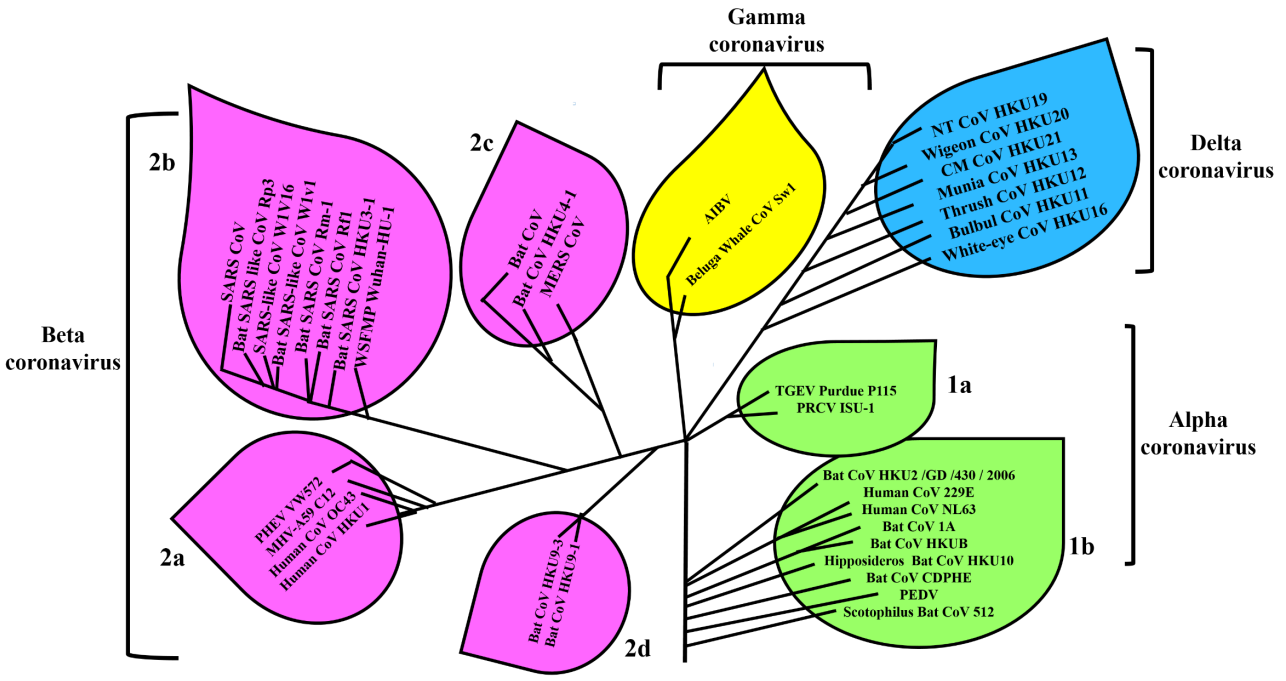



**Fig. (2).** The phylogenetic tree of CoVs, which has been classified into four genera (prototype shown): α-CoVs (yellow), β-CoVs (red), γ-CoVs (green) as well as δ-CoVs (blue), and their subclass are depicted as 1a and 1b for α-CoVs and 2a, 2b, 2c and 2d for the β-CoVs. This tree has been reproduced from the published tree of Coronavirinae. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

understand better and provide references for follow-up research, treatment, and prevention of this novel infectious disease.

## 2. VIRAL STRUCTURE AND ITS LIFE-CYCLE

CoVs have the largest 30 kb positive sense, the single-stranded RNA (ssRNA) genome has a spherical shape, and pleomorphic virus typically contains crown-shaped peplomers of 80-160 nm in size [21]. As the CoVs have RNA as their genetic material, two-thirds of the genome consists of RNA synthesis materials. The two large-sized non-structural polyproteins show no involvement in host response modulation (ORF1a-ORF1b) and encode viral polymerase (RdRp). The rest one–third part of the genome encodes three main structural proteins: the inner phosphorylated nucleocapsid protein (N), a different transmembrane glycoprotein (M), and the vast (200 K) glycoprotein S (for spike) constituting bulky peplomers (15-20 nm) on the surface of the viral envelope and a small transmembrane protein (E) [22], as shown in Fig. (**4**).

Through various transmission ways, the virus first gets into the respiratory system. It attacks the alveoli and attaches to the specific receptors on the type-2 pneumocyte, *i.e.*, ACE-2, by the specific spike proteins (S-spike) present. After binding with the ACE-2 receptor, the virus is engulfed by the cell, and then it releases its ss-RNA (+ve sense ssRNA) into the cytoplasm of the type-2 pneumocyte cell. After entering the host cells, it uses the host cell's ribosome to convert mRNA into protein and form a bunch of different poly-protein through translation. Further, this poly-protein makes all the components of the virus, *i.e.*, nucleic capsid, specific capsid enzyme, S-spikes proteins, *etc.*, with the help of proteinases enzyme, which breakdown poly-protein into different individual proteins needed for making whole viral structure. On the other side, it uses RNA-dependent RNA polymerase, *i.e.*, RDRP, to synthesize more copies of ssRNA from the ssRNA of the virus. Thus, all the viral components are assembled and then bud off from the type-2 pneumocyte making tons of virus particles and destroying the type-2 pneumocyte. When type-2 pneumocyte is damaged, it releases specific inflammatory mediators that activate macrophages.

Further, the activated macrophages secreted specific cytokine-like interleukin-1 (IL-1), IL-6, TNFα, *etc.*, which comes into the bloodstream, causes endothelial cells to dilate, and increases capillary permeability to release all the inflammatory mediators, which lead to vasodilation. Due to this, the plasma fluid starts flowing out, leaking into the interstitial spaces and potentially into the alveoli as all the fluid accumulates outside the alveoli leading to its compression. Finally, alveoli is collapsed due to increased surface tension and collapsing pressure. Due to the alveolar collapse, tons of fluid accumulating around this alveolus impairs the alveolar membrane and the respiratory membrane (allowing gas exchange), leading to hypoxemia. Hypoxemia is a hazardous condition, especially when it becomes a refractor and results in multisystem organ failure (MSOF). The whole process of

the virus life cycle and its mode of action is described in Fig. (**5**).

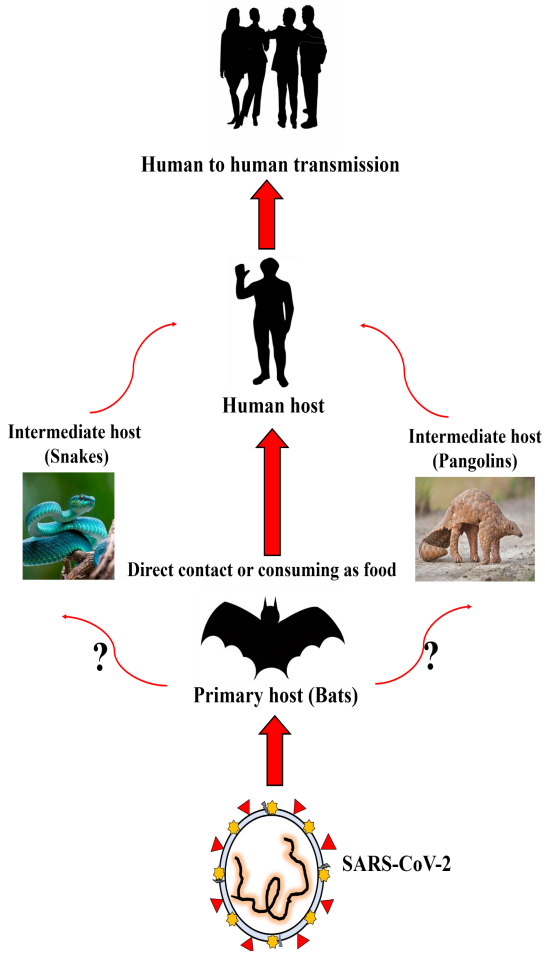



**Fig. (3).** Schematic illustration showing the CoVs potential reservoirs and transmission mode [14]. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

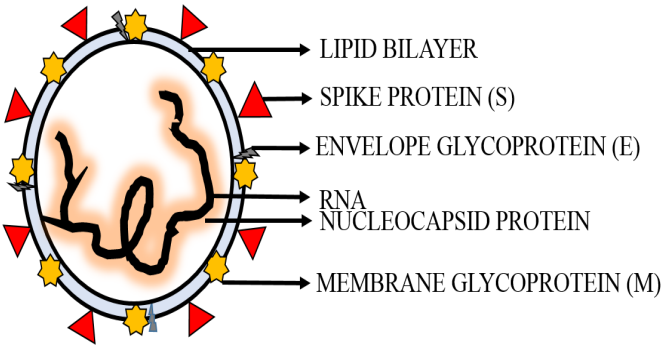



**Fig. (4).** This diagram shows the structure of a CoV virion along with the supercoiled nucleocapsid. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

The effect of different geographies, temperatures, and latitudes has also played an essential role in transmitting the virus worldwide. This situation seems worrisome as COVID-19 is enlarging its deadly roots rapidly to each corner of the world. There has been consistent transmission in the west and east at the community level. The new epicenters of the virus were roughly around the 30-50˚N" zone; Iran, Northern Italy, South Korea, and Japan [23]. In January and February 2020, *i.e.*, in Wuhan and other affected countries, there was a striking similarity in the measures of relative humidity (R.H., 47-79%) and average temperature (5-11˚C) [23] and general laboratory conditions that are conducive to CoVs survival (4˚C and 20-80 % R.H.) [24]. Also, none of the affected cities have minimum temperatures going below 0˚C. Therefore, the similarities between temperatures and the towns affected by COVID-19 deserve special attention. Human CoVs usually show symptoms like the common cold, indicate strong winter seasonality from April to December, and are unnoticed in the temperate regions during the summer months [25]. Recently, an integrated sequence-based analysis of SARS-CoV-2 genomes from various geographical locations (mainly from the USA, Wuhan, India, Nepal, and Italy) was performed by Rahila Sardar and his co-workers to find the distinct traits lacking in SARS-CoV and other related CoVs family genomes, discussing the facilitation of communication, distinctive infection, immunogenic characteristics, and virulence of the virus. They observed that all the genomes (countries mentioned above) share 99 % similarity with Wuhan (SARS-CoV-2) genome, excluding the genome from Nepal as it shares 100 % similarity with the Wuhan genome. They also confirmed that all the SARS-CoV-2 genomes are under the SARS-CoV clade, which shares 83.3672 % amino acid and 79.5772 % nucleotide identity with the SARS-CoV genome.

Further, they recognized numerous distinctive features of the SARS-CoV-2 genome, including a novel mutation in the spike surface glycoprotein in the Indian SARS-CoV-2, lacking in other strains. They also exposed nine host micro RNAs (miRNAs), which can only target SARS-CoV-2 genes without targeting SARS and MERS genomes. Similarly, hsa-miR-27b is the only miRNA with a target gene in the Indian SARS-CoV-2 genome [26]. Furthermore, this overabundance of information is utilized for vaccine and drug research developments.

## 3. PATHOGENESIS AND CLINICAL SYMPTOMS

The spread of the SARS-CoV-2 virus can occur through the symptomatic or asymptomatic person coming in contact with healthy persons or touching an infected surface. It can cause infection in two ways: first is through fecal contamination, and the other one is by coughing, sneezing, *i.e.*, airborne mode, or respiratory droplet. A respiratory droplet can spread from an affected person to a healthy person within 3-6 feet. Also, respiratory droplets could stay on surfaces and survive for approximately 24 hours. If someone goes and touches that surface and then touches one of their mucous membranes, they can inoculate themselves. Another prominent thing is that these can also be asymptomatic carriers, so sometimes, if someone is not showing symptoms, they do

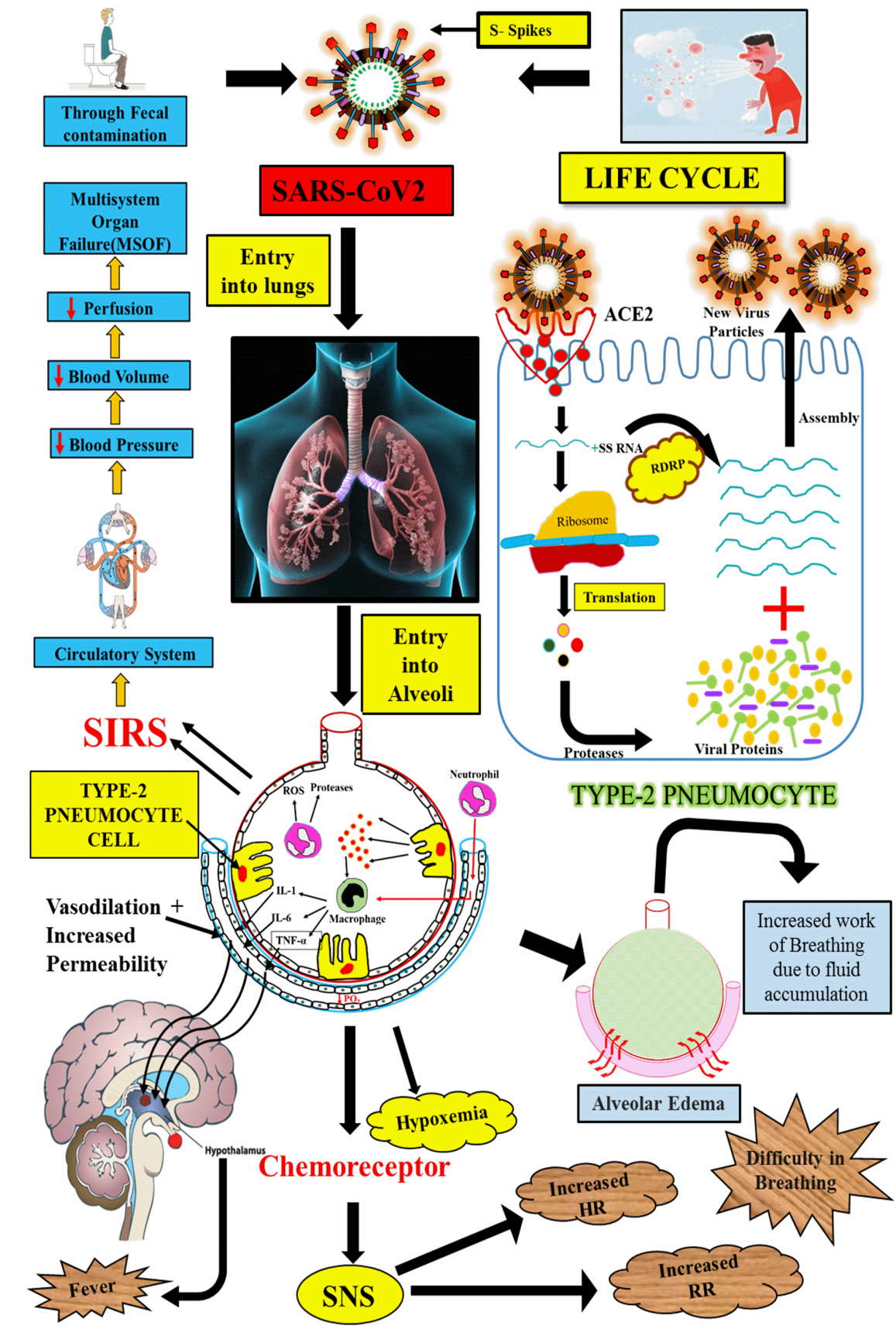



**Fig. (5).** The life cycle of SARS-CoV-2 in host cells and its mode of action leading to multisystem organ failure. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

potentially have the ability to spread these CoVs, which is very alarming.

After the 5-day incubation period, symptoms of CoV illness arose [27]. The time from the onset of COVID-19 symptoms to mortality varies between 6 and 41 days, with a mean of 14 days [28]. Fever, tiredness, and cough are common signs between COVID-19 and previous β-CoVs, in addition to hemoptysis, diarrhoea, lymphopenia, sputum production, dyspnoea, and headache. COVID-19, on the other hand, showed specific clinical features associated with a lower airway infection, as evidenced by symptoms of the upper respiratory tract like rhinorrhoea, sneezing, diarrhoea, and sore throat. Infected individuals with the SARS-CoV-2 virus have irregular respiratory findings, pro-inflammatory cytokines in their plasma, and leukocytes. According to one case report, on the fifth day of illness, the COVID-19 patient had a body temperature of 39.0 °C, cough, and coarse breathing sounds in both lungs. Ground-glass opacities, severe pneumonia, immediate cardiac damage, and RNAaemia that led to mortality were notable pathogenesis of COVID-19 infection focusing on the respiratory system. In addition, individuals with COVID-19 infection had increased cytokines and levels of chemokines in blood like GMCSF, MIP1, IFN MIP1, MCP1, IP10, IL8, VEGFA GCSF, IL9, basic FGF2, PDGFB, TNF, IL10, IL7, and IL1RA. As a result, appropriate methodologies must be developed immediately to generate therapies to decrease and prevent disease spread by distinguishing the various transmission routes from feces and urine samples.

## 4. DIAGNOSTIC TESTING

The clinical characteristics of COVID-19 show variation from asymptomatic to acute respiratory syndrome and leading to death in some cases. COVID-19 can affect all age groups; however, casualties are reported in the older population (50-70 % of mortality rate), while mild cases are described in the case of infants, neonates, and children [29]. Therefore, accurate and rapid detection is essential to control outbreaks of COVID-19 disease in hospitals and at the community level [30]. Recently, some tests are available for diagnosing CoVs that comprise reverse transcription loop-mediated isothermal amplification, *i.e*., RT-LAMP, real-time reverse-transcription polymerase chain reaction, *i.e*., rRT-PCR, and reverse-transcription polymerase chain reaction, *i.e*., RT-PCR [31, 32]. The laboratory examination of the China National Health Commission revealed that oropharyngeal and nasopharyngeal swab tests have become standard criteria for COVID-19 detection.

### 4.1. RT-PCR Based Test

RT-PCR is a technique used to detect the presence of specific genetic material in the sample. To test COVID-19, the swab collected from the patient's mouth or nose is treated with various chemical substances to remove fats and protein and isolate RNA. This technique provides positive results in 2 to 8 days [33]. The schematic view of a detailed procedure of RT-PCR is shown in Fig. (**6**). This technique was adopted to detect two diverse (N and ORF1b) SARS-CoV-2 genomes [34]. Additionally, three novel RT-PCR assays were also developed for targeting the spike (S), nucleocapsid (N) genes, and RdRp/helicase (Hel) of SARS-CoV-2. In all the three novel assays, RdRp/Hel assay showed highly sensitive, specific, and lowest detection limits *in vitro*, thus improving the laboratory diagnosis for COVID-19 detection [35]. The automated results from molecular diagnostics help handle many patient samples to keep pace with fluctuating requirements [36-38]. The required molecular analysis performance to detect SARS-CoV-2 was witnessed with minimal hands-on time and reliable and fast results [39, 40]. Also, the patient suffering from COVID-19 has to be analyzed with typical chest computerized tomography (C.T.) characteristics after the outcome of RT-PCR comes negative if the patient in recent exposure suffers from coughing, dyspnea, fever, fatigue, or sore throat [41].

### 4.2. Antibody-based Test

There have been 11 point-of-care (POC) diagnostics for COVID-19 described so far, most of which are antibody and molecular-based testing. Numerous researchers and companies are working on developing antibody tests, such as colloidal gold immunoassays, *i.e*., Assay Genie POC kit as well as VivaDiag COVID-19 IgG-IgM test, ELISA, lateral flow immunoassays, *i.e*., BioMedomics rapid test as well as Surescreen rapid test cassette, and, time-resolved fluorescence immunoassays, *i.e*., Goldsite diagnostics kit and so on [42]. This test is straightforward in which a small amount of blood sample collected from the patient is added to a test tube coated with a virus “mash” or purified viral components. If the patient is infected and carries antibodies against the virus, they will get bound to the test tube. The test tube is later developed to see if any antibodies are present and if present, it confirms the infection. The first test for COVID-19 measured the antibodies in the blood, authorized by the U.S. Food and Drug Administration (FDA). This test could identify the people who recovered from COVID-19 as they have immunity against COVID-19 by developing a vaccine. These blood tests search for the particular protein, *i.e*., antibodies generated against CoVs, not for other common diseases or seasonal flu.

### 4.3. Other Diagnostic Developments

Last year, the shortage of commercial kits and time-consuming testing procedures resulted in a delay in diagnosing COVID-19 infection. Recently, Mylab Discovery Solutions Pvt. Ltd., a Pune-based molecular diagnostics company specializing in molecular diagnostic kits, has fabricated the first made-in-India test kit for COVID-19 in a record period of six weeks. Mylab is the only Indian company to attain specificity and 100 % sensitivity even in early-stage detection by the ICMR evaluation with a reasonable and cost-effective price. This kit is made according to the guidelines of WHO/CDC and with the support of central and local governments that have emphasized ‘Make in India.’ Moreover, Mylab’s COVID-19 test kit could detect the virus within 2.5 h compared to other developed kits that usually took more than seven h for the same.

In another work, scientists from India have successfully developed an inexpensive, paper-strip test that utilized the CRISPR-C as a nine cutting edge gene-editing tool for de-

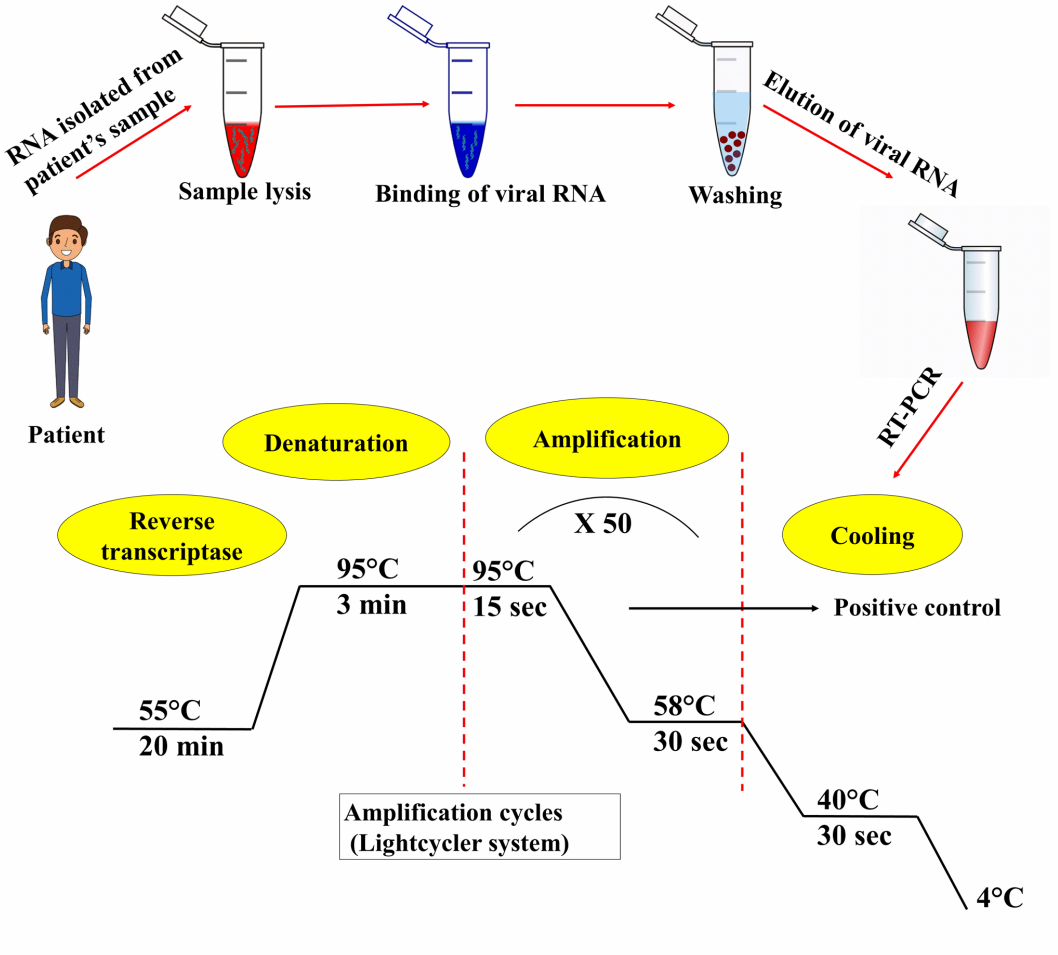



**Fig. (6).** Schematic illustration of the protocol for COVID-19 detection using RT-PCR. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

tecting COVID-19 disease within an hour in the samples of suspected people. The kit bore a resemblance to a portable paper-strip test utilized to confirm pregnancy, did not require various particular abilities to carry out the analysis, and is reasonably less sophisticated. It does not depend upon the expensive RT-PCR devices for RNA isolation, DNA conversion, and amplification. Rapid testing would undoubtedly be the critical method to ensure prompt isolation of positive cases to contain the infection from spreading out fast. Many test kits based on molecular and antibody-based techniques were developed in different countries, listed in Table **1** [42].

## 5. TREATMENT

The pandemic of COVID-19 has drawn significant global attention caused by SARS-CoV-2. As a result, scientists and clinicians worldwide have been studying SARS and MERS to learn more about this new emerging disease and its epidemiology to help the development of COVID-19 prevention and treatment drugs. The research community reacted quickly with a possible list of existing COVID-19 therapeutic drugs based on genomic sequence data coupled with protein structure modeling. Besides the currently available treatments for COVID-19 and its problems, scientists are actively developing new promising therapeutic approaches, such as protease inhibitors, vaccines, interferon-based therapies, peptides, monoclonal antibodies, and small molecule drugs to alleviate the pandemic. Based on the earlier experience of treating patients with SARS-CoV and MERS-CoV, some antivirals like ribavirin, ritonavir, and lopinavir have been used [43]. Intravenous immunoglobulin, chloroquine, hydroxychloroquine, arbidol (anti-viral drug accessible in Russia and China), patient's plasma from CoVs, and interferons are other drugs proposed for its therapy [44-46].

It is critical to identify the targets in order to recognize the medications that can be used to treat the COVID-19 infection. Table **2** shows the list of some of the promising target candidates and their role in causing a viral infection [47]. PLpro and 3CLpro are two viral proteases that target candidates for splitting peptides of a virus to active subunits for virus replication and assembly inside infected cells. Another target, RdRp, is RNA polymerase, which is accountable for synthesizing RNA in the virus. The interaction of structural viral protein S with the receptor ACE-2 present in the host cells results in endocytosis of the virus into the cells. It can also act as a feasible drug target. Moreover, the protease

**Table 1. List of some FDA-approved Diagnostic Kits and their details.**

| Product | Manufacturer/ Location | Sample Type | FDA Approval | Time to Result | Storage Temperature (°C) | Type | Target |
|---|---|---|---|---|---|---|---|
| **Molecular Technique-based Point-of-care Diagnostics** | | | | | | | |
| Xpert SARS-CoV-2 | Cepheid (US/Worldwide distribution) | Nasopharyngeal swab, nasal aspirate | Yes | 45 mins - 1hr | 2-28 | RT-PCR | SARS-CoV-2 |
| VitaPCR COVID-19 assay | Credo (Singapore) | Nasopharyngeal or oropharyngeal swabs | pending | 20 mins | 15-30 | RT-PCR | SARS-CoV-2 |
| RapiPrep COVID-19 | Microsens Dx (London) | Sputum or swabs | Yes | 30 mins | - | Lamp amplification technology | SARS-CoV-2 |
| ePlex SARS-CoV-2 | GenMark Diagnostics (United States) | Nasopharyngeal swab | Yes | - | 2-8 | RT-PCR | SARS-CoV-2 RNA |
| Accula SARS-CoV-2 | Messa Biotech (United States) | Throat and nasal swabs | Yes | 30 mins | 15-30 | RT-PCR+ lateral flow | SARS-CoV-2 RNA |
| ID NOW COVID-19 | Abbott Diagnostics (Worldwide) | Throat, nasal, nasopharyngeal, and oropharyngeal swabs | Yes | 13 mins | 15-30 | Isothermal nucleic acid amplification | SARS-CoV-2 nucleic acid |
| **Antibody-based POC Diagnostics** | | | | | | | |
| GT-100 SARS-CoV-2 IgG/IgM kit | Goldsite Diagnostics Inc. (China) | Human serum and plasma (20 uL) | - | 12 mins | - | Time-resolved fluorescence Immunoassay | IgG/IgM |
| rapid POC kit | Genie (Acro Biotech, Inc) (Ireland) | Blood, serum, and plasma | - | 15 mins | 2-30 | Colloidal gold Immunochromatography | IgG/IgM |
| COVID-19 IgM-IgG Rapid Test | BioMedomics, BD (United States) | Finger prick / venous blood | - | 15 mins | RT | Lateral flow immunoassay | IgG/IgM |
| COVID-19 Rapid Test Cassette | SureScreen Diagnostics (England) | Finger prick | - | 1-15 mins | 2-30 | Lateral flow immunoassay | IgG/IgM |
| VivaDiag COVID-19 IgG - IgM test | VivaChek (China) | Finger-prick/ venous blood, plasma, or serum | - | 15 mins | 2-30 | Colloidal gold Immunochromatography | IgG/IgM |

**Table 2. Critical proteins and their role in causing a viral infection.**

| Target Candidate | Full Name | Role During Viral Infection | Refs. |
|---|---|---|---|
| AT2 | angiotensin AT2 receptor | A critical effector involved in controlling the blood volume and pressure of the cardiovascular system | [47] |
| 3CLpro | CoVs main protease 3CLpro | A protease for converting the viral polyprotein into functional units 3CLpro by proteolysis | [48, 49] |
| PLpro | papain-like protease PLpro | A protease for converting the viral polyprotein into functional units by proteolysis | [48, 49] |

**Table 2 Contd….**

| Target Candidate | Full Name | Role During Viral Infection | Refs. |
|---|---|---|---|
| ACE-2 | angiotensin-converting enzyme 2 | A protein present as a viral receptor on the host cells for binding to viral S protein | [47, 50, 51] |
| S-protein | viral spike glycoprotein | A viral protein present on the surface to bind with the host cell receptor ACE-2 | [47, 50, 51] |
| TMPRSS2 | transmembrane protease, serine 2 | A protease produced by the host cell to primes S protein for facilitating its binding to ACE-2 | [52] |
| RdRp | RNA-dependent RNA polymerase | To replicate the genome of the virus | [47, 48, 53-55] |

TMPRSS2 plays a significant role in the proteolysis of S protein produced by the host cells, enabling it to bind with the ACE-2 receptor in human cells.

### 5.1. Antiviral Drugs

COVID-19 has been treated using a variety of nucleoside analogs, which are FDA-approved antiviral medicines. RdRp is aimed at nucleoside analogs that are inhibitors of RNA synthesis, which decrease viral RNA replication by blocking nucleotide synthesis and causing mutations [56]. Inhibition of viral replication is caused by Type I interferons in various ways, which are antiviral cytokines by nature [57]. Interferons, along with nucleoside analog combinations, can manage COVID-19.

Ritonavir and Lopinavir are anti-HIV-approved remedies that can prevent the protease activity of CoVs in animals, and *in vitro* were connected to reducing the death rate [58]. They have also been proven to have antiviral properties versus SARS and MERS. Nelfinavir, a selective HIV inhibitor that is protease in nature, has been demonstrated to mitigate SARS-CoV, indicating that it could be used to treat COVID-19 [59].

Ribavirin is an antiviral compound (a guanosine analog) used to treat numerous virus infections, including hepatitis C, hemorrhagic fevers, and respiratory syncytial virus. Additionally, ribavirin targets the SARS-CoV-2 RdRp model after docking, modeling, and sequence analysis [60], making it a more potent antiviral drug against SARS-CoV-2. In culture, it synergistically inhibits SARS-CoV with IFN-b [61, 62]. It increases the inborn antiviral response beside SARS-CoV-2 in conjunction with pegylated interferon. In China, COVID-19 patients were medicated with IFN-a in conjunction with ribavirin administered twice a day *via* vapor aspiration [63].

Prophylactic efficacy of remdesivir has been clinically observed, and *in-vitro*, it has superior antiviral efficacy over lopinavir and ritonavir [64]. Remdesivir, a broad-spectrum antiviral drug, targeted the Ebola virus [65] and numerous other RNA viruses to compete for RdRp [64]. In a recent case report, aemdesivir and arbidol, ritonavir/lopinavir, and interferon b were related to noteworthy improvements and operational in treating patients suffering from COVID-19 [66]. This drug was primarily utilized in the United States for treating the first patient with COVID-19, thereby improving the patient's clinical condition after treatment of one day only [67]. The China-Japan Friendship Hospital and the University of Nebraska Medical Center conducted phase III and II clinical trials of Remdesivir, which demonstrated that Remdesivir decreased recovery durations and reduced infectivity in the experimental group. The United States Food and Drug Administration (FDA) has granted Gilead Sciences, Inc., USA, urgent usage permission (EUA) for the experimental antiviral Remdesivir in COVID-19 tereatment [68-70].

Barcitinib drug was also proposed due to its potential capacity to diminish viral entrance and its anti-inflammatory consequences [71]. Furthermore, Toyama Chemical of Japan originally developed favipiravir, a purine nucleoside analog known for erroneous viral RNA synthesis [72], which has recently got approval for clinical trial as a drug to treat this novel virus [49]. In COVID-19 patients, favipiravir was utilized with other antiviral medicines such as baloxavir. Glenmark Pharmaceuticals Ltd. began phase III clinical studies on the antiviral favipiravir in May 2020, after receiving consent from India's drug regulator DCGI in mid-April 2020 [73]. The trials are expected to be finished in July-August 2020.

Azithromycin is a macrolide antibiotic that suppresses the 50S ribosomal subunit, affecting protein synthesis in Gram-negative and Gram-positive bacteria, giving it an antibacterial tag. Despite the lack of clear evidence of azithromycin's effectiveness in treating COVID-19, few scientific groups have claimed that azithromycin's antibacterial characteristics are still beneficial in the experimental therapy of community-acquired pneumonia in patients suffering from COVID-19 [74]. It can improve the hydroxychloroquine potency in treating severe COVID-19 patients [75], and a mixture of azithromycin and hydroxychloroquine could be a promising solution to remdesivir for treatment of COVID-19 illness.

Ivermectin, an FDA-approved antiparasitic drug that has recently shown antiviral action *in vitro* against a broad spectrum of viruses [76, 77], also reduces SARS-CoV-2 multiplication *in vitro* [78]. A research group at Australia's

Monash Biomedicine Discovery Institute attempts to determine the ivermectin role in managing COVID-19. Thus, assessing these antiviral drugs in terms of safety and efficiency for COVID-19 detection needs further clinical trials. WHO has registered 34 trials of antiviral drugs in patients suffering from COVID-19 pandemic. Table **3** summarises the current drugs with therapeutic abilities for COVID-19 (Drug Repurposing) [79-100].

### 5.2. Vaccine and Vaccination Strategies for COVID-19

Vaccines are one of the most important advancements in human healthcare for reducing the prevalence and effect of infectious diseases.Vaccines are often used to control epidemics and are an excellent way to lower pandemic fatality [101]. The COVID-19 vaccine was developed at a faster rate than any preceding vaccine. However, due to the urgent need researchers generated vaccines that could be used to protect the citizens from SARS-CoV-2 [102]. Around 1.3 billion vaccination doses had been circulated globally as of 29$^{th}$ September 2021, having an average of 17 doses/ 100 people (Coronavirus COVID-19 (2019-nCoV) (arcgis.com)) (Fig. **7**). Even though vaccinations are still in short supply worldwide, the mainstream governments have focused their timely immunization campaigns on high-risk populations, including the clinically vulnerable, persons in their 60s, 70s, older, and front-line workers like nurses and doctors [103]. There have been few reported side effects from clinical trials of the SARS-CoV-2 vaccine at this hour. The effectiveness of vaccines that have been properly investigated appears to be promising.Vaccination programs have a crucial role in spreading SARS-CoV-2, which is more common in younger age groups, limiting the spread of the virus and possibly defending those who are most susceptible [103, 104]. The diversity of technologies being evaluated in the present production system of the COVID-19 vaccine is noteworthy, comprising live attenuated nucleic acids and protein vaccines. A total of 257 vaccines are recently in the development stage, out of which 78 are in clinical testing, following the tracker of the COVID-19 vaccine (https://www.covid-19vaccinetracker.org). Most COVID-19 vaccine candidates are intended to stimulate anti-viral S protein neutralizing antibodies, avoiding infection by inhibiting absorption through the human ACE-2 receptor [105].

### 5.3. Leading Vaccines

National regulatory bodies have awarded use authorizations to 22 COVID-19 vaccines as of September 2021, with eight vaccinations licensed for full utilization and six vaccines in limited or early use (as described in Table **4**). The WHO Strategic Advisory Group of Experts on Immunization has determined that these vaccines are safe and effective [106].

#### *5.3.1. The Pfizer-BioNTech*

The Pfizer-BioNTech vaccine is an mRNA vaccine administered intramuscularly in 2 dose series after 21 days of a time interval. According to published clinical trial findings, the Pfizer-BioNTech vaccine is 95 percent effective in preventing laboratory-confirmed COVID-19 disease in patients who were not affected by the virus previously [107].

#### *5.3.2. The Johnson and Johnson*

The Johnson & Johnson vaccine, on the other hand, is a viral vector vaccine that may be given intramuscularly in a single dosage. According to published clinical trial results, two weeks after vaccination, the Johnson & Johnson vaccine is found to be 66.3 percent effective in preventing laboratory-confirmed COVID-19 in people who have never had the virus [107].

#### *5.3.3. AstraZeneca Vaccine*

The University of Oxford and its spin-off business, Vaccitech, co-invented the AstraZeneca vaccine in particular. It is a weak variant of a common cold virus, *i.e*., adenovirus, which infects chimps and contains the genetic material of SARS-CoV-2 virus-containing spike protein. The vaccination is administered intramuscularly in two doses over four weeks. The AstraZeneca vaccine was 79% and 100% effective in averting symptomatic COVID-19 and hospitalization and severe disease in phase III clinical trial, respectively [107].

#### *5.3.4. Moderna Vaccine*

The US BARDA and NIAID developed the COVID-19 Moderna vaccine. It is a lipid nanoparticle-encapsulated mRNA-1273 vaccine given intramuscularly in two doses after 28 days. The phase III clinical research found that controlling COVID-19 sickness, including severe disease, was 94.1 percent effective [108].

#### *5.3.5. Sinopharm BIBP Vaccine*

The BIBP and CNPG (Sinopharm) developed a COVID-19 vaccine called Sinopharm BIBP. It is a viral vaccination that is inactivated and given intramuscularly in a 2-dose series after 21 days. The efficacy of COVID-19 in symptomatic and hospitalized patients was reported to be 79 percent in phase III clinical trial [107].

#### *5.3.6. Covaxin Vaccine*

In collaboration with the Indian Council of Medical Research (ICMR) and the National Institute of Virology, Bharat Biotech Company produced Covaxin, India's first vaccine (NIV). The vaccine contains inactivated SARS-CoV-2 virus administered intramuscularly over 28 days in a two-dose series. After the 2$^{nd}$ dose, the National Library of Medicine (U.S.) revealed that clinical trial data of Covaxin phase III showed 81 percent short-term efficacy in avoiding COVID-19 disease in patients without past infection [107].

#### *5.3.7. Sputnik V Vaccine*

Sputnik V is Russia's 1$^{st}$ vaccine against COVID-19 to be authorized. It is a DNA-based adenoviral vaccination. The efficacy of Sputnik V in a phase III trial against COVID-19 disease patients was reported to be 91.6 percent [109]. Table **2** summarises the vaccines now in clinical studies.

Long-term storage of vaccines is essential for preserving their stability and immunogenicity. BioNTech/Pfizer and Moderna, two promising vaccines of COVID-19 mRNA, are to be kept in ultra-cold settings at 18°C and 70°C, respectively. Owing to an absence of cold-chain infrastructure, the BioNTech/Pfizer vaccine's ultra-cold storage would be

**Table 3. Current Drugs with therapeutic abilities for COVID-19 (Drug repurposing).**

| Drug Agent | Target | Probable Process for COVID-19 Actions | Infection Sign | Trialled for COVID-19? | Refs. |
|---|---|---|---|---|---|
| **Under Clinical Trial/Re-purposed/off Label** | | | | | |
| Barcitinib | JAK kinase | A JAK inhibitor that may interface with the inflammatory processes | Approved drug for rheumatoid arthritis | Yes | [71] |
| Lopinavir | Viral protease:3CL pro or PLpro | Protease inhibitors that may inhibit the viral protease:3CLpro or PLpro | Lopinavir and ritonavir are approved drug combinations for HIV infection | Yes | [48] |
| Ritonavir | - | - | - | Yes | [48, 79] |
| Darunavir | - | - | Approved drug for HIV infection | Yes | [47] |
| Favipiravir (favilavir) | RdRp | A purine nucleoside that acts as an alternate substrate leading to inaccurate viral RNA synthesis | Viral infections | Yes | [54, 72] |
| Remdesivir | - | A nucleotide analog that may block viral nucleotide synthesis to stop viral replication | Ebola virus infection | Yes | [48, 54, 80] |
| Ribavirin | - | - | RSV infection, hepatitis C, some viral hemorrhagic fevers | Yes | [49, 53, 54] |
| Galidesivir | - | - | Hepatitis C, Ebola virus, Marburg virus | Yes | [81] |
| BCX-4430 (salt form of galidesivir) | - | - | Hepatitis C, Ebola virus, Marburg virus | Yes | [81] |
| Nitazoxanide | N/A | A drug that may inhibit viral protein expression | Various helminthic, protozoal, and viral infection-caused diarrhea | Yes | [54] |
| Corticosteroids | - | - | Steroid hormones | Yes | [82] |
| Ritonavir + lopinavir | - | - | - | Yes | [83] |
| Ribavirin + ritonavir + lopinavir | | | - | - | [84, 85] |
| Darunavir | Antiretroviral, protease inhibitor. Used with low doses of cobicistat to increase bioavailability and half-life | - | HIV infection | Yes | [84, 86] |
| Emtricitabine + tenofovir | Non-nucleoside reverse transcriptase inhibitor + nucleotide reverse transcriptase inhibitor | - | HIV infection | Yes | [87] |

**(Table 3) Contd….**

| Drug Agent | Target | Probable Process for COVID-19 Actions | Infection Sign | Trialled for COVID-19? | Refs. |
|---|---|---|---|---|---|
| Ruxolitinib | Myelofibrosis and polycythemia vera treatment | - | Myelofibrosis and polycythemia vera | Yes | [88] |
| Sirolimus | mTOR inhibitor IL2, immunosuppress sant | - | Anti-rejection medicine in those aged >=13 who received a kidney transplant. Also used to treat LAM | - | [89] |
| IFN + ribavirin | Combination antiviral + proteins made and released by host cells | - | - | - | [90, 91] |
| IFN + ribavirin + steroids | Combination of proteins made and released by host cells + antiviral + steroid hormones | - | - | - | [90] |
| Lopinavir + ritonavir + IFN + ribavirin | combination of protease inhibitor + proteins made and released by host cells + antiviral | - | - | - | [92, 93] |
| Baloxavir marboxil | Antiviral (endonuclease inhibitor) | - | In the U.S., licensed for uncomplicated acute influenza, and in Japan for all influenza | Yes | [94] |
| **Licensed in the Country of Origin for other Diseases** | | | | | |
| Favipiravir | Experimental antiviral drug. Pyrazinecarboxamide derivative viral RNA polymerase inhibitor | - | Influenza (licensed in Japan) | Yes | [95] |
| Enisamium iodide | - | - | Antiviral on the market in Ukraine | - | [96, 97] |
| Arbidol | S protein/ACE2 | An inhibitor that may disrupt the binding of viral envelope protein to host cells and prevent viral entry to the target cell | Influenza antiviral drug | Yes | [47, 51] |
| Novaferon | Recombinant protein produced by DNA-shuffling of IFN-α | - | Licensed in China hepatitis B | Yes | [98] |
| Phase 2/Phase 3/Observational | | | | | |
| GS5734/Remdesivir | Nucleoside Inhibitor | - | N/A | Yes | [67, 99] |
| Alferon | natural, human interferon alpha protein | - | N/A | - | [100] |

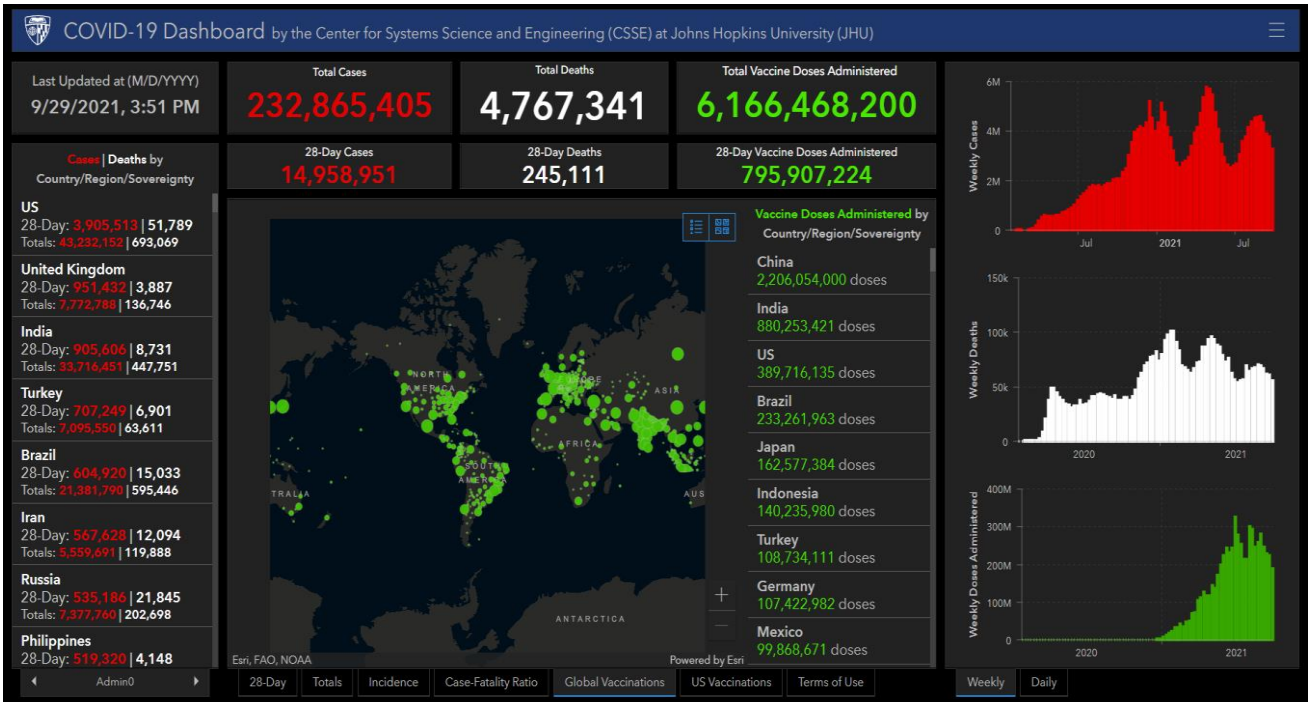



**Fig. (7).** Share of the population fully vaccinated (have received all doses prescribed by the vaccination protocol) against COVID-19, September 29, 2021. Retrieved from: Coronavirus COVID-19 (2019-nCoV) (arcgis.com) [Online Resource]. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

**Table 4. Landscape list of authorized/approved COVID-19 vaccines.**

| S. No. | Vaccine Type | Vaccine Name | Developers | Country of Origin |
|---|---|---|---|---|
| 1. | mRNA | Comirnaty | Pfizer, BioNTech; Fosun Pharma | Multinational |
| 2. | mRNA | Moderna COVID-19 Vaccine | Moderna, BARDA, NIAID | US |
| 3. | Adenovirus | COVID-19 Vaccine Astra-Zeneca; also known as Vaxzevria and Covishield | BARDA, OWS | UK |
| 4. | Inactivated vaccine | Covaxin | Bharat Biotech, ICMR; Ocugen; ViroVax | India |
| 5. | Recombinant adenovirus | Sputnik V | Gamaleya Research Institute, Acellena Contract Drug Research and Development | Russia |
| 6. | Recombinant adenovirus | Sputnik Light | Gamaleya Research Institute, Acellena Contract Drug Research and Development | Russia |
| 7. | Non-replicating viral vector | COVID-19 Vaccine Janssen | Janssen Vaccines (Johnson & Johnson) | The Netherlands, US |
| 8. | Inactivated vaccine (formalin with alum adjuvant) | CoronaVac | Sinovac | China |
| 9. | Inactivated vaccine | BBIBP-CorV | Beijing Institute of Biological Products; China National Pharmaceutical Group (Sinopharm) | China |
| 10. | Peptide vaccine | EpiVacCorona | Federal Budgetary Research Institution State Research Center of Virology and Biotechnology | Russia |

**(Table 4) Contd….**

| S. No. | Vaccine Type | Vaccine Name | Developers | Country of Origin |
|---|---|---|---|---|
| 11. | Recombinant vaccine (adenovirus type 5 vector) | Convidicea | CanSino Biologics | China |
| 12. | DNA vaccine (plasmid) | ZyCoV-D | Zydus Cadila | India |
| 13. | Protein subunit vaccine | MVC-COV1901 | Medigen Vaccine Biologics Corp.; Dynavax | Taiwan |
| 14. | Conjugate vaccine | Soberana 02/Soberana Plus | Finlay Institute of Vaccines; Pasteur Institute | Cuba, Iran |
| 15. | Protein subunit vaccine | Abdala | Center for Genetic Engineering and Biotechnology | Cuba |
| 16. | Inactivated vaccine | Unnamed vaccine | Chinese Academy of Medical Sciences, Institute of Medical Biology | China |
| 17. | Inactivated vaccine | COVIran Barekat | Shifa Pharmed Industrial Group | Iran |
| 18. | Inactivated vaccine | Unnamed vaccine | Minhai Biotechnology Co.; Kangtai Biological Products Co. Ltd. | China |
| 19. | Inactivated vaccine | QazVac (QazCovid-in) | Research Institute for Biological Safety Problems | Kazakhstan |
| 20. | Recombinant vaccine | ZF2001 | Anhui Zhifei Longcom Biopharmaceutical, Institute of Microbiology of the Chinese Academy of Sciences | China, Uzbekistan |
| 21. | Inactivated vaccine | CoviVac | Chumakov Federal Scientific Center for Research and Development of Immune and Biological Products | Russia |
| 22. | Inactivated vaccine | WIBP-CorV | Wuhan Institute of Biological Products; China National Pharmaceutical Group (Sinopharm) | China |

(Source: https://www.nytimes.com/interactive/2020/science/coronavirus-vaccine-tracker.html.
https://www.raps.org/news-and-articles/news-articles/2020/3/covid-19-vaccine-tracker.
Emergency use. Early use. Stopped use.)

impractical for middle and low -income countries. AstraZeneca's vaccine, in contrast, can be kept, transported, and held for at least six months at standard refrigerated temperatures (2–8 °C). Furthermore, AstraZeneca's vaccine is relatively inexpensive (about $2-4 for each dosage) compared to other vaccines. As a result, it may be produced for a wider population and distributed in developing nations utilizing existing medical facilities.

## CONCLUSION

COVID-19 is a severe contagious infection caused by a novel CoV, SARS-CoV-2. The common signs of COVID-19 disease include fever, breathlessness, and cough, just like in the case of SARS, and the primary host of the virus was supposed to be bats. This novel CoV is exceptionally infectious and spreads *via* close contact and droplets. The outbreak of this disease has created a global threat to human health and society as it is a life-threatening disease. This COVID-19 has affected various developed countries' public health, economics, and medical infrastructure. It is critical to stop the global epidemic from spreading and reduce the mortality rate as soon as possible. As per WHO and ECDC guidelines, avoiding contact with the sick person and market/public places is the best way to fight against COVID-19. It can now be controlled by implementing preventative measures in addition to the utilization of approved vaccinations and antiviral medications to reduce COVID-19's global impact. Finally, we must work hard to advance our research so that specific treatments can be proposed to reduce disease mortality and morbidity and save people's lives. Until today, vaccination programs have been the only way to diminish COVID-19's effects and develop herd immunity.

## LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| CoVs | = | Coronaviruses |
| CoV2 | = | Coronaviruse 2 |
| WHO | = | World health organization |
| SARS | = | Severe acute respiratory syndrome |
| MERS | = | Middle East respiratory syndrome |
| ssRNA | = | single-stranded ribonucleic acid |
| COVID-19 | = | Coronavirus disease 2019 |
| IL | = | Interleukin |
| TNF-α | = | Tumor necrosis factor-alpha |
| ACE-2 | = | Angiotensin-converting enzyme 2 |
| TMPRSS2 | = | Transmembrane serine protease 2 |

## CONSENT FOR PUBLICATION

Not applicable.

## FUNDING

PRS is thankful to the Biomedical Device and Technology Development (BDTD), Department of Science and Technology (DST), New Delhi, India; Department of Biotechnology (DBT), New Delhi, India, for the Indo-Russia project (DBT/IC-2/Indo-Russia/2017-19/02); and Indian Council of Medical Research (ICMR), New Delhi, India, for financial support. Amit K. Yadav recognized the Ministry of Education, Govt. of India for the Prime Minister Research Fellowship for the financial assistants.

## CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

## ACKNOWLEDGEMENTS

Declared none.

## REFERENCES

[1] Yadav AK, Verma D, Kumar A, Kumar P, Solanki PR. The perspectives of biomarker-based electrochemical immunosensors, artificial intelligence and the internet of medical things toward COVID-19 diagnosis and management. Mater Today Chem 2021; 20: 100443. http://dx.doi.org/10.1016/j.mtchem.2021.100443 PMID: 33615086

[2] Bogoch II, Watts A, Thomas-Bachli A, Huber C, Kraemer MU, Khan K. Pneumonia of unknown etiology in wuhan, china: Potential for international spread *via* commercial air travel. J Travel Med 2020; 27(2): taaa008. http://dx.doi.org/10.1093/jtm/taaa008

[3] Lu H, Stratton CW, Tang YW. Outbreak of pneumonia of unknown etiology in Wuhan China: The mystery and the miracle. J Med Virol 2020; 92(4): 401-2. http://dx.doi.org/10.1002/jmv.25678

[4] Singhal T. A review of coronavirus disease-2019 (COVID-19). Indian J Pediatr 2020; 87(4): 281-6. http://dx.doi.org/10.1007/s12098-020-03263-6 PMID: 32166607

[5] Organization WH. Novel Coronavirus ( 2019-nCoV): Situation Report. 2020; 3. Available from: https://apps.who.int/iris/handle/10665/330762

[6] Chan JFW, To KKW, Tse H, Jin DY, Yuen KY. Interspecies transmission and emergence of novel viruses: Lessons from bats and birds. Trends Microbiol 2013; 21(10): 544-55. http://dx.doi.org/10.1016/j.tim.2013.05.005 PMID: 23770275

[7] Zhu N, Zhang D, Wang W, *et al*. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020; 382(8): 727-33. http://dx.doi.org/10.1056/NEJMoa2001017 PMID: 31978945

[8] Corman VM, Ithete NL, Richards LR, *et al*. Rooting the phylogenetic tree of middle East respiratory syndrome coronavirus by characterization of a conspecific virus from an African bat. J Virol 2014; 88(19): 11297-303. http://dx.doi.org/10.1128/JVI.01498-14 PMID: 25031349

[9] Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 2019; 17(3): 181-92. http://dx.doi.org/10.1038/s41579-018-0118-9 PMID: 30531947

[10] Rahman A, Sarkar A. Risk factors for fatal middle east respiratory syndrome coronavirus infections in Saudi Arabia: Analysis of the WHO line list, 2013-2018. Am J Public Health 2019; 109(9): 1288-93. http://dx.doi.org/10.2105/AJPH.2019.305186 PMID: 31318592

[11] Memish ZA, Zumla AI, Assiri A. Middle East respiratory syndrome coronavirus infections in health care workers. N Engl J Med 2013; 369(9): 884-6. http://dx.doi.org/10.1056/NEJMc1308698 PMID: 23923992

[12] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol 2020; 5(4): 536-44. http://dx.doi.org/10.1038/s41564-020-0695-z PMID: 32123347

[13] Kupferschmidt K, Cohen J. Will novel virus go pandemic or be contained? Science 2020; 367(6478): 610-1. http://dx.doi.org/10.1126/science.367.6478.610

[14] Mishra A, Nair N, Yadav AK, Solanki P, Majeed J, Tripathi V. Coronavirus Disease 2019 (COVID-19): Origin, impact, and drug development. In: (Ed.), SARS-CoV-2 Origin and COVID-19 Pandemic Across the Globe. Intech Open 2021. http://dx.doi.org/10.5772/intechopen.98358

[15] Liu Y, Wang K, Massoud TF, Paulmurugan R. SARS-CoV-2 vaccine development: An overview and perspectives. ACS Pharmacol Transl Sci 2020; 3(5): 844-58. http://dx.doi.org/10.1021/acsptsci.0c00109 PMID: 33062951

[16] Poland GA, Ovsyannikova IG, Crooke SN, Kennedy RB, Eds. SARS-CoV-2 vaccine development: Current status. Mayo Clin Proc 2020; 95(10): 2172-88. http://dx.doi.org/10.1016/j.mayocp.2020.07.021

[17] Seow J, Graham C, Merrick B, Acors S, Pickering S, Steel KJ. Longitudinal evaluation and decline of antibody responses in SARS-CoV-2 infection. Nat Microbiol 2020; 5(12): 1598. http://dx.doi.org/10.1038/s41564-020-00813-8 PMID: 33106674

[18] Wang K, Long QX, Deng HJ, *et al*. Longitudinal dynamics of the neutralizing antibody response to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection. Clin Infect Dis 2021; 73(3): e531-9. http://dx.doi.org/10.1093/cid/ciaa1143 PMID: 32745196

[19] To K, Hung I, Ip JD, Chu A, Chan WM, Tam AR, *et al*. COVID-19 re-infection by a phylogenetically distinct SARS-coronavirus-2 strain confirmed by whole genome sequencing. Clin Infect Dis 2021; 73(9): e2946-51. http://dx.doi.org/10.1093/cid/ciaa1275 PMID: 32840608

[20] Venkat Kumar G, Jeyanthi V, Ramakrishnan S. A short review on antibody therapy for COVID-19. New Microbes New Infect 2020; 35: 100682. http://dx.doi.org/10.1016/j.nmni.2020.100682 PMID: 32313660

[21] Woo PC, Huang Y, Lau SK, Yuen K-Y. Coronavirus genomics and bioinformatics analysis. Viruses 2010; 2(8): 1804-20. http://dx.doi.org/10.3390/v2081803 PMID: 21994708

[22] Jalandra R, Yadav AK, Verma D, *et al*. Strategies and perspectives to develop SARS-CoV-2 detection methods and diagnostics. Biomed Pharmacother 2020; 129: 110446. http://dx.doi.org/10.1016/j.biopha.2020.110446 PMID: 32768943

[23] Sajadi MM, Habibzadeh P, Vintzileos A, Shokouhi S, Miralles-Wilhelm F, Amoroso A. Temperature, humidity and latitude analysis to predict potential spread and seasonality for COVID-19. SSRN 2020; 3550308. http://dx.doi.org/10.2139/ssrn.3550308 PMID: 32714105

[24] Casanova LM, Jeon S, Rutala WA, Weber DJ, Sobsey MD. Effects of air temperature and relative humidity on coronavirus survival on surfaces. Appl Environ Microbiol 2010; 76(9): 2712-7. http://dx.doi.org/10.1128/AEM.02291-09 PMID: 20228108

[25] Gaunt ER, Hardie A, Claas EC, Simmonds P, Templeton KE. Epidemiology and clinical presentations of the four human coronaviruses 229E, HKU1, NL63, and OC43 detected over 3 years using a novel multiplex real-time PCR method. J Clin Microbiol 2010; 48(8): 2940-7. http://dx.doi.org/10.1128/JCM.00636-10 PMID: 20554810

[26] Sardar R, Satish D, Birla S, Gupta D. Comparative analyses of SAR-CoV-2 genomes from different geographical locations and other coronavirus family genomes reveals unique features potentially consequential to host-virus interaction and pathogenesis. bioRxiv 2020. http://dx.doi.org/10.1101/2020.03.21.001586

[27] Li Q, Guan X, Wu P, *et al*. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. N Engl J Med 2020; 382(13): 1199-207. http://dx.doi.org/10.1056/NEJMoa2001316 PMID: 31995857

[28] Wang W, Tang J, Wei F. Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China. J Med Virol 2020; 92(4): 441-7.
http://dx.doi.org/10.1002/jmv.25689 PMID: 31994742

[29] Zeng LK, Tao XW, Yuan WH, Wang J, Liu X, Liu ZS. First case of neonate infected with novel coronavirus pneumonia in China. Zhonghua Er Ke Za Zhi 2020; 58(0): E009.
PMID: 32065520

[30] To KKW, Tsang OTY, Yip CCY, *et al*. Consistent detection of 2019 novel coronavirus in saliva. Clin Infect Dis 2020; 71(15): 841-3.
http://dx.doi.org/10.1093/cid/ciaa149 PMID: 32047895

[31] Bhadra S, Jiang YS, Kumar MR, Johnson RF, Hensley LE, Ellington AD. Real-time sequence-validated loop-mediated isothermal amplification assays for detection of Middle East respiratory syndrome coronavirus (MERS-CoV). PLoS One 2015; 10(4): e0123126.
http://dx.doi.org/10.1371/journal.pone.0123126 PMID: 25856093

[32] Chan JFW, Choi GKY, Tsang AKL, *et al*. Development and evaluation of novel real-time reverse transcription-PCR assays with locked nucleic acid probes targeting leader sequences of human-pathogenic coronaviruses. J Clin Microbiol 2015; 53(8): 2722-6.
http://dx.doi.org/10.1128/JCM.01224-15 PMID: 26019210

[33] Huang P, Liu T, Huang L, *et al*. Use of chest CT in combination with negative RT-PCR assay for the 2019 novel coronavirus but high clinical suspicion. Radiology 2020; 295(1): 22-3.
http://dx.doi.org/10.1148/radiol.2020200330 PMID: 32049600

[34] Chu DKW, Pan Y, Cheng SMS, *et al*. Molecular diagnosis of a novel coronavirus (2019-nCoV) causing an outbreak of pneumonia. Clin Chem 2020; 66(4): 549-55.
http://dx.doi.org/10.1093/clinchem/hvaa029 PMID: 32031583

[35] Chan JFW, Yip CCY, To KKW, *et al*. Improved molecular diagnosis of COVID-19 by the novel, highly sensitive and specific COVID-19-RdRp/Hel real-time reverse transcription-PCR assay validated *in vitro* and with clinical specimens. J Clin Microbiol 2020; 58(5): e00310-20.
http://dx.doi.org/10.1128/JCM.00310-20 PMID: 32132196

[36] Cobb B, Simon CO, Stramer SL, *et al*. The Cobas® 6800/8800 System: A new era of automation in molecular diagnostics. Expert Rev Mol Diagn 2017; 17(2): 167-80.
http://dx.doi.org/10.1080/14737159.2017.1275962 PMID: 28043179

[37] Marlowe EM, Hardy D, Krevolin M, *et al*. High-throughput testing of urogenital and extragenital specimens for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* with Cobas® CT/NG. Eur J Microbiol Immunol (Bp) 2017; 7(3): 176-86.
http://dx.doi.org/10.1556/1886.2017.00018 PMID: 29034107

[38] Greub G, Sahli R, Brouillet R, Jaton K. Ten years of R&D and full automation in molecular diagnosis. Future Microbiol 2016; 11(3): 403-25.
http://dx.doi.org/10.2217/fmb.15.152 PMID: 27028061

[39] Eigner U, Reucher S, Hefner N, *et al*. Clinical evaluation of multiplex RT-PCR assays for the detection of influenza A/B and respiratory syncytial virus using a high throughput system. J Virol Methods 2019; 269: 49-54.
http://dx.doi.org/10.1016/j.jviromet.2019.03.015 PMID: 30946852

[40] Pfefferle S, Reucher S, Nörz D, Lütgehetmann M. Evaluation of a quantitative RT-PCR assay for the detection of the emerging coronavirus SARS-CoV-2 using a high throughput system. Euro Surveill 2020; 25(9): 2000152.
http://dx.doi.org/10.2807/1560-7917.ES.2020.25.9.2000152 PMID: 32156329

[41] Xie X, Zhong Z, Zhao W, Zheng C, Wang F, Liu J. Chest CT for typical 2019-nCoV pneumonia: Relationship to negative RT-PCR testing. Radiology 2020; 296(2): E41-5.
http://dx.doi.org/10.1148/radiol.2020200343

[42] Green K, Graziadio S, Turner P, Fanshawe T, Allen J. Molecular and antibody point-of-care tests to support the screening, diagnosis and monitoring of COVID-19. The Centre for Evidence-Based Medicine 2020. Available from: https://www.cebm.net/covid-19/molecular-and-antibody-point-of-care-tests-to-support-the-screening-diagnosis-and-monitoring-of-covid-19/

[43] Chen N, Zhou M, Dong X, *et al*. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. Lancet 2020; 395(10223): 507-13.
http://dx.doi.org/10.1016/S0140-6736(20)30211-7 PMID: 32007143

[44] Expert consensus on chloroquine phosphate for the treatment of novel coronavirus pneumonia. Zhonghua Jie He He Hu Xi Za Zhi 2020; 43(3): 185-8.
PMID: 32164085

[45] Jin YH, Cai L, Cheng ZS, *et al*. A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version). Mil Med Res 2020; 7(1): 4.
http://dx.doi.org/10.1186/s40779-020-0233-6 PMID: 32029004

[46] Zhang L, Liu Y. Potential interventions for novel coronavirus in China: A systematic review. J Med Virol 2020; 92(5): 479-90.
http://dx.doi.org/10.1002/jmv.25707 PMID: 32052466

[47] Liu C, Zhou Q, Li Y, *et al*. Research and development on therapeutic agents and vaccines for COVID-19 and related human coronavirus diseases. ACS Cent Sci 2020; 6: pp. (3)315-31.
http://dx.doi.org/10.1021/acscentsci.0c00272

[48] Sheahan TP, Sims AC, Leist SR, *et al*. Comparative therapeutic efficacy of remdesivir and combination lopinavir, ritonavir, and interferon beta against MERS-CoV. Nat Commun 2020; 11(1): 222.
http://dx.doi.org/10.1038/s41467-019-13940-6 PMID: 31924756

[49] Maxmen A. More than 80 clinical trials launch to test coronavirus treatments. Nature 2020; 578(7795): 347-8.
http://dx.doi.org/10.1038/d41586-020-00444-3 PMID: 32071447

[50] Kadam RU, Wilson IA. Structural basis of influenza virus fusion inhibition by the antiviral drug Arbidol. Proc Natl Acad Sci USA 2017; 114(2): 206-14.
http://dx.doi.org/10.1073/pnas.1617020114 PMID: 28003465

[51] Qiu R, Wei X, Zhao M, Zhong C, Zhao C, Hu J, *et al*. Outcome reporting from protocols of clinical trials of coronavirus disease 2019 (COVID-19): A review. medRxiv 2020.
http://dx.doi.org/10.1101/2020.03.04.20031401

[52] Hoffmann M, Kleine-Weber H, Schroeder S, *et al*. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 2020; 181(2): 271-280.e8.
http://dx.doi.org/10.1016/j.cell.2020.02.052 PMID: 32142651

[53] Morse JS, Lalonde T, Xu S, Liu WR. Learning from the past: Possible urgent prevention and treatment options for severe acute respiratory infections caused by 2019-nCoV. ChemBioChem 2020; 21(5): 730-8.
http://dx.doi.org/10.1002/cbic.202000047 PMID: 32022370

[54] Guo D. Old weapon for new enemy: Drug repurposing for treatment of newly emerging viral diseases. Virol Sin 2020; 35(3): 253-5.
http://dx.doi.org/10.1007/s12250-020-00204-7 PMID: 32048130

[55] Arabi YM, Mandourah Y, Al-Hameed F, *et al*. Corticosteroid therapy for critically ill patients with Middle East respiratory syndrome. Am J Respir Crit Care Med 2018; 197(6): 757-67.
http://dx.doi.org/10.1164/rccm.201706-1172OC PMID: 29161116

[56] Wang Q, Mu F, Xie J, Cheng J, Fu Y, Jiang D. A single ssRNA segment encoding RdRp is sufficient for replication, infection, and transmission of ourmia-like virus in fungi. Front Microbiol 2020; 11: 379.
http://dx.doi.org/10.3389/fmicb.2020.00379 PMID: 32256466

[57] Samuel CE. Antiviral actions of interferons. Clin Microbiol Rev 2001; 14(4): 778-809.
http://dx.doi.org/10.1128/CMR.14.4.778-809.2001 PMID: 11585785

[58] Falzarano D, de Wit E, Rasmussen AL, *et al*. Treatment with interferon-α2b and ribavirin improves outcome in MERS-CoV-infected rhesus macaques. Nat Med 2013; 19(10): 1313-7.
http://dx.doi.org/10.1038/nm.3362 PMID: 24013700

[59] Yamamoto N, Yang R, Yoshinaka Y, *et al*. HIV protease inhibitor nelfinavir inhibits replication of SARS-associated coronavirus. Biochem Biophys Res Commun 2004; 318(3): 719-25.
http://dx.doi.org/10.1016/j.bbrc.2004.04.083 PMID: 15144898

[60] Elfiky AA. Anti-HCV, nucleotide inhibitors, repurposing against COVID-19. Life Sci 2020; 248: 117477.
http://dx.doi.org/10.1016/j.lfs.2020.117477 PMID: 32119961

[61] So LK, Lau AC, Yam LY, *et al*. Development of a standard treatment protocol for severe acute respiratory syndrome. Lancet 2003; 361(9369): 1615-7.

http://dx.doi.org/10.1016/S0140-6736(03)13265-5 PMID: 12747883

[62] Morgenstern B, Michaelis M, Baer PC, Doerr HW, Cinatl J Jr. Ribavirin and interferon-β synergistically inhibit SARS-associated coronavirus replication in animal and human cell lines. Biochem Biophys Res Commun 2005; 326(4): 905-8. http://dx.doi.org/10.1016/j.bbrc.2004.11.128 PMID: 15607755

[63] Zumla A, Chan JF, Azhar EI, Hui DS, Yuen K-Y. Coronaviruses - drug discovery and therapeutic options. Nat Rev Drug Discov 2016; 15(5): 327-47. http://dx.doi.org/10.1038/nrd.2015.37 PMID: 26868298

[64] Tchesnokov EP, Feng JY, Porter DP, Götte M. Mechanism of inhibition of Ebola virus RNA-dependent RNA polymerase by remdesivir. Viruses 2019; 11(4): 326. http://dx.doi.org/10.3390/v11040326 PMID: 30987343

[65] Mulangu S, Dodd LE, Davey RT Jr, *et al*. A randomized, controlled trial of Ebola virus disease therapeutics. N Engl J Med 2019; 381(24): 2293-303. http://dx.doi.org/10.1056/NEJMoa1910993 PMID: 31774950

[66] Wang Z, Chen X, Lu Y, Chen F, Zhang W. Clinical characteristics and therapeutic procedure for four cases with 2019 novel coronavirus pneumonia receiving combined Chinese and Western medicine treatment. Biosci Trends 2020; 14(1): 64-8. http://dx.doi.org/10.5582/bst.2020.01030 PMID: 32037389

[67] Holshue ML, DeBolt C, Lindquist S, *et al*. First case of 2019 novel coronavirus in the United States. N Engl J Med 2020; 382(10): 929-36. http://dx.doi.org/10.1056/NEJMoa2001191 PMID: 32004427

[68] Coronavirus F. Update: FDA Issues Emergency Use Authorization for Potential COVID-19 Treatment. Press Announc 2020. Available from: https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-issues-emergency-use-authorization-potential-covid-19-treatment

[69] Musa A, Warbasse E, Baron DA, *et al*. Addendum to systematic review of remdesivir for the treatment of COVID-19. West J Emerg Med 2020; 21(4): 742-3. http://dx.doi.org/10.5811/westjem.2020.5.48121 PMID: 32726231

[70] Williamson BN, Feldmann F, Schwarz B, *et al*. Clinical benefit of remdesivir in *rhesus macaques* infected with SARS-CoV-2. Nature 2020; 585(7824): 273-6. http://dx.doi.org/10.1038/s41586-020-2423-5 PMID: 32516797

[71] Richardson P, Griffin I, Tucker C, *et al*. Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. Lancet 2020; 395(10223): e30-1. http://dx.doi.org/10.1016/S0140-6736(20)30304-4 PMID: 32032529

[72] Mifsud EJ, Hayden FG, Hurt AC. Antivirals targeting the polymerase complex of influenza viruses. Antiviral Res 2019; 169: 104545. http://dx.doi.org/10.1016/j.antiviral.2019.104545 PMID: 31247246

[73] Gleumark Pharmaceutical Ltd. Clinical trials on antiviral drug Favipiravir for COVID-19 patients in India, 2019.

[74] Sultana J, Cutroneo PM, Crisafulli S, Puglisi G, Caramori G, Trifirò G. Azithromycin in COVID-19 patients: Pharmacological mechanism, clinical evidence and prescribing guidelines. Drug Saf 2020; 43(8): 691-8. http://dx.doi.org/10.1007/s40264-020-00976-7 PMID: 32696429

[75] Gautret P, Lagier JC, Parola P, *et al*. Hydroxychloroquine and azithromycin as a treatment of COVID-19: Results of an open-label non-randomized clinical trial. Int J Antimicrob Agents 2020; 56(1): 105949. http://dx.doi.org/10.1016/j.ijantimicag.2020.105949 PMID: 32205204

[76] González Canga A, Sahagún Prieto AM, Diez Liébana MJ, Fernández Martínez N, Sierra Vega M, García Vieitez JJ. The pharmacokinetics and interactions of ivermectin in humans--a mini-review. AAPS J 2008; 10(1): 42-6. http://dx.doi.org/10.1208/s12248-007-9000-9 PMID: 18446504

[77] Tay MY, Fraser JE, Chan WK, *et al*. Nuclear localization of dengue virus (DENV) 1-4 non-structural protein 5; protection against all 4 DENV serotypes by the inhibitor ivermectin. Antiviral Res 2013; 99(3): 301-6. http://dx.doi.org/10.1016/j.antiviral.2013.06.002 PMID: 23769930

[78] Caly L, Druce JD, Catton MG, Jans DA, Wagstaff KM. The FDA-approved drug ivermectin inhibits the replication of SARS-CoV-2 *in vitro*. Antiviral Res 2020; 178: 104787. http://dx.doi.org/10.1016/j.antiviral.2020.104787 PMID: 32251768

[79] Zeldin RK, Petruschke RA. Pharmacological and therapeutic properties of ritonavir-boosted protease inhibitor therapy in HIV-infected patients. J Antimicrob Chemother 2004; 53(1): 4-9. http://dx.doi.org/10.1093/jac/dkh029 PMID: 14657084

[80] Wang M, Cao R, Zhang L, *et al*. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) *in vitro*. Cell Res 2020; 30(3): 269-71. http://dx.doi.org/10.1038/s41422-020-0282-0 PMID: 32020029

[81] Warren TK, Wells J, Panchal RG, *et al*. Protection against filovirus diseases by a novel broad-spectrum nucleoside analogue BCX4430. Nature 2014; 508(7496): 402-5. http://dx.doi.org/10.1038/nature13027 PMID: 24590073

[82] Sahu KK, Mishra AK, Lal A. Comprehensive update on current outbreak of novel coronavirus infection (2019-nCoV). Ann Transl Med 2020; 8(6): 393. http://dx.doi.org/10.21037/atm.2020.02.92 PMID: 32355837

[83] Tang B, Li S, Xiong Y, *et al*. Coronavirus disease 2019 (COVID-19) pneumonia in a hemodialysis patient. Kidney Med 2020; 2(3): 354-8. PMID: 32292904

[84] Pang J, Wang MX, Ang IYH, *et al*. Potential rapid diagnostics, vaccine and therapeutics for 2019 novel coronavirus (2019-nCoV): A systematic review. J Clin Med 2020; 9(3): 623. http://dx.doi.org/10.3390/jcm9030623 PMID: 32110875

[85] Chau TN, Lee KC, Yao H, *et al*. SARS-associated viral hepatitis caused by a novel coronavirus: Report of three cases. Hepatology 2004; 39(2): 302-10. http://dx.doi.org/10.1002/hep.20111 PMID: 14767982

[86] Coleman CM, Sisk JM, Mingo RM, Nelson EA, White JM, Frieman MB. Abelson kinase inhibitors are potent inhibitors of severe acute respiratory syndrome coronavirus and Middle East respiratory syndrome coronavirus fusion. J Virol 2016; 90(19): 8924-33. http://dx.doi.org/10.1128/JVI.01429-16 PMID: 27466418

[87] Park HY. Primary registry of the WHO International clinical trial registry platform: Clinical Research Information Service (CRIS). J Korean Med Assoc 2011; 54(1): 92-7. http://dx.doi.org/10.5124/jkma.2011.54.1.92

[88] Cong Y, Hart BJ, Gross R, *et al*. MERS-CoV pathogenesis and antiviral efficacy of licensed drugs in human monocyte-derived antigen-presenting cells. PLoS One 2018; 13(3): e0194868. http://dx.doi.org/10.1371/journal.pone.0194868 PMID: 29566060

[89] Kindrachuk J. Selective inhibition of host cell signaling for rotavirus antivirals: PI3K/Akt/mTOR-mediated rotavirus pathogenesis. Virulence 2018; 9(1): 5-8. http://dx.doi.org/10.1080/21505594.2017.1356539 PMID: 28723236

[90] Al Ghamdi M, Alghamdi KM, Ghandoora Y, *et al*. Treatment outcomes for patients with Middle Eastern Respiratory Syndrome coronavirus (MERS CoV) infection at a coronavirus referral center in the Kingdom of Saudi Arabia. BMC Infect Dis 2016; 16(1): 174. http://dx.doi.org/10.1186/s12879-016-1492-4 PMID: 27097824

[91] Arabi YM, Shalhoub S, Mandourah Y, *et al*. Ribavirin and interferon therapy for critically Ill patients with middle East respiratory syndrome: A multicenter observational study. Clin Infect Dis 2020; 70(9): 1837-44. http://dx.doi.org/10.1093/cid/ciz544 PMID: 31925415

[92] Arabi YM, Alothman A, Balkhy HH, *et al*. Treatment of Middle East respiratory syndrome with a combination of lopinavir-ritonavir and interferon-β1b (MIRACLE trial): Study protocol for a randomized controlled trial. Trials 2018; 19(1): 81. http://dx.doi.org/10.1186/s13063-017-2427-0 PMID: 29382391

[93] Chu H, Zhou J, Wong BHY, *et al*. Middle East respiratory syndrome coronavirus efficiently infects human primary T lymphocytes and activates the extrinsic and intrinsic apoptosis pathways. J Infect Dis 2016; 213(6): 904-14. http://dx.doi.org/10.1093/infdis/jiv380 PMID: 26203058

[94] Lyssenko V, Jonsson A, Almgren P, *et al*. Clinical risk factors, DNA variants, and the development of type 2 diabetes. N Engl J Med 2008; 359(21): 2220-32. http://dx.doi.org/10.1056/NEJMoa0801869 PMID: 19020324

[95] Cai Q, Yang M, Liu D, *et al*. Experimental treatment with favipiravir for COVID-19: An open-label control study. Engineering (Beijing) 2020; 6(10): 1192-8. http://dx.doi.org/10.1016/j.eng.2020.03.007 PMID: 32346491

[96] Boltz D, Peng X, Muzzio M, Dash P, Thomas PG, Margitich V. Activity of enisamium, an isonicotinic acid derivative, against influenza viruses in differentiated normal human bronchial epithelial cells. Antivir Chem Chemother 2018; 26: 2040206618811416. http://dx.doi.org/10.1177/2040206618811416 PMID: 30466301

[97] Cocking D, Cinatl J, Boltz DA, *et al*. Antiviral effect of a derivative of isonicotinic acid enisamium iodide (FAV00A) against influenza virus. Acta Virol 2018; 62(2): 191-5. http://dx.doi.org/10.4149/av_2018_211 PMID: 29895160

[98] Rosa SGV, Santos WC. Clinical trials on drug repositioning for COVID-19 treatment. Rev Panam Salud Publica 2020; 44: e40. http://dx.doi.org/10.26633/RPSP.2020.40 PMID: 32256547

[99] Qin C, Zhou L, Hu Z, *et al*. Dysregulation of immune response in patients with COVID-19 in Wuhan, China. Clin Infect Dis 2020 Jul 28; 71(15): 762-8. http://dx.doi.org/10.1093/cid/ciaa248 PMID: 32161940

[100] Colson P, Rolain JM, Lagier JC, Brouqui P, Raoult D. Chloroquine and hydroxychloroquine as available weapons to fight COVID-19. Int J Antimicrob Agents 2020; 55(4): 105932. http://dx.doi.org/10.1016/j.ijantimicag.2020.105932 PMID: 32145363

[101] Lawes-Wickwar S, Ghio D, Tang MY, *et al*. A rapid systematic review of public responses to health messages encouraging vaccination against infectious diseases in a pandemic or epidemic. Vaccines (Basel) 2021; 9(2): 72. http://dx.doi.org/10.3390/vaccines9020072 PMID: 33498395

[102] Doroftei B, Ciobica A, Ilie OD, Maftei R, Ilea C. Mini-review discussing the reliability and efficiency of COVID-19 vaccines. Diagnostics (Basel) 2021; 11(4): 579. http://dx.doi.org/10.3390/diagnostics11040579 PMID: 33804914

[103] Carvalho T, Krammer F, Iwasaki A. The first 12 months of COVID-19: A timeline of immunological insights. Nat Rev Immunol 2021; 21(4): 245-56. http://dx.doi.org/10.1038/s41577-021-00522-1 PMID: 33723416

[104] Kim JH, Marks F, Clemens JD. Looking beyond COVID-19 vaccine phase 3 trials. Nat Med 2021; 27(2): 205-11. http://dx.doi.org/10.1038/s41591-021-01230-y PMID: 33469205

[105] Koirala A, Joo YJ, Khatami A, Chiu C, Britton PN. Vaccines for COVID-19: The current state of play. Paediatr Respir Rev 2020; 35: 43-9. PMID: 32653463

[106] Acharya KP, Ghimire TR, Subramanya SH. Access to and equitable distribution of COVID-19 vaccine in low-income countries. NPJ Vaccines (Basel) 2021; 6(1): 54. PMID: 33854072

[107] Venkadapathi J, Govindarajan VK, Sekaran S, Venkatapathy S. A minireview of the promising drugs and vaccines in pipeline for the treatment of COVID-19 and current update on clinical trials. Front Mol Biosci 2021; 8: 637378. http://dx.doi.org/10.3389/fmolb.2021.637378 PMID: 34179072

[108] Baden LR, El Sahly HM, Essink B, *et al*. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. N Engl J Med 2021; 384(5): 403-16. http://dx.doi.org/10.1056/NEJMoa2035389 PMID: 33378609

[109] Jones I, Roy P. Sputnik V. COVID-19 vaccine candidate appears safe and effective. Lancet 2021; 397(10275): 642-3. http://dx.doi.org/10.1016/S0140-6736(21)00191-4 PMID: 33545098



View publication stats

Electrical,Electronic and Telecommunications Engineering

# Recent Trends in the Optimization of Renewable Distributed Generation: A Review

## Tendencias recientes en la optimización de la generación distribuida renovable: una revisión

Vivek Saxena[1,2], Narendra Kumar[3], Uma Nangia[4]

**ABSTRACT**

Distributed energy resources, or distributed generation (DG), are the mainstay of modern power systems aiming towards green energy generation via the effective integration of renewable energy sources. DG involvement in traditional power systems comprises the enhancement of power quality parameters, the utilization of renewable sources, cost optimization, and stable and reliable energy generation. The advantages of such a revolutionary approach can be achieved with the optimal sizing and allocation of DG by means of adequate optimization techniques, constraints, and optimized parameters. In this study, a comprehensive review of DG optimization is presented in light of recent developments, and a comparison is carried out based on the adopted optimization techniques, test system, enhanced parameters, and outcome evaluations.

**Keywords:** energy generation, distributed energy resources, renewable sources, optimization approach

**RESUMEN**

Los recursos energéticos distribuidos, o generación distribuida (GD), son el pilar de los sistemas energéticos modernos, que tienen como objetivo generar energía verde a través de la integración efectiva de fuentes de energía renovables. La participación de la GD en el sistema eléctrico tradicional abarca la mejora de los parámetros de calidad de la energía, la utilización de fuentes renovables, la optimización de costos y la generación de energía estable y confiable. Las ventajas de un enfoque tan revolucionario se pueden alcanzar asignando el tamaño y la ubicación óptimos de la GD mediante técnicas de optimización adecuadas, restricciones y parámetros optimizados. En este estudio se presenta una revisión completa de la optimización de GD en vista del desarrollo reciente, y se hace una comparación basada en las técnicas de optimización adoptadas, el sistema de prueba, los parámetros mejorados y las evaluaciones de resultados.

**Palabras clave:** generación de energía, recursos energéticos distribuidos, fuentes renovables, enfoque de optimización

**Received: August 07th 2021**
**Accepted: March 31th 2022**

## Introduction[1 2]

Distributed generation (DG) is different from the centralized generation of electrical energy, which is attributed to an approach that involves generation through small units situated near the end users. DG is also known as dispersed generation, embedded generation, or decentralized generation. Renewable and non-renewable energy resources can be utilized due to their technical, financial, and environmental advantages via the optimal establishing of DG. The qualities of electrical parameters that can be consigned in electrical power systems are real and reactive power losses, voltage stability, the tap setting of a transformer, power factor correction, reliability, fault level, and total harmonic reduction with the assimilation of DG.

Environmental pollution has been minimized by maximizing the use of renewable energy sources. Integrated generation models have been designed as an alternative to centralized generation, which demonstrate the potential of solar photovoltaic systems, wind power generation, biogas, and other natural energy sources. The emission level of hazardous pollutants can also be an objective parameter in a multi-objective function for DG allocation. The challenges of renewable energy amalgamation are the intermittent nature of energy sources, which are highly dependent on meteorological conditions and geographical structures (Saxena *et al.*, 2021).

### *Merits of DG planning*

For optimal DG allocation, the objective function has been optimized by considering various constraints in order to determine

1 Research Scholar, Department of Electrical Engineering, Delhi Technological University, Delhi, India.
2 Assistant Professor, Department of Electrical and Electronics Engineering, A.B.E.S. Engineering College, Ghaziabad, India.
3 Professor, Department of Electrical Engineering, Delhi Technological University, Delhi, India.
4 Professor, Department of Electrical Engineering, Delhi Technological University, Delhi, India.

**How to cite:** Saxena, V., Kumar, N., and Nangia, U. (2022). Recent trends in the optimization of renewable distributed generation: A review. *Ingeniería e Investigación*, *42*(3), e97702.

the reliability, stability, and utility of the system. As per the desired level of operation, constraints such as load flow, voltage profile, line losses, transformer capacity, short circuit level, power quality, number, size, and capacity of DG may be a part of a single objective or multi-objective functions. As far as financial objectives are concerned, DG capital, operation, and maintenance costs have been considered for cost reduction and maximization of profit. In this process, the objective function includes revenue besides the expenditure, the penalty factor for constraints violations, the integrated cost parameters, and fixed and variable cost parameters.

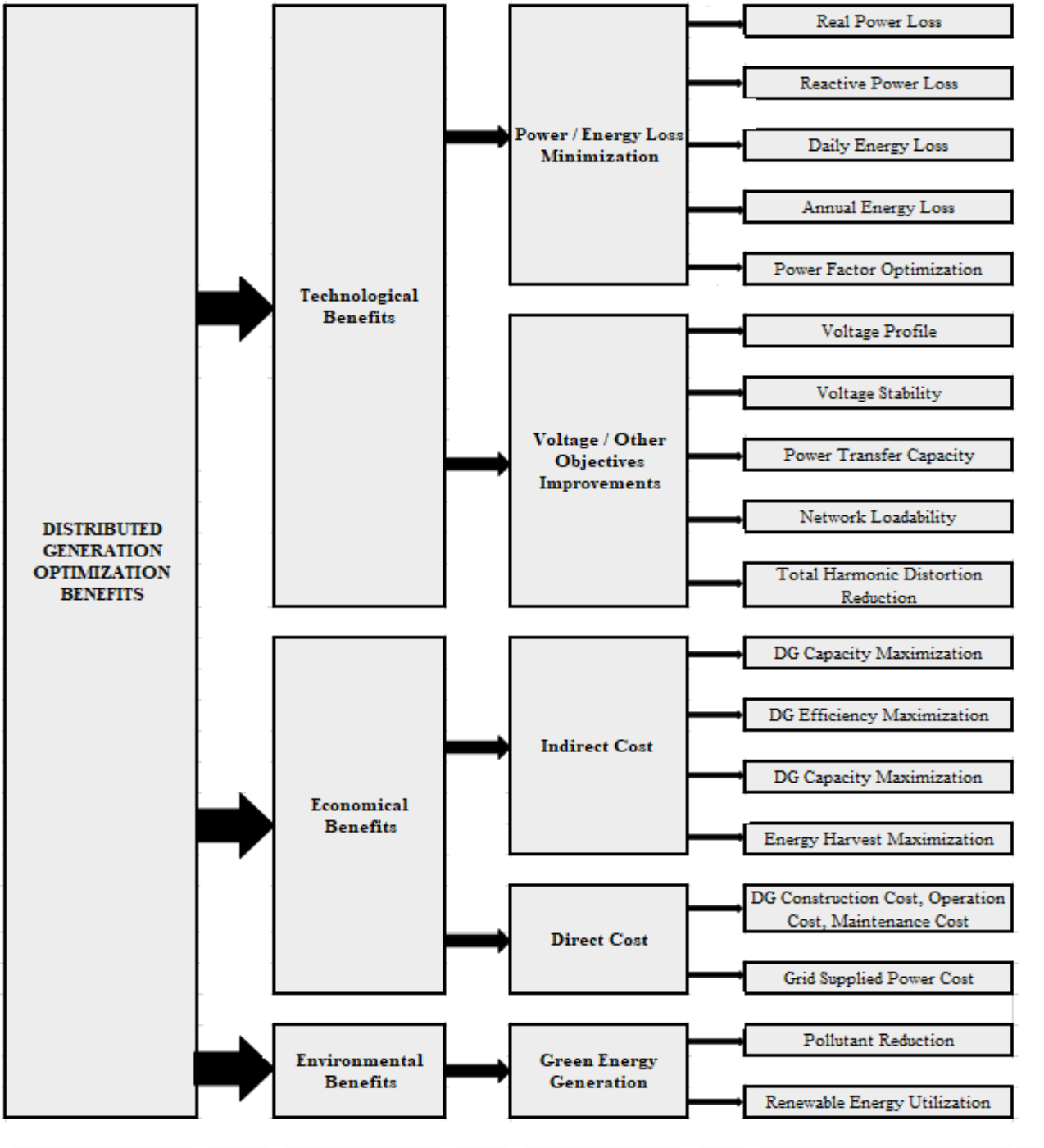



**Figure 1.** Benefits of DG optimization

**Source:** Authors

The benefits of DG optimization are summarized in Figure 1. The planning of renewable DG in distribution networks has several advantages, and broad classifications are given as per the following highlights.

### *Technological benefits*

DG integration in distribution networks reduces system losses and increases the efficacy, reliability, and optimal dispatch of the system. Moreover, the voltage profile and power factor are also

simultaneously optimized with power loss minimization. The objective function may be single-objective or multi-objective and subjected to various equality and inequality constraints. Technological benefits are achieved after optimized DG allocation.

*Economic benefits*

The demand for electrical energy is increasing day by day, and it is not economic to expand centralized generation due to its various limitations. The incorporation of DG may lead down a path that is cost-effective and provides technological benefits, reducing energy losses and satisfying the energy demand through natural resources project DG as a futuristic approach. Capacity maximization and energy harvesting within the distribution network is also involved in financially optimized systems.

*Environmental benefits*

The high impact of renewable energy resources in the power network enlightens the pathway towards pollution-free, green energy generation. The maximization of extracted renewable energy is possible through the integration of DG in the distribution system.

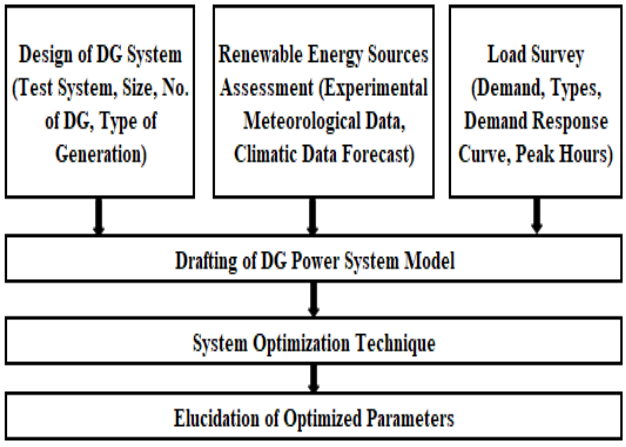



**Figure 2.** Framework for renewable DG planning

**Source:** Authors

The key segments of renewable energy-based DG planning are shown in Figure 2.

## DG system design

Designing the DG system involves the selection of the energy generation category, the storage technique, and the bus system. Energy generation is broadly categorized based on renewable and non-renewable resources. Solar energy, wind energy, biomass energy, and fuel cells are the main renewable energy sources, whereas combustion is utilized in fossil fuel-based energy extraction (Allan *et al.,* 2015).

Despite the intermittent nature of renewable energy, there are numerous stable integrations of distributed energy generation and utilization, in addition to the incorporation of storage devices. Moreover, crop-based energy is also employed in DG.

The different types of bus planning can be categorized as follows:

*Bus system with DC*

The generated energy is delivered easily to DC load via a DC bus. However, inverter devices may be used to deliver energy to AC loads.

*Bus system with AC*

This architecture is used to directly supply AC loads through AC buses, while the DC demand is completed by using rectifiers.

*Bus system with AC and DC*

It is a more complex but highly efficient system to deliver AC and DC supply concurrently. Energy storage is also possible with the help of supply converters.

A categorization of renewable energy sources is presented in Figure 3 (DTI, 2007).

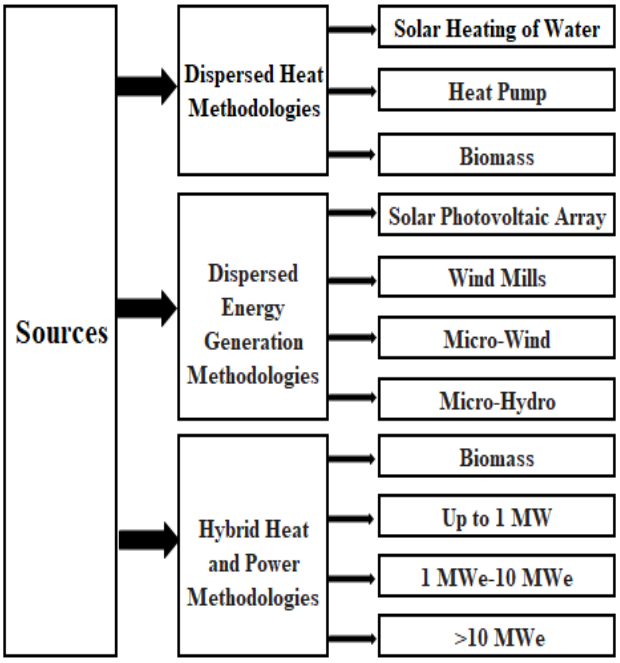



**Figure 3.** Categorization of renewable energy sources

**Source:** Authors

## Renewable energy assessment

The assessment of renewable energy is a key component of DG implementation because of the intermittent nature of such resources in comparison with fossil fuel-based generation. To enhance the impact of green energy generation, it is necessary to prepare a framework of renewable energy-based DGs to neutralize the impact of intermittency. The weather conditions, topographical properties, experimental results, data records, soil grade, and seasonal variation are the most important parameters for the analysis involved in the elaboration of the renewable DG framework.

In this vein, biomass energy is dependent on atmospheric conditions, but storage of feedstock can provide continuous energy generation. In turn, wind and solar are unlimited sources of energy but have a high level of intermittency. It is a very difficult task to develop a model that can ensure uninterrupted energy generation.

The intermittency of wind and solar power can be reduced by the following methods (Theo *et al.,* 2017):

*Experiential meteorological data*

Experiential data of solar and wind function can be collected both directly and indirectly by means of visiting the site and consulting published research, meteorological laboratory records, relevant

offices, and authorities. The measured metrological data related to solar energy are the irradiance level (hourly, daily, weekly, monthly, seasonal, and global), the intensity, and the power output. Moreover, the wind data collected are the hourly average, the monthly mean, the energy generated per day, and the frequency.

### Climatic data forecast

It is necessary to forecast reliable metrological information, even in light of the following constraints, which may hamper the assessment:

- The proficiency and resources are limited in a particular area;
- low financial assistance limits the complete, highly accurate seasonal data observation; and
- the failure of an observational system may lead to the absence of regular data recording for a long period.

## Load survey

Load demand analysis is mandatory for generation planning and utilization. The nature of the load may be continuous or discrete, while its measurement may be conducted in different time frames or seasons. It consists of analyzing the demand at low, medium, and high levels, and it varies from residential to industrial sectors. Studying the load curve from peak to valley and valley to peak is also required to maintain the consistency between demand and supply.

## Optimized DG allocation

A predicted capacity-based solar power plant in an unbalanced power distribution model was recommended. The particle swarm optimization (PSO) technique was used to optimize the size and location to get the plant's capacity and determine its power capacity per day. The process started by modeling the varying load, followed by assuming an unbalanced distribution system, and concluded with the determination of the DG's probable capacity. The optimized parameters were compared to the absence of distributed generation (ADG) (Jana *et al.,* 2020).

A metaheuristic, hybrid-approach PSO coupled with a grey wolf optimization (PSOGWO) approach was implemented to enhance the voltage profile and optimize the real power losses. The photovoltaic and fuel cells and variable capacitor types were selected for the optimization process, and the results were significant for voltage profile improvement and real power losses reduction (Kumar *et al.,* 2020).

A simulation test was conducted on a physical 10 kV Chinese feeder for the execution of the cluster coordinated optimization (CCO) method, which exhibits a dual approach: determining voltage handling capacity of internal constellation reactive generators and employing constellation sovereign optimization. Moreover, complex circuits can be converted into an evident network based on essential observational information and Ohmic parameters, as well as autonomously evaluated for every cluster. The concept of this optimization basically involves choosing a fundamental load among two nearby nodes, and the effectiveness of its outcomes have been demonstrated in comparison with other optimization techniques such as centralized optimization (CO), multiplier-based distributed inter-cluster optimization (MBDICO), and cluster independent optimization (CIO) (Lu *et al.,* 2020).

Heuristic nature-inspired algorithms, namely dragonfly-particle swarm optimization (DFPSO) show prominent results in the enhancement of power losses and voltage profiles. Hybrid optimization techniques exhibit a dual approach: dragonfly for optimizing switching sequences and PSO for DG allocation in view of different scenarios of loss reduction (only active power, only kVAR, or complex power). The outcomes are impressive in comparison without those without DG reconfiguration (WRDG) (Rafi *et al.,* 2020).

It is necessary to optimize a mixed-integer nonlinear programming objective function to increase the penetration level of renewable energy sources in the distribution network. This objective function was solved with the help of a two-stage technique named coyote optimization algorithm (COA) at variable load levels to minimize the energy losses of the system. This methodology started with a load flow analysis, followed by determining the DG location, size, and loss reduction ratio (LRR). Furthermore, a methodology was examined for two types of cases: the IEEE 123 bus benchmark and the Tai power 137 bus system. The test results were compared with those of mixed-integer linear programming (MILP), genetic algorithm (GA), PSO, biogeography-based optimization, (BBO), and grey wolf optimization (GWO) (Chang *et al.,* 2020).

The challenges of voltage regulation and power losses were addressed in HV and MV distribution networks (DN) via generalized benders decomposition (GBD), and a simulation test was carried out in the Jinzhai distribution network of China. The effectiveness was examined with global centralized optimization (GCO) and independent optimization (IO). Crest voltage compensation was adopted in order to obtain enhanced results in a master-slave network with the increased precision of the LinDistFlow equation. Unlike other optimization techniques, this one facilitates the complete utilization of the voltage regulation capacity, which reduces energy losses during off-grid generation and yields good results in terms of cost-saving and computational calculation. A multi-objective function was disintegrated into a main universal problem and auxiliary problems in order to generate an iterative process (Chai *et al.,* 2020).

A DG management approach (DGMA) was proposed as a supervisor unit for controlling and optimizing the operations of multiple inverter modules in a power grid in order to operate simultaneously in different dispatch segments. The proposed supervisor unit allows independent and subsidiary services to enable various modes of operation: maximum real power, power factor improvement subjected to connection/disconnection authorization, constant reactive power, and active power (Santos *et al.,* 2020).

A two-stage data-driven distributed robust optimization (DDRORO) model was implemented to overcome the challenges of intermittency in renewable source integration given the lack of informative characteristics as compared to robust and stochastic optimization approaches. Green energy generation and profit maximization were accomplished by increasing the penetration level of renewable energy. Consequently, in comparison with the other stochastic optimization (SO) models, cost-saving reached 20% via the optimal sizing and allocation of DG (Fathabad *et al.,* 2020).

The PSO method was studied in a 22 kV distribution system (DS) in Thailand to reduce the system energy losses caused by the inappropriate allocation of DG. Additionally, the concept of load transfer was used, and, to this effect, the tie-switch location was identified through this nature-inspired optimization algorithm. A comparison of results demonstrated reduced energy losses and a voltage profile increase of 40,38 and 6,59%, respectively, which took 3,13 h (Karaaom *et al.,* 2020).

To enrich the generating capacity of DG and smoothen voltage fluctuations, reactive power optimization was modeled via the general algebraic modeling system approach (GAMSA), although the utilization capacity of DG was reduced. The recognized multi-objective function (considering the voltage variant, the real power losses, and the DG capacity) was converted into a single objective function, which was subjected to various constraints such as power balance, balanced node, line power, node voltage, DG output, and reactive power compensation. The variations in active power, reactive power, and node voltage irrespective of load demand were demonstrated in optimized results (Wang *et al.,* 2020).

Compared to the absence of DG optimization (WDG), the importance of DG planning and allocation was justified with the help of the PSO approach in view of its enormous advantages, which include improved voltage profiles, reliability, and power quality. This was carried out in the Basuki Rahmat Surabaya feeder by using a graphical user interface system. Voltage and current-based injection matrices were used for power flow analysis, and the particle swarm optimization technique was then demonstrated with suitable results (Saidah *et al.,* 2020).

Cost minimization and voltage regulation were achieved by trivial alteration using a moth search optimization technique called corrected moth search optimization (CMSO). A multi-objective function was developed via the optimal placement of DG and shunt capacitors while concurrently optimizing and controlling the tap position of grid transformer. The results were also presented in comparison with other established optimization techniques such as the harmony search algorithm (HAS), the fireworks algorithm (FA), and the dynamic node priority list-genetic algorithm (DNPLGA) for the IEEE 33 and 118 test systems at different load conditions (peak, normal, and light) (Singh *et al.,* 2020).

A qualitative merging of genetic algorithm (GA) and ant colony optimization (ACO) resulted in a restructured algorithm dubbed genetic ant colony optimization (GACO), an approach that exhibited reduced energy losses and improved voltages. Load flow analysis was completed via the forward-backward method, and the newly developed technique was implemented with a faster convergence and free from local optima tracking. The results obtained were compared to those of renowned metaheuristic optimization techniques, which resulted in better voltage levels, average iterations (AI), reduced losses, and a faster convergence rate (Yang, 2020).

A multi-objective modified symbiotic organism search algorithm (MMSOSA) was executed for the optimization of the capacity and placement of renewable DG, with the purpose of achieving cost minimization (the cost of electricity for yearly purchases, the yearly expenditure, and the cost of operation), voltage regulation, and annual energy loss reduction at various load levels (constant, residential, commercial, and industrial) subjected to various constraints such as the power balance equation, the node voltage limit, and the DG capacity limit. Hourly demand variation and the total voltage deviation (TVD) curve were also significant for the improved results. Moreover, the results were quite impressive as compared to multi-objective symbiotic organism search (MOSOS) and the non-dominated sorting genetic algorithm-II (NDSGA) (Saha *et al.,* 2020).

The frequency regulation of a DG was carried out with the help of the moth swarm optimization technique, in which a series of PI-PD controllers were used through the incorporation of renewable energy-based sources (such as solar photovoltaic and wind power) and energy storage devices along with an electric vehicle. PI works in the primary stage and PD in the secondary one. Such a cascaded operation reduced the steady-state error of the system in comparison with conventional PID controllers, which may lead to an unstable transient response. For the execution in MATLAB and SIMULINK, the gain and time constant parameters of solar photovoltaic, wind power, fuel cell, and diesel engine generators, the electric vehicle, and the battery storage were taken as the nominal parameter of DG (Khamari *et al.,* 2020).

A battery storage architecture was proposed for the effective integration of distributed solar photovoltaic systems in order to avoid the intermittency of renewable energy sources. The flexibility of the grid, regardless of faults (circuit scale fault, substation scale fault, and generation plant scale factor), has also been understood to cause disturbances. Numerous types of battery storage architectures were identified based on different types of working principles: In-building Distributed Storage Architecture, Circuit/Distribution Storage Architecture, Substation/Microgrid Storage Architecture, Generation Plant Storage Architecture, and Utility/Grid Storage Architecture. As an outcome, grid resilience was maximized by optimizing the combination of battery storage systems to increase the penetration level (PL) of renewable energy (Confrey *et al.,* 2020).

A mathematical apprehension technique was proposed to enhance power and voltage quality via the amalgamation of DG in power distribution. This operation was carried out with the help of a flexible multi-level switch. The loading frequency and the power handling capability of the feeder were considered for the effective assimilation of DG through equalization. Moreover, a control framework for a PI controller and a steady-state converse model were developed in order to optimize the different modes of a flexible multi-level switch (Yu *et al.,* 2020).

The coordinated scheduling of renewable DG with the constraints of a futuristic smart grid was presented, and such a combination was focused on sorting out the disturbance caused by the intermittency of renewable energy sources. Firstly, renewable virtual sources were considered for the procedural implementation, and then the obtained characteristics were simulated in order to design the framework of the strategy. The recommended model showed potential regarding the use of pumped storage energy in order to minimize the intermittency of renewable energy generation. Effective forecasting was achieved with source shedding with the purpose of maximizing profits (Dong *et al.,* 2020).

A novel filter was used for the prediction of distributed solar power generation, in which the geographical proximity of the energy system affects the observations. Moreover, the intermittency in cloud formation and propagation was considered for the model at a resolution of 1 minute. This bi-level methodology consists of estimating PV power and measuring the lower frequency of the sampled data (Alam *et al.,* 2020).

A dumping cost to evaluate soil contamination was proposed along with a new framework for delivering the electrical energy at a unitary power factor level in order to minimize distribution energy losses. Consequently, the cost required for power congestion management was also reduced, thus enhancing the level of futuristic smart city power projects (Parida *et al.,* 2020).

The voltage regulation of a distribution system was improved with the use of solar PV DG as a reactive element that was independent from information exchange and feedback assessment. The backward/forward sweep algorithm was used while considering the irradiance level of solar energy and ambient temperature (Ammar *et al.,* 2019).

The cost of acquiring land is increasing day by day for the installation of a solar PV system as DG. An appropriate method was recommended by the authors for solar PV installation so that land could be utilized in a better way. The unusable land near the railway track area, water storage land, and highways could be used while offering remuneration to the owner instead of purchasing new land. A geographic information system was also developed to get the details of such lands having the appropriate irradiation level of solar energy. (Asanov *et al.,* 2019).

The increased risk power failure of renewable DG due to the intermittency of the climatological and atmospheric conditions has been highlighted with the proposal of an adaptive forecaster subgroup assortment strategy for enhancing the forecast. In this two-level approach, a binary genetic algorithm was used for selecting the characteristic, and a regression-based vector was applied to calculate the suitability level of the estimator. The accuracy was enhanced by 58,4% in comparison with the real prediction method (Eseye *et al.,* 2019).

A case study of wind power generation in DG to meet the demand of rural America was analyzed. In this context, 2015 separates the pre-wind and the post-wind eras. A graphical presentation was included showing the variation in the development of residential, commercial, and industrial customers over the last five years of the span. Along with this, wind speed, atmospheric temperature, and load demand were also taken into account (Madala *et al.,* 2019).

A simulation test was carried out to investigate the survival of wind power generation as distributed energy sources during the abnormal conditions of a power failure. A seven-scenario system (without DG, downstream, mid-way, upstream, and four types of wind power generation) was adopted to validate the results (Gumilar *et al.,* 2019).

Due to the progression of large heat pump projects, many countries are required to advance in the traditional grid system, which offers the possibilities of renewable energy sources. For the effective integration of heat pumps with wind power generation in the distribution network, a mathematical model was presented with the purpose of maximizing profits and reducing costs (Cui *et al.,* 2019).

Wind power generation in DG has been used for voltage profile enhancement, loss reduction, and ecological benefits in the distribution network. A power control curve optimization approach was adopted for the speed regulation of a windmill rotor. The optimized parameters were energy losses and voltage quality (Eltamaly *et al.,* 2019).

A multi-objective function of DG size and allocation optimization was considered regarding the operating, capital, environmental, wind, and light abandonment costs, which were subjected to voltage, current, equation flow, and DG capacity constraints. This was carried out with the implementation of the particle swarm optimization method in wind turbine, solar PV, and gas fueled micro turbine generators (Ma *et al.,* 2019).

The reliability of the radial distribution network could be increased by facilitating DG integration in the formation of multiple energy sources. PSO and the gravitational search algorithm (GSA) were used to obtain the optimal allocation of single and multiple DG. The reliability of the system was calculated in terms of numerous indices. The outcomes were evaluated in the IEEE 33-bus system for the various DG categories and compared to different metaheuristic approaches. The results were more significant in real and reactive power injection than in real power injection. The annual cost of energy losses (AELS) was better in comparison with GSA, PSO, the iterative algorithm (IA), and the krill herd algorithm (KHA) (Parihar *et al.,* 2021).

The optimal allocation and sizing of renewable energy-based DG could be found with the use of two different approaches in the same distribution network, namely the PSO and weighted-sum multi-objective (WSMO) approaches. The impact of optimal DG allocation was evaluated in the presence of harmonics distortion and validated in the IEEE 33-bus test system. In this strategy, a bi-level objective function was converted into a single objective function, and PSO was then applied. This approach demonstrated reduced energy losses, enhanced voltage profiles, and better reliability in the distorted distribution network. The results were compared with a GA, a multi-objective GA (MOGA), the analytical method (AM), and the backtracking search optimization algorithm (BSOA) (Parihar *et al.,* 2022).

Power quality parameters were also enhanced by incorporating a fuzzy logic controller (FLC) with hybrid optimization approaches. A hybrid optimization approach was proposed which integrates the features of the ant lion optimization algorithm (ALOA) and PSO. The multi-objective function was optimized in the presence of a training dataset offered by FLC, which resulted in optimized DG allocation. Such a process was evaluated in terms of load flow in the IEEE 33 test system. The results showed significant improvements stemming from the use of hybrid optimization techniques for different renewable energy-based power generators across a wide range of power factors. Said results were evaluated in comparison with those of ALOA and the proposed technique without FLC (Samala *et al.,* 2020b).

Another hybrid optimization approach was proposed by integrating the grasshopper optimization algorithm (GOA) and cuckoo search (CS) for the enhancement of various parameters of a distribution network in the presence of DG. This technique was evaluated in the IEEE 33 and 69 test systems under half load, full load, and overload conditions. The results were compared to those obtained via salp swarm optimization (SSO) and the lightning search algorithm (LSA) (Suresh *et al.,* 2020).

## Comparative assessment

An comparative assessment of renewable DG optimization is given in Table 1. The analysis was carried out in terms of the DG optimization technique, the test system, the parameters evaluated, and the optimized outcomes. The Table provides an insight into the literature on DG allocation in distribution networks.

**Table 1.** A comparative analysis of DG optimization

| Reference | Optimization technique | Test system | Evaluated parameters | Optimized results | Renewable DG | Non-renewable DG | Compared approaches | DG size (DGS) | DG location (DGL) | System losses (SL) | Voltage profile | Total cost (TC) | Calculation time (CT) | Fitness value (FV) | Power factor (pf) | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lu *et al.* (2020) | CCO | IEEE 123 | SL (MW) | 0,8319 | ✓ | | CO, CI, MBDICO | | | ✓ | ✓ | | ✓ | | | |
| | | | Vmax (p.u.) | 1,05 | | | | | | | | | | | | |
| | | | CT | 16,69 | | | | | | | | | | | | |
| Rafi *et al.* (2020) | DFPSO | IEEE 33, DG=1 | DGL | 29 | | ✓ | WRDG | ✓ | ✓ | ✓ | ✓ | | | | | |
| | | | DGS (kW) | 830 | | | | | | | | | | | | |
| | | | SL (kVA) | 135,44 + j 83,69 | | | | | | | | | | | | |
| | | IEEE 33, DG=2 | DGL | 29, 24 | | | | | | | | | | | | |
| | | | DGS (kW) | 830, 340 | | | | | | | | | | | | |
| | | | SL (kVA) | 118,40 + j 81,6 | | | | | | | | | | | | |
| | | IEEE 33, DG=3 | DGL | 29, 24, 20 | | | | | | | | | | | | |
| | | | DGS (kW) | 830, 340, 720 | | | | | | | | | | | | |
| | | | SL (kVA) | 116,7 + j 80,44 | | | | | | | | | | | | |
| Chang *et al.* (2020) | COA | IEEE 123 | SL (kW) | 18,2081 | ✓ | | MILP, GA, PSO, BBO, GWO | | | ✓ | ✓ | | ✓ | ✓ | | |
| | | | LRR | 77,01 | | | | | | | | | | | | |
| | | | V (p.u.) | 0,9939 – 1,0023 | | | | | | | | | | | | |
| | | | FV | 0,2299 | | | | | | | | | | | | |
| Chai *et al.* (2020) | GBD | HV and MV DN | SL (MW) | 0,6877, 0,0273, 0,0463, 0,055 | ✓ | | GCO, IO | | | ✓ | ✓ | ✓ | | | ✓ | |
| | | | TC (Yuan) | 333,1 | | | | | | | | | | | | |
| Santos *et al.* (2020) | DGMA | IEEE 13 | SL, p.f. | Near to the reference vaue | ✓ | | | | | ✓ | | | | | ✓ | ✓ |
| Fathabad *et al.* 2020 | DDDRO | IEEE 33 | TC (K$) | 17 015 | ✓ | | SO | | | | | ✓ | | | | |
| Karaaom *et al.* (2020) | PSO | 22 kV DS | SL (MWh) | 36,38 | | ✓ | WDG | | | ✓ | ✓ | | | | | |
| | | | Vrise (p.u.) | 0,97 | | | | | | | | | | | | |

| Reference | Optimization technique | Test system | Evaluated parameters | Optimized results | Renewable DG | Non-renewable DG | Compared approaches | DG size (DGS) | DG location (DGL) | System losses (SL) | Voltage profile | Total cost (TC) | Calculation time (CT) | Fitness value (FV) | Power factor (pf) | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Wang *et al.* (2020) | GAMSA | IEEE 33 | DGL | 4, 5, 6, 11, 25 | ✓ | | WDG | | ✓ | ✓ | ✓ | | | | | ✓ |
| Saidah *et al.* (2020) | PSO | IEEE 33 | DGL | 25 | | ✓ | WDG | ✓ | ✓ | ✓ | ✓ | | | | | |
| | | | DGS (MVA) | 9.5 | | | | | | | | | | | | |
| | | | SL (kW) | 1.56 | | | | | | | | | | | | |
| Singh *et al.* (2020) | CMSO | IEEE 33 | DGL | 13, 25, 30 | | ✓ | HAS, FA, DNPLGA | ✓ | ✓ | ✓ | ✓ | ✓ | | | | |
| | | | DGS (MW / MVAr) | 1,459, 0,987, 1,795 | | | | | | | | | | | | |
| | | | PL | 60.65% | | | | | | | | | | | | |
| | | | SL (MW) | 0,1964 | | | | | | | | | | | | |
| | | | TC (USD) | 64820 | | | | | | | | | | | | |
| | | IEEE 118 | DGL | 33, 38, 52, 71, 80, 96, 109 | | | | | | | | | | | | |
| | | | DGS (MW / MVAr) | 2,910, 3,225, 2,677, 4,414, 2,276, 2,159, 4,642 | | | | | | | | | | | | |
| | | | SL (MW) | 1,4622 | | | | | | | | | | | | |
| | | | TC (USD) | 478 229 | | | | | | | | | | | | |
| Yang (2020) | GACO | IEEE 33 | SL / kW | 139,4715 | | ✓ | ACO, GA | | | ✓ | ✓ | | ✓ | | | |
| | | | AI | 17,4 | | | | | | | | | | | | |
| Saha *et al.* (2020) | MMSOSA | IEEE 69 | DGL | 24, 26, 62 | ✓ | | MOSOS, NDSGA | ✓ | ✓ | ✓ | ✓ | ✓ | | | | |
| | | | DGS (kW) | 540,250, 1 325,5 | | | | | | | | | | | | |
| | | | SL (kWh) | 299 850 | | | | | | | | | | | | |
| | | | TC ($) | 783 670 | | | | | | | | | | | | |
| | | | TVD | 5 666 | | | | | | | | | | | | |
| Parihar *et al.* (2022) | WSMO, PSO | IEEE -33 | DGS (kW) | 2 588,4 | ✓ | | AM, MOGA, BSOA, GA | ✓ | ✓ | ✓ | ✓ | ✓ | | | | ✓ |
| | | | DGL | 6 | | | | | | | | | | | | |
| | | | Improved V % | 4,18 | | | | | | | | | | | | |
| | | | PDG cost ($/Hr) | 52,01 | | | | | | | | | | | | |
| | | | AELS ($) | 54 499,46 | | | | | | | | | | | | |

| Reference | Optimization technique | Test system | Evaluated parameters | Optimized results | Renewable DG | Non-renewable DG | Compared approaches | DG size (DGS) | DG location (DGL) | System losses (SL) | Voltage profile | Total cost (TC) | Calculation time (CT) | Fitness value (FV) | Power factor (pf) | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Parihar *et al.* (2021) | PSO & GSA | IEEE 33, DG=1 (Injects P) | DGS (kW) | 2 605 | ✓ | | IA, GSA, PSO, KHA etc. | ✓ | ✓ | ✓ | ✓ | ✓ | | | | ✓ |
| | | | DGL | 6 | | | | | | | | | | | | |
| | | | improved V % | 4.35 | | | | | | | | | | | | |
| | | | AELS ($) | 52 597 | | | | | | | | | | | | |
| | | IEEE 33, DG=3 (Injects P & Q) | DGS (kW) | 3 150 | | | | | | | | | | | | |
| | | | DGL | 6 | | | | | | | | | | | | |
| | | | improved V % | 6,19 | | | | | | | | | | | | |
| | | | AELS ($) | 76 774,6 | | | | | | | | | | | | |
| | | IEEE 33, DG=3 (Injects P & Q) | DGS (kW) | 859,2, 1 031,6, 605,3 | | | | | | | | | | | | |
| | | | DGL | 24, 30, 14 | | | | | | | | | | | | |
| | | | Improved V % | 10,07 | | | | | | | | | | | | |
| | | | AELS ($) | 101 477,8 | | | | | | | | | | | | |
| Samala *et al.* (2020b) | ALOA +PSO+ FLC | IEEE-33 (PV=2, u.p.f) | DGS (kW) | 385, 2 154 | ✓ | | PSO, ALOP+ PSO | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | |
| | | | DGL | 32, 7 | | | | | | | | | | | | |
| | | | TC ($) | 12 062 | | | | | | | | | | | | |
| | | IEEE-33 (WT=2, u.p.f) | DGS (kW) | 951, 696 | | | | | | | | | | | | |
| | | | DGL | 31, 17 | | | | | | | | | | | | |
| | | | TC ($) | 8 496 | | | | | | | | | | | | |
| | | IEEE-33 (PV=2, p.f.=0,85 lag.) | DGS (kW + KVAr) | 924 + j1223, 665 + j710 | | | | | | | | | | | | |
| | | | DGL | 32, 14 | | | | | | | | | | | | |
| | | | TC ($) | 8 037 | | | | | | | | | | | | |
| | | IEEE-33 (WT=2, p.f.=0,85 lag.) | DGS (kW + KVAr) | 993 + j1667, 606 + j913 | | | | | | | | | | | | |
| | | | DGL | 8, 30 | | | | | | | | | | | | |
| | | | TC ($) | 8 038 | | | | | | | | | | | | |
| Samala *et al.* (2020a) | ALOA +PSO | IEEE-33 | DGL | 5, 17, 31 | ✓ | ✓ | ALOA, GSA | ✓ | ✓ | ✓ | ✓ | | ✓ | | | ✓ |
| | | | DGS (MW) | 0,4571, 0,2478, 0,9987 | | | | | | | | | | | | |
| | | | SL (MVA) | 1,5257 | | | | | | | | | | | | |

| Reference | Optimization technique | Test system | Evaluated parameters | Optimized results | Renewable DG | Non-renewable DG | Compared approaches | DG size (DGS) | DG location (DGL) | System losses (SL) | Voltage profile | Total cost (TC) | Calculation time (CT) | Fitness value (FV) | Power factor (pf) | Reliability |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Suresh *et al.* (2020) | GOA+CS | IEEE-33, half load | DGL | 2 | | ✓ | SSO, LSA, GA | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| | | | DGS (kW) | 1 716 | | | | | | | | | | | | |
| | | IEEE-33, full load | DGL | 24 | | | | | | | | | | | | |
| | | | DGS (kW) | 926,99 | | | | | | | | | | | | |
| | | IEEE-33, 150% load | DGL | 24 | | | | | | | | | | | | |
| | | | DGS (kW) | 926,99 | | | | | | | | | | | | |
| | | IEEE-69, half load | DGL | 17 | | | | | | | | | | | | |
| | | | DGS (kW) | 1930,7 | | | | | | | | | | | | |
| | | IEEE-69, full load | DGL | 6 | | | | | | | | | | | | |
| | | | DGS (kW) | 1990,7 | | | | | | | | | | | | |
| | | IEEE-69, 150% load | DGL | 12 | | | | | | | | | | | | |
| | | | DGS (kW) | 1890,6 | | | | | | | | | | | | |
| Jana *et al.* (2020) | PSO | IEEE 34 | DGS (kW) | 58,928 | ✓ | | ADG | ✓ | ✓ | ✓ | ✓ | | | | | |
| Kumar *et al.* (2020) | PSOGWO | IEEE 33 | DGL | 6 | ✓ | | ADG | ✓ | ✓ | ✓ | ✓ | | | | | |
| | | | DGS (MW) | 2,576 | | | | | | | | | | | | |
| | | | SL (kW) | 103,966 | | | | | | | | | | | | |
| | | IEEE 33 | DGL | 30 | | ✓ | | | | | | | | | | |
| | | | DGS (MVAr) | 1,25 | | | | | | | | | | | | |
| | | | SL (kW) | 143,6 | | | | | | | | | | | | |
| | | IEEE 69 | DGL | 61 | ✓ | | | | | | | | | | | |
| | | | DGS (MW) | 1,869 | | | | | | | | | | | | |
| | | | SL (kW) | 83,9013 | | | | | | | | | | | | |
| | | IEEE 69 | DGL | 61 | | ✓ | | | | | | | | | | |
| | | | DGS (MVAr) | 1,328 | | | | | | | | | | | | |
| | | | SL (kW) | 152,4 | | | | | | | | | | | | |

**Source:** Authors

## Research gap

After assessing the planning and optimization of renewable DG systems, it was concluded that the intermittency of renewable energy resources is not properly considered in order to establish a stable and reliable distribution system. The integration of energy storage systems and demand response schemes are also underevaluated during optimal power dispatch. The utilization of energy storage devices can reduce the intermittent nature of natural resources, although the overall cost is increased. It is known that energy saving is also regarded as energy generation, so the end-user can shift their demand as per the demand response scheme. Load shifting from peak to valley hours enhances the quality of optimal power dispatch to reduce the system losses and imbalances between demand and supply. Consequently, battery storage and demand response integration are necessary for renewable DG

planning in order to increase the impact of natural energy sources and coordination between distribution companies and consumers.

## Conclusions

A comprehensive evaluation of recent trends in DG optimization is demonstrated in this paper, with its massive benefits subject to various constraints. These enormous benefits are associated with power quality, economic, and atmospheric parameters. The literature was reviewed based on optimization techniques, methodology, test system, optimized parameters, type of generation, enhanced outcomes, and compared approaches. The optimization techniques can be classified as classical, nature-inspired, and hybrid approaches depending on their methodology. Moreover, these techniques could be compared in terms of their convergence time, available literature, the dimensions of their objective function, and the complexity of their implementation. DG has the potential to exploit the dissemination of renewable energy sources in order to meet the present challenges as it contributes to safe environmental conditions. Moreover, the integration of a battery storage system can reduce the intermittency of renewable energy sources. Demand response is also an important aspect of the modern power grid. Renewable energy-based DG, battery storage, demand response, capacitor banks, and electric vehicles could be used for optimal power dispatch in distribution networks.

## References


Alam, S. S., Florita, A. R., and Hodge, B. (2020). Distributed PV generation estimation using multi-rate and event-driven Kalman kriging filter. *IET Smart Grid, 3*(4), 538-46. https://doi.org/10.1049/iet-stg.2018.0246

Allan, G., Eromenko, I., Gilmartin, M., Kockar, I., and McGregor, P. (2015). The economics of distributed energy generation: A literature review. *Renewable and Sustainable Energy Reviews, 42*, 543-556. https://doi.org/10.1016/j.rser.2014.07.064

Ammar, M., and Sharaf, A. M. (2019). Optimized use of PV distributed generation in voltage regulation: A probabilistic formulation. *IEEE Transactions on Industrial Informatics, 15*(1), 247-256. https://doi.org/10.1109/TII.2018.2829188

Asanov, I., Loktionov, E., and Sharaborova E. (2019, March 25-29). *Evaluation of potential distributed solar generation at right-of-way land of infrastructure objects* [Conference presentation]. 2019 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), Sochi, Russia. https://doi.org/10.1109/ICIEAM.2019.8742979

Chai, Y., Guo, L., Wang, C., and Y. Liu. (2020). Hierarchical distributed voltage optimization method for HV and MV distribution networks. *IEEE Transactions on Smart Grid, 11*(2), 968-980. https://doi.org/10.1109/TSG.2019.2928701

Chang, G. W., and N. Cong Chinh. (2020). Coyote optimization algorithm-based approach for strategic planning of photovoltaic distributed generation. *IEEE Access, 8*, 36180-36190. https://doi.org/10.1109/ACCESS.2020.2975107

Confrey, J., Etemadi, A. H., and Eveleigh, T. J. (2020). Energy storage system architecture optimization for grid resilience with high penetration of distributed photovoltaic generation. *IEEE Systems Journal, 14*(1), 1135-1146. https://doi.org/10.1109/JSYST.2019.2918273

Cui, Q., Bai, X., and Dong, W. (2019). Collaborative planning of distributed wind power generation and distribution network with large-scale heat pumps. *CSEE Journal of Power and Energy System, 5*(3), 335-347. https://doi.org/10.1109/JSYST.2019.2918273

Dong, H., Li, S., and Z. Tian. (2020). Coordinated scheduled strategy for distributed generation considering uncertainties in smart grids. *IEEE Access, 8*, 86171-86179. https://doi.org/10.1109/ACCESS.2020.2992342

DTI (2007). *Review of distributed generation. A joint Government/ Ofgem report*. https://www.ofgem.gov.uk/publications/review-distributed-generation-joint-governmentofgem-report

Eltamaly, A. M., Mohamed, Y. S., and Elghaffar, A. N. A. (2019, April 15-17). *Analyzing of wind distributed generation configuration in active distribution network* [Conference presentation]. 2019 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO), Manama, Bahrain. https://doi.org/10.1109/ICMSAO.2019.8880291

Eseye, A. T., Lehtonen, M., and Millar, R. J. (2019). Adaptive predictor subset selection strategy for enhanced forecasting of distributed PV power generation. *IEEE Access, 7*, 90652-90665. https://doi.org/10.1109/ACCESS.2019.2926826

Fathabad, A. M., Cheng, J., Pan, K., and Qiu, F. (2020). Data-driven planning for renewable distributed generation in distribution systems. *IEEE Transactions on Power Systems, 35*(6), 4357-4368. https://doi.org/10.1109/TPWRS.2020.3001235

Gumilar, L., Habibi, M. A., and Nugroho, W. S. (2020, February 20). *Analysis of short circuit on four types wind power plants as distributed generation* [Conference presentation]. 2020 International Conference on Smart Technology and Applications (ICoSTA), Surabaya, Indonesia. https://doi.org/10.1109/ICoSTA48221.2020.1570599089

Jana, C., Bala, M. J., and Goswami, S. K. (2020). *Finding of the probable size and location of distributed generation in unbalanced distribution system with demand uncertainty* [Conference presentation]. 2020 IEEE Calcutta Conference (CALCON), Kolkata, India. https://doi.org/10.1109/CALCON49167.2020.9106559

Karaaom, C., Jirapong, P., and Tantrapon, K. (2020, March, 4-6). *Optimal allocation of tie switch in distribution systems for energy loss reduction using particle swarm optimization* [Conference presentation 2020 8th International Electrical Engineering Congress (iEECON), Chiang Mai, Thailand. https://doi.org/10.1109/iEECON48109.2020.229528

Khamari, D., Sahu, R. K., and Panda, S. (2020, January 3-5). *Moth Swarm Algorithm based cascade PI-PD controller for frequency regulation of distributed power generation system with electric vehicle* [Conference presentation]. 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India. https://doi.org/10.1109/ICPC2T48082.2020.9071491

Kumar, K., Ramswaroop, R., Yadav, L. K., and Joshi, P. (2020, February 28 – March 1). *Optimal planning of distributed generation using hybrid metaheuristic approach* [Conference presentation]. 2020 IEEE 9th Power India International Conference (PIICON), Sonepat, India. https://doi.org/10.1109/PIICON49524.2020.9112955

Lu, H., Y. Chai, Y., J. Hao, J., and Chen, X. (2020). Network simplification-based cluster coordinated optimization method for distributed PVs with inadequate measurement. *IEEE Access, 8*, 65283-65293. https://doi.org/10.1109/ACCESS.2020.2984709

Ma, Q., Zhang, Y., and Wen, Y. (2019, December 1-4). *Research on optimization for siting and sizing of distributed generation considering wind and light abandonment and environment cost* [Conference presentation]. 2019 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), Macao, China. https://doi.org/10.1109/APPEEC45492.2019.8994588

Madala, S., Ramaekers, T., and Herink, A. (2019, April 28 – May 1). *The story of a rural public power district – Before and after distribution grid conneceted wind power generation* [Conference presentation]. 2019 IEEE Rural Electric Power Conference (REPC), Bloomington, MN, USA. https://doi.org/10.1109/REPC.2019.00018

Mahmoud, H. A., and Rama, K. (2017). A review of the optimal allocation of distributed generation: Object, constraints, methods, and algorithms. *Renewable and Sustainable Energy Reviews, 75*, 293-312. https://doi.org/10.1016/j.rser.2016.10.071

Parida, A., Choudhury, S., and Chatterjee, D. (2020, February 28 – March 1). *Optimal solar PV based distributed generation system*

*suitable for cost-effective energy supply* [Conference presentation]. 2020 IEEE 9th Power India International Conference (PIICON), Sonepat, India.

https://doi.org/10.1109/PIICON49524.2020.9113018

Parihar S. S., and Malik M. (2021). Optimal allocation of multiple DGs in RDS using PSO and its impact on system reliability. *Electronics and Energetics, 34*(2), 219-237. https://doi.org/10.2298/FUEE2102219P

Parihar S. S., and Malik M. (2022). Analysing the impact of optimally allocated solar PV-based DG in harmonics polluted distribution network. *Sustainable Energy Technologies and Assessments, 49*, 101784. https://doi.org/10.1016/j.seta.2021.101784

Rafi, V., and Dhal, P. K. (2020, January 20-22). *Loss minimization based distributed generator placement at radial distributed system using hybrid optimization technique* [Conference presentation]. 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India. https://doi.org/10.1109/ICCCI48352.2020.9104145

Saha, S., and Mukherjee, V. (2020, February 28 – March 1). *A novel meta-heuristic for optimal allocation of distributed generation in balanced distribution network considering hourly load variation* [Conference presentation]. 2020 IEEE 9th Power India International Conference (PIICON), Sonepat, India. https://doi.org/10.1109/PIICON49524.2020.9113054

Saidah, and Masrufun M. (2020, February 20). *Optimization of DG placement and size using PSO based on GUI* [Conference presentation]. 2020 International Conference on Smart Technology and Applications (ICoSTA), Surabaya, Indonesia. https://doi.org/10.1109/ICoSTA48221.2020.1570615982

Samala, R. K., and Kotapuri, M. R. (2020a). Distributed generation allocation in distribution system using particle swarm optimization based ant-lion. *International Journal of Control and Automation, 13*, 414-426. http://sersc.org/journals/index.php/IJCA/article/view/6004

Samala, R. K., and Kotapuri, M. R. (2020b). Optimal allocation of multiple photo-voltaic and/or wind-turbine based distributed generations in radial distribution system using hybrid technique with fuzzy logic controller. *SN Applied Sciences, 2*, 191. https://doi.org/10.1007/s42452-020-1957-3

Santos, L. N., Sousa, G. G., Salvatti, G. A., Carati, E. G., da Costa, J. P., Cardoso, R., Stein, C. M. O., and, Nadal, Z. L. I. (2020, February 26-28). *A distributed generation manager with support for distributed network operator commands* [Conference presentation]. 2020 IEEE International Conference on Industrial Technology (ICIT), Buenos Aires, Argentina. https://doi.org/10.1109/ICIT45562.2020.9067315

Saxena, V., Kumar, N., and Nangia, U. (2021, February 11-13). *Smart Grid: A Sustainable Smart Approach* [Conference presentation]. 2020 3rd International Conference on Computational & Experimental Methods in Mechanical Engineering ICCEMME (2021), Uttar Pradesh, India. https://doi.org /10.1088/1742-6596/2007/1/012042

Singh, P., Bishnoi, S. K., and Meena, N. K. (2020). Moth search optimization for optimal DERs integration in conjunction to OLTC tap operations in distribution systems. *IEEE Systems Journal, 14*(1), 880-888. https://doi.org/10.1109/JSYST.2019.2911534

Suresh, M. C. V., and Edward J. B. (2020). A hybrid algorithm based optimal placement of DG units for loss reduction in the distribution system. *Applied Soft Computing, 91*, 106191. https://doi.org/10.1016/j.asoc.2020.106191

Wang, Z., Liu, H., and Li, J. (2020, June 4-7). *Reactive power planning in distribution network considering the consumption capacity of distributed generation* [Conference presentation]. 2020 5th Asia Conference on Power and Electrical Engineering (ACPEE), Chengdu, China. https://doi.org/10.1109/ACPEE48638.2020.9136183

Yang, J. (2020, March 18-20). *Research on optimized reconfiguration of distributed distribution network based on ant colony optimization algorithm* [Conference presentation]. 2020 International Conference on Computer Engineering and Application (ICCEA), Guangzhou, China. https://doi.org/10.1109/ICCEA50009.2020.00012

Yu, C., Zhang, G., and Peng, B. (2020, June 4-7). *Feeder power flow control strategy for flexible multi-state switch with joint access to the distributed generation* [Conference presentation]. 2020 5th Asia Conference on Power and Electrical Engineering (ACPEE), Chengdu, China. https://doi.org/10.1109/ACPEE48638.2020.9136556

FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 423-428 | e-ISSN: 2347-470X

# Simulation and Optimization of Ultrasonic Transducer

**Arkya Bagchi[1], Akash Chowdhary[2], Abhishek Takki[3] and Raminder Kaur[4]**

*[1,2,3,4]Department of Applied Chemistry, Delhi Technological University, Delhi, India, [1]arkyabagchi@gmail.com, [2]akashchowdhary2211@gmail.com, [3]abhishek1999245@gmail.com, [4]raminderkaur@dce.ac.in*

***Correspondence:** Arkya Bagchi; Email: arkyabagchi@gmail.com

**ABSTRACT-** The ultrasonic transducers have numerous applications in industries, including medical probes for performing ultrasound scans. One of the significant drawbacks of the ultrasonic transducer is the wastage of a large portion of energy, due to high acoustic impedance, while transmitting ultrasonic waves to the target object. The present study is aimed to investigate the material design of the piezo-composite transducer and improve its performance. Different piezo-composite transducers were simulated in the COMSOL environment by varying input parameters, and three key performance indicators (KPI) were calculated. Many constraint-based multivariable optimization algorithms have been used to maximize the KPIs. A set of parameters, such as Sensitivity and Fractional Bandwidth, have been found to increase the performance of piezo-composite transducer model and its overall efficiency. This study is intended to impinge unidirectional property to the transducer which is found to be beneficial in more accurate medical as well as structural reports and cost savings.

**General Terms:** Optimisation

**Keywords:** Ultrasonic transducer; Piezo-composite; Regression; Optimisation Algorithm.

ARTICLE INFORMATION

**Author(s)**: Arkya Bagchi, Akash Chowdhary, Abhishek Takki and Raminder Kaur
**Received**: 05203/2022; **Accepted**: 15/07/2022; **Published**: 10/08/2022;
**e-ISSN**: 2347-470X;
**Paper Id**: IJEER100303;
**Citation:** 10.37391/IJEER.100303
**Webpage-link**:
https://ijeer.forexjournal.co.in/archive/volume-10/ijeer-100303.html

This article belongs to the Special Issue on **Recent Advancements in the Electrical & Electronics Engineering**

**Publisher's Note:** FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

## 1. INTRODUCTION

An ultrasonic transducer is a device used for the production of ultrasonic vibration. Ultrasonic transducers are classified into many categories, based on energy source and the mode through which these ultrasonic waves are propagated. Electromechanical transducers are far more adaptable than mechanical or liquid driven transducers and include piezoelectric and magneto strictive devices. An ultrasonic transducer works when an electrical signal is applied to the piezoelectric crystal, and the crystal starts to vibrate, and this vibration leads to the formation of acoustic waves, also called Ultrasound waves with a frequency range in 2-8 MHz. The reverse process receives the echo generated back to the crystal and converts it to the electrical signal. This ultrasonic transducer has a variety of applications in many domains, and one of the significant usages is in the medical industry for ultrasound imaging. Medical Ultrasonic probes consist of piezoelectric material, backing material, matching layer and an acoustic lens. One of the essential characteristics of piezoelectric material is the smooth conversion of electrical to mechanical energy. When the element is excited by the external electrical source, it shows a high acoustic impedance. A large portion of acoustic energy gets wasted while transmitting the ultrasonic waves onto the target object. Initially, Piezoelectric ceramics like Lead Zirconate titanate (PZT), Lead-based relaxor family and piezoelectric single crystals: LN, PMN-PT, PIN-PMN-PT were commonly employed as they have high electromechanical coupling and low dielectric losses. With advancement in technology over the time, the piezo-polymers such as polyvinylidene difluoride and its copolymer with trifluoro ethylene were used.

But, nowadays, with advancements in material design, piezocomposites give far better performance than piezo-ceramics and piezo-polymers. The matching layer is attached, which helps in acoustic matching with the target object, provides good electromechanical coupling and reduces the electrical and mechanical losses. Alumina and silver loaded epoxy are commonly used for designing the matching layer. The backing layer is used to damp out the vibrations from the piezoelectric crystal, and it also acts as a supporting material for the transducer component. The transducer's performance (also called KPI) depends on three different parameters i.e sensitivity, fractional bandwidth, and the center frequency. The KPIs are calculated from the impedance curve of the transducer.

The main objective of this study is to optimize these KPIs and improve the performance of ultrasonic transducers. We studied various materials among conventional piezoelectric ceramics, piezoelectric polymers and piezoelectric composite. The behaviour of piezocomposites was the best suitable to carry out our research on ultrasonic transducers. Different piezo-composite transducers were simulated in the COMSOL environment by varying input parameters. The associated KPIs

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 423-428 | e-ISSN: 2347-470X

(Key Performance Indicators) were calculated from the Impedance vs Frequency graph obtained from the simulation. Constraint-based multivariable optimization algorithms were used to optimize the KPIs of transducers. Our work is unique in the sense that we determined and optimized the KPIs using optimization algorithms that impinges unidirectional property to the transducer. No such work has been done earlier within this particular domain. The overall paper consists of 5 different components. The first part comprised of introduction; the second part presents the past studies in this field. The third part details the methodology and implementation respectively. The fourth part states the experimental results and summarises the conclusion.

## 2. LITERATURE REVIEW

B. Hailu et al. [1] Presents the comparison of different piezoelectric Materials for the design of transducers. Majorly three piezoelectric materials were proposed which were, lead zirconate titanate, modified lead titanate, and polyvinylidene difluoride, in combination with various 1-3 piezoelectric composite configurations. Then the finite element analysis was carried out to compare each embedded transducer configuration as a function of various different parameters to examine the potential behaviour and capabilities of the device in many different circumstances. With the use of the Pzflex modelling environment they were able to monitor the phenomenon. This method is not very efficient and time consuming. In our case we implemented optimization algorithms from Scipy library package which efficiently found the right parameters. [2] highlighted the need for artificial neural network models to describe the mapping relation between piezocomposite design parameters and performance parameters. Due to high time consumption of FEM simulations, artificial neural network models were established. In our case, we have used Comsol software to carry out FEM simulations. Also, we've carried out simulations for a larger set of data points for more accurate results and tried out different optimization algorithms to obtain the most suitable result. Bareiro et al. [3] focuses on optimising a group of parameters in ultrasound transducers using a genetic algorithm (GA) which are assessed by simulating B-mode scans for every single transducer configuration. The paper also emphasises on the significance of optimisation in the medical industry. The paper proposes a software-based and fully automatic optimization of transducer arrays so as to enhance ultrasound image quality by the help of ultrasound simulator and a genetic algorithm. [4] presents an electrical equivalent circuit model for piezoelectric transducers to analyse the performance of ultrasonic transducers by providing a close agreement between simulated and experimentally observed dynamic responses of the piezoelectric transducers which were designed and fabricated with composite sensor materials and epoxy-tungsten composite as a backing material. It introduces important characteristics of the transducers, like, electrical and acoustic impedance matching along with frequency response were simulated. Qifa et al. [5] underlines the excellent properties of piezoelectric single crystals that have extensively been employed for various sensors and actuators applications. Recent development of both the single-element and array transducers fabricated using the single crystals were presented. In our paper, we've targeted majorly 3 important key factors and summarised how we can optimise the performance of ultrasonic transducers to achieve better results. [11] demonstrated the design of ultrasound imaging probe and emphasises upon the importance of pulse-echo measurement for the precise measurement of electroacoustic properties of the probe. The significance of the Finite Element Method to deeply comprehend the model design, electro-acoustical performance parameters and their optimisation has also been explained. The transducer used was of a single array element type and the design consisted of a hard rubber backing substrate, four acoustic matching layers, standard piezoelectric ceramic and silicon rubber lens. In our case, we've used a similar approach, but instead of a single array element, we've carried out the simulations on a piezocomposite for several advantages like significant reduction in lateral resonances, improved electromechanical coupling coefficient and improved Tx/Rx Sensitivity. Grimaldi [12] discusses about the downsides of ultrasonic sensor performance and optimization of its parameters to improve the performance. This optimization technique is based on signal processing format. The only disadvantage is that it is only applicable to those which can be represented by transfer function whereas in our case we have used optimization algorithms which are computationally efficient and determines optimised parameters easily.

## 3. METHODOLOGY

A technique for estimating BP using 1D-CNN is proposed. CNN is effective technique when the input signal is highly correlated between neighboring samples. As shown in *Figure 2*, neighboring samples of PPG are correlated. Unlike the other deep natural networks, CNN maintains temporal information of input data so that CNN is suitable for the locally correlated input data such as PPG signals. 3 CNN model are proposed and compared.

### 3.1 Material Selection

After going through the literature review survey, we studied majorly three different categories of materials which are piezoceramics, piezopolymers and piezocomposites. Out of these piezocomposites were found to have the most outstanding performance. For designing the matching layer material Alumina loaded epoxy was considered. It has excellent acoustic matching properties with the target object, providing good electromechanical coupling and reducing the electrical and mechanical losses. For designing the backing layer Tungsten loaded epoxy is considered. They damp out the vibrations from the piezoelectric crystal, and also acts as a supporting material for the transducer component. In piezocomposites a single transducer block having multiple materials layered polymer matrix is placed next to each other and the pillars are mainly made up of PZT material and are part of the transducer. The choice of the polymer matrix depends also on the type of design we require. For example, when designing phased array devices, we avoid crosstalk (energy coupling laterally across the device)

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 423-428 | e-ISSN: 2347-470X

which is unwanted. Crosstalk adds noise to receive signals and can disrupt the directivity. Soft materials have better characteristics for reducing crosstalk as their attenuative properties are generally higher than stiffer materials.

## 3.2 Parameter Selection

We have studied many parameters to understand the transducer working. The results that allow us to analyse the performance of a transducer in general are called Key Performance Indicators (KPIs). The Sensitivity and Fractional Bandwidth are two important figures of merit, two KPIs of ultrasonic transducers are determined by matching layer thickness (MLT), composite volume fraction (CVF) and central frequency of Ultrasonic Transducers. Sensitivity is defined as the ratio of transmitter signal to receive signal after the 1st reflection of the ultrasonic waves. Fractional bandwidth is a fraction of a wide range of returning echo frequencies that are captured from a receive pulse. Matching layer is basically the layer located between piezoelectric element and the acoustic lens to help transfer the ultrasound energy from the elements to the medium. It is essential for steady penetration of the acoustic energy from the transducer to the body tissue and for easily tracing the reflected acoustic waves (the returning echo) back without much interruption to the transducer for detection. It also prevents the loss of transmission.

## 3.3 Model Design and Simulation

We started by designing a simple model of piezoelectric transducer in Onscale Solve. It comprised majorly of 5 components. The outer layer which was assigned with structural steel as the material. The insulation layer of Silicone Rubber. The Backing layer of Tungsten-loaded epoxy and the Matching Layer of unidirectional piezocomposite. Then, we configured the coordinate system to the Global coordinate system and defined all the key points as well as contact regions. We further created a mesh for the complete model and optimised it by analysing the element quality. Then, we inserted a time dependent forcing function to define the voltage loads across the piezoelectric material and other boundary conditions so that we can simulate the output parameters. Further it was replicated to form a linear patterned composite and the respective materials were assigned to the blocks. We exported the model to COMSOL in which we carried out the simulation. *Figure 1* shows the model design of piezoelectric transducer.

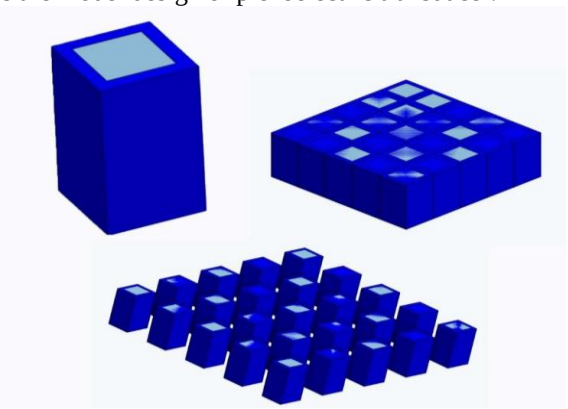


**Figure 1:** Piezoelectric Transducer Model Design

## 3.4 Dataset Formation

After performing the simulation, we created the dataset which involved the input parameters (matching layer thickness and composite volume fraction) and obtained the output parameters (sensitivity and fractional bandwidth). Results obtained from simulation were used to create each row of the dataset. Initially 25 simulations were performed with constrained input parameters (matching layer thickness: (0.4,0.8), composite volume fraction: (40,60), central frequency close to 1 MHz).

## 3.5 Optimization

Our main objective was to maximise the KPI's and with increase in volume fraction the sensitivity increases but the fractional bandwidth decreases. There should be an optimum value of inputs at which these KPIs are maximised. For this purpose, we took a single feature by multiplying both sensitivity & fractional bandwidth. Now for optimising this product term we formed an equation which satisfies both inputs values and the product term. This was done with the help of regression analysis in MS Excel. Fig. 2 shows the variation of sensitivity and fractional bandwidth with matching layer thickness.

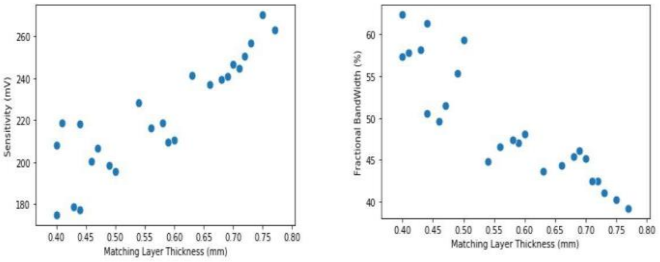



**Figure 2:** Variation of MLT with Fractional bandwidth and sensitivity

### 3.5.1 Linear Regression

In Linear regression [6], features of degree 1 are used to fit the data points with a straight line. Within Excel, Analysis ToolPak is used to perform the regression. The required features and output variable are selected and this ToolPak automatically performs the regression analysis. The problem with the linear regression model is that it is prone to overfitting and not robust with outliers. We used the R2 metric and residual plot to evaluate its performance. The coefficient values represent the coefficient of features.

### 3.5.2 Polynomial Regression

In polynomial regression [7], higher degree of features used to fit the data points. 2- and 3-degree terms are generated from matching layer thickness & volume fraction. Different combinations of terms are used to form the equation for optimization.

The results of the regression analysis is mentioned within the results and discussion section. From regression analysis we obtained the equation to be optimised for maximising the KPI's. This is a multivariate constrained optimization problem. Defined the bound constraints for matching layer thickness and volume fraction. The open-source Python library for scientific computing called SciPy provides a suite of global & local

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 423-428 | e-ISSN: 2347-470X

optimization algorithms. Three local optimization algorithms are used in our study. They are Trust Region constrained Algorithm [8], Sequential quadratic programming (SQP) [9] and BFGS algorithm [10]

# 4. RESULTS AND DISCUSSION

## 4.1 Simulation Results

This section discusses the graphs obtained by simulating the piezoelectric transducer on COMSOL environment. *Figure 3* shows the receive signal on the device. It is the excitation pulse from which we get two return/receive signals. From this we are going to look at the first one in detail. *Figure 4* shows the impedance curve. We post-processed this first reflection by plotting Voltage inputs as a function of Time and applying Fast Fourier Transform on the receive pulse to obtain Piezo Impedance plot as a function of Frequency so as to look at the frequency content and the amplitude of the signal. Then, the two plots were superimposed to calculate the sensitivity. *Figure 5* shows the frequency spectrum of the received pulse which is superimposed on receive voltage pulse. Impedance spectrum of the piezocomposite device was again plotted separately as shown in *Figure 6* to determine the improved performance of the device in terms of lateral resonances, electromechanical coupling coefficient and sensitivity. *Figure 6* shows impedance spectrum of piezo-composite device.

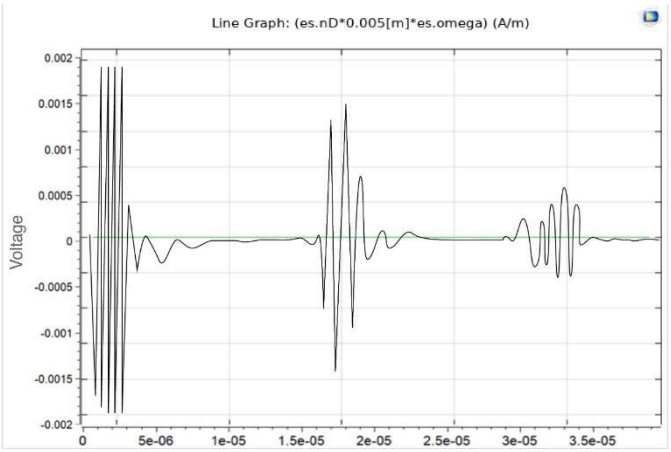



**Figure 3:** Receive Signal on Device

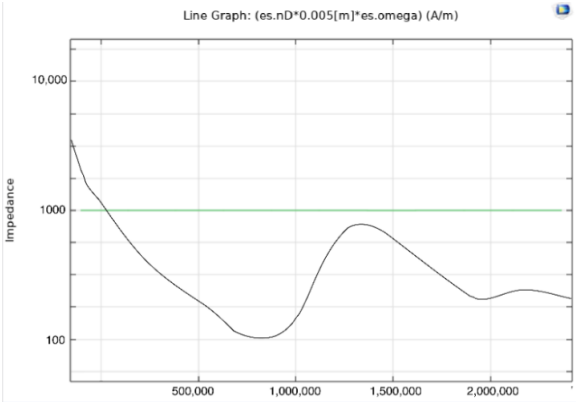



**Figure 4:** Impedance Curve

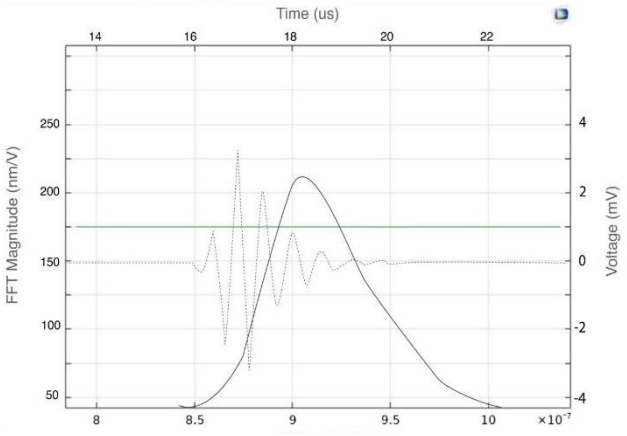



**Figure 5:** Frequency spectrum of receive pulse imposed on voltage pulse

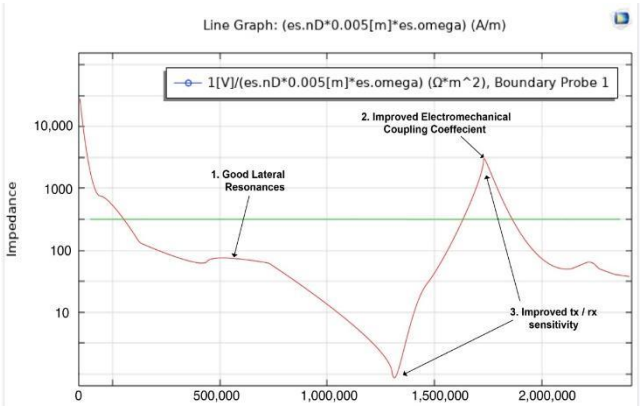



**Figure 6:** Impedance spectrum of piezocomposite device

## 4.2 Optimization Results

This section discusses the outputs after performing the optimization. After calculating the KPIs in order to understand the variation of inputs with outputs, we plotted several graphs which shows the relationship between inputs & outputs. *Figure 7, 8, 9* shows the variation of frequency, fractional bandwidth & sensitivity as function of input variables *i.e.* MLT and CVF *Figure 10* plots the variation of inputs with outputs along with sensitivity bandwidth product as pair-grid. Matching layer thickness shows certain correlation with KPIs.

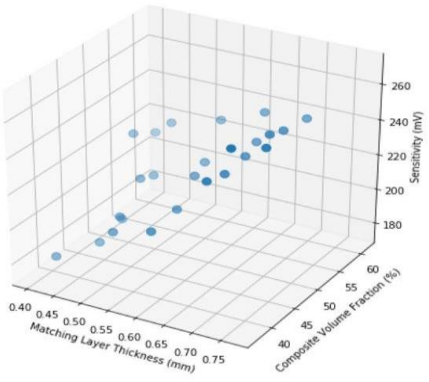



**Figure 7:** Variation of sensitivity with input parameters

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 423-428 | e-ISSN: 2347-470X

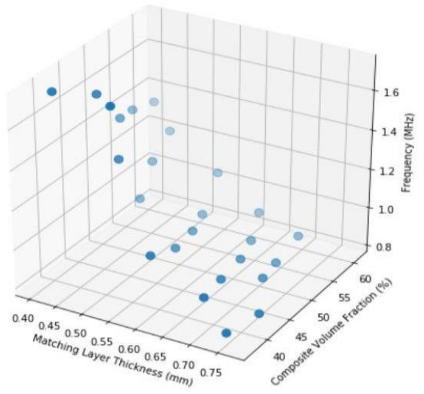



**Figure 8:** Variation of frequency with input parameters

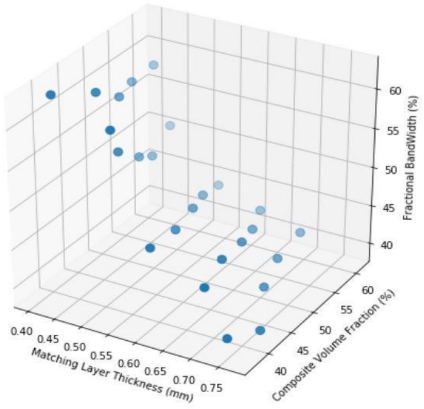



**Figure 9:** Variation of fractional bandwidth with input parameters

Linear & Polynomial Regression was used to create the equation for optimization function. The results of the regression analysis are mentioned below. $R^2$ value is calculated for each type in order to compare each model performance. *Table 1* shows the different combination of 1,2- and 3-degree features for linear and polynomial regression along with adjusted $R^2$ value.

**Table 1: Regression Analysis of Input Features**

| Regression Type | Adjusted $R^2$ value | Features Involved |
|---|---|---|
| Linear Regression | 0.03 | MLT, CVF |
| Polynomial Regression | 0.20 | MLT, $MLT^2$, CVF |
| Polynomial Regression | 0.19 | MLT, $MLT^2$, CVF, $CVF^2$ |
| Polynomial Regression | 0.29 | MLT, $MLT^2$, $CVF^2$ |
| Polynomial Regression | 0.20 | MLT, $MLT^2$, $CVF^3$ |

Polynomial regression model with MLT, $MLT^2$, $CVF^2$ fits the data points really well with high $R^2$ value. *Fig. 11, 12, 13* plots the residuals of the accepted features. Different optimization algorithms are used to maximise the sensitivity bandwidth product with the help of objective function obtained from regression analysis. *Table 2* shows the optimisation algorithm used then optimised values for MLT & CVF and its corresponding sensitivity bandwidth product.

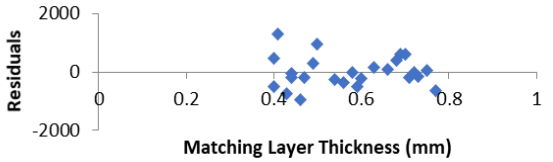



**Figure 10:** Residual plot for MLT

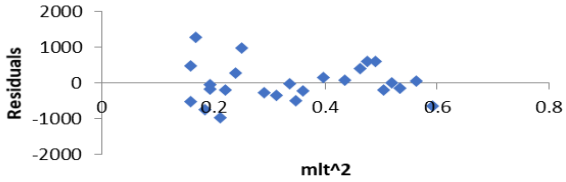



**Figure 11:** Residual plot for MLT2

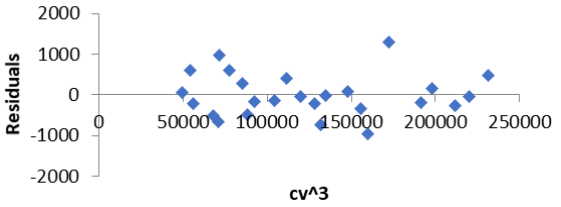



**Figure 12:** Residual plot for CV3

**Table 2: Optimisation Algorithm Results**

| Algorithm | Optimised Values (mm, %) | Sensitivity-Bandwidth Product |
|---|---|---|
| Trust Region constrained algorithm | (0.40, 61.39) | 11931.8 |
| Sequential Least Squares Programming | (0.40, 40.78) | 10909.5 |
| BFGS algorithm | (0.44, 45.23) | 10878.9 |

## 5. CONCLUSION AND FUTURE REMARKS

This paper has found the parameters (Matching Layer Thickness and Composite Volume Fraction) which optimise the KPIs (Sensitivity, Fractional Bandwidth and Central Frequency) by the help of various optimisation algorithms and thus result in increased performance and overall efficiency of the piezocomposite transducer model. There are certain drawbacks in our analysis, as we have performed limited number of simulations. This may hamper the accuracy of optimized figures resulting from optimization. Apart from that more computationally efficient, robust and effective optimisation techniques can be used. In our study, we used MLT and CVF as input parameters. More variables could result in better performance of transducer. We can conclude that with

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 3 | Pages 423-428 | e-ISSN: 2347-470X

optimized values, ultrasonic transducer can be designed in laboratory which would have several different applications. In medical field, a high-performance transducer could result in good resolution in eye imaging, noninvasive imaging of skin morphology, multi-modality in photoacoustic imaging etc. In industries more advanced automated machinery, object navigation etc. It has more diverse future applications in agriculture, military, aerospace, chemical and construction. With use of optimization algorithms like genetic algorithm, particle swarm optimization better results can be derived.

## ACKNOWLEDGEMENTS

We would like to express our gratitude to our supervisors, Dr. Raminder Kaur for their continuous guidance throughout this project. Their timely advice, meticulous scrutiny, scholarly advice and scientific approach have helped us to a great extent to accomplish this task. We would also like to thank S.G Warker for helping us and providing the infrastructure for completion of our project.

## REFERENCES

[1] B. Hailu, G. Hayward, A. Gachagan, A. McNab, and R. Farlow, "Comparison of different piezoelectric materials for the design of embedded transducers for structural health monitoring applications," Proc. IEEE Ultrason. Symp., vol. 2, pp. 1009–1013, 2000, doi: 10.1109/ULTSYM.2000.921495.

[2] K. Wierzbanowski, J. Tarasiuk, and A. Lodini, "Optimization of Material Properties Using Genetic Algorithms," Mater. Sci. Forum, vol. 652, pp. 1–6, 2010, doi: 10.4028/WWW.SCIENTIFIC.NET/MSF.652.1.

[3] "(16) (PDF) Development of the Fabrication Process and Characterization of Piezoelectric BaTiO 3 /Epoxy Composite Used for Coated Ultrasonic Transducer Patterns in Structural Health Monitoring." https://www.researchgate.net/publication/320161385_Development_of_the_Fabrication_Process_and_Characterization_of_Piezoelectric_BaTiO_3_Epoxy_Composite_Used_for_Coated_Ultrasonic_Transducer_Patterns_in_Structural_Health_Monitoring (accessed Apr. 28, 2022).

[4] "Simulation and optimization of piezocomposite transducers with a practical example | OnScale." https://onscale.com/blog/simulation-and-optimization-of-piezocomposite-transducers-with-a-practical-example/ (accessed Apr. 28, 2022).

[5] Q. Zhou, K. H. Lam, H. Zheng, W. Qiu, and K. K. Shung, "Piezoelectric single crystals for ultrasonic transducers in biomedical applications," Prog. Mater. Sci., vol. 66, p. 87, 2014, doi: 10.1016/J.PMATSCI.2014.06.001.

[6] "Linear Regression for Machine Learning." https://machinelearningmastery.com/linear-regression-for-machine-learning/ (accessed May 02, 2022).

[7] "Understanding Polynomial Regression!!! | by Abhigyan | Analytics Vidhya | Medium." https://medium.com/analytics-vidhya/understanding-polynomial-regression-5ac25b970e18 (accessed May 02, 2022).

[8] "Trust-region methods - optimization." https://optimization.mccormick.northwestern.edu/index.php/Trust-region_methods (accessed May 02, 2022).

[9] "Sequential Quadratic Programming | NEOS." https://neos-guide.org/content/sequential-quadratic-programming (accessed May 02, 2022).

[10] "A Gentle Introduction to the BFGS Optimization Algorithm." https://machinelearningmastery.com/bfgs-optimization-in-python/ (accessed May 02, 2022).

[11] Spicci, L. (2013). FEM Simulation for 'pulse-echo' Performances of an Ultrasound Imaging Linear Probe.

[12] Grimaldi, U., and M. Parvis. "Enhancing Ultrasonic Sensor Performance by Optimization of the Driving Signal." Measurement 14, no. 3-4 (1995): 219–28. doi:10.1016/0263-2241(94)00028-6

© 2022 by Arkya Bagchi, Akash Chowdhary, Abhishek Takki and Raminder Kaur. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Solar Power Driven Vehicles

***Prajakta Tiwari, Anjali, Zeenat Aara, Dr. J.P. Kesari***

Delhi Technological University

## 1.Introduction

With the fast depletion of fossil fuels and growing pollution, it has come to be critical for us to offer an exchange choice to the primitive fossil fuels. An alternative that's renewable in addition to secure for our environment. The first aspect which involves our thoughts is sun energy. Solar power may be transformed into electric power and effectively utilised for walking the automobiles. Moreover, it even doesn't launch any pollutant in an effort to be dangerous to us and our environment. Using batteries to energy automobiles and shop electric powered power at the grid have been as soon as idea to be absolutely uneconomic propositions. But nowadays the use of batteries to carry out those features isn't always handiest possible, however gives benefits over the incumbent technology of fossil gasoline technology and inner combustion

## 2.The advent of solar cars and electric vehicles

It turned into in 1955 that the primary sun automobile version turned into displayed on the Chicago Powerama conference with the aid of using William G. Cobb, a General Motors employee. In 1976, the primary sun automobile that would be legally pushed turned into launched. In 1981 Hans Tholstrup and Larry Perkins made a sun powered race car; quiet achiever "Solar trek" s. In 1982, the pair has become the primary to move a continent in a sun automobile, from Perth to Sydney, Australia. Hans Tholstrup have become the writer of the World Solar Challenge competition. Following that, a whole lot of sun automobile races have been deliberate to elevate public awareness. It turned into accompanied with the aid of using numerous big improvements within side the sun automobile technology. But it's miles within side the previous few a long time that foremost contributions and tendencies had been made within side the sun electricity utilisation.

## 3. Materials and Methods

### 3.1Components of a solar car

The Components are Solar Array, Power Trackers, Batteries, Motor and Controller, Solar car brakes. They collect the energy from the sun and convert it into usable electrical energy.

#### 3.1.1 Solar Photovoltaic

A sun panel (photovoltaic module or photovoltaic panel) is a packaged, linked meeting of sun cells, additionally referred to as photovoltaic cells. Because a unmarried sun panel can produce best a restrained quantity of power, many installations include numerous panels. A photovoltaic machine generally consists of an array of sun panels, an inverter, and on occasion a battery and interconnection wiring.Solar panels use mild energy (photons) from the solar to generate strength via the photovoltaic effect. The structural (load carrying) member of a module can both be the pinnacle layer or the returned layer. The majority of modules use wafer-primarily based totally crystalline silicon cells or thin-movie cells primarily based totally on cadmium telluride or silicon. The carrying out wires that take the modern-day off the panels can also additionally include silver, copper or different non-magnetic conductive transition metals.

#### 3.1.2. Power Trackers

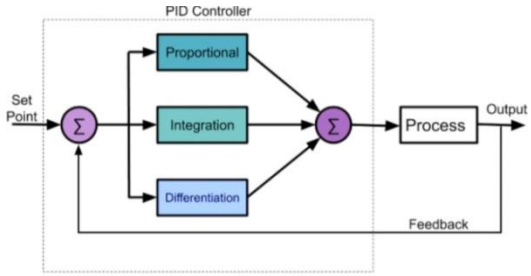
A Power tracker converts the output of sun mobileular into right voltage to be utilized by batteries to force the motor. In the absence of a energy tracker, the DC motor machine might bring about terrible performance. In order to hold the output of the sun mobileular consistent and excessive, most energy factor trackers are regularly utilized in maximum a hit sun automobiles. This is completed via way of means of the usage of a chip (sensor) within side the MPTT to constantly degree the output voltage of the sun mobileular, evaluating it with the constant battery voltage, after which figuring out the first-rate voltage to price the battery. This permits the power from the sun cells to attain performance more than 92%. Motor controllers have emerge as crucial digital additives of hybrid and electric powered automobiles via way of means of distinctive feature of switching big quantity of currents at excessive velocity with none transferring parts. A energy tracker can even modify the quantity of strength furnished via way of means of the battery structures to save you them from being damaged. The output of the energy tracker is hooked up to the batteries in which the energy may be

saved for the motor controller

**3.1.3. Batteries**

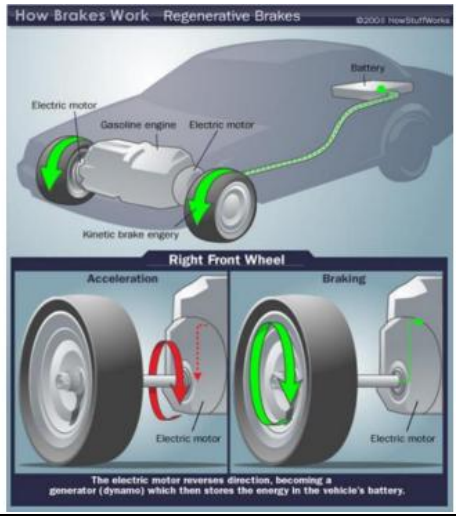
Batteries in sun motors are used as opposed to gas tanks in a traditional vehicle. The battery shops the incoming sun electricity in a chemical shape while the motor of the sun vehicle isn't always running. Batteries offer solid electricity to the DC cars and as a result, the cars will paintings maximum efficaciously as there aren't anyt any interruptions in the quantity of electricity supplied. The intermittent electricity fluctuations brought about due to daylight interference are triumph over via way of means of a battery electricity supply. Power from the batteries is likewise had to save electricity for night time time or cloudy day operation. Different kind of batteries including lead–acid battery, nickel cadmium (NiCd) battery, nickel steel hydride (NiMH) battery, lithium ion battery, and lithium polymer battery may be used for sun motors. The maximum typically used batteries are crafted from lead–acid, lithium-ion, and nickel–cadmium. Common lead–acid batteries of the sort used withinside the common own circle of relatives vehicle are too heavy at the same time as others are light-weight and greater green however are expensive. Solar-powered motors commonly perform in batteries that variety from eighty to a hundred and seventy

**3.1.4 PID Controller:**




PID Control (proportional-fundamental-spinoff) is through a ways the widest kind of automated manage utilized in industry. Even aleven though it has a fantastically easy algorithm/structure, there are numerous diffused versions in how it's miles implemented in industry. A proportional fundamental spinoff controller (PID controller) is a common manage loop remarks mechanism broadly utilized in business manage systems. A PID controller will accurate the mistake among the output and the favored enter or set factor through calculating and supply an output of correction as a way to alter the technique accordingly.

**3.1.5 Solar car brakes**




Regenerative braking is a braking technique that utilises the mechanical strength from the motor through changing kinetic strength into electric strength and fed again into battery supply. Theoretically, the regenerative braking device can convert an amazing fraction of its kinetic strength to price up the battery, the use of the equal precept as an alternator. In regenerative braking technique, it makes use of the motor to sluggish down the automobile whilst the motive force applies pressure to the brake pedal, then the electrical motor works in opposite direction, thereby slowing the automobile.

**3.2 Working Principle**

The sun module established at the pinnacle of vehicle is used to price the batteries thru price controller. A a hundred and forty WP sun module is used with output starting from 24V to 25V at STC. The batteries are to begin with completely charged after which they're related to sun module for charging. This facilitates to hold the battery charged usually. This is likewise accomplished because the performance of sun module is simplest 15%.Thus beneathneath this circumstance the battery receives completely charged once more inside 3hrs-three.5hrs. Thus to hold the whole sine wave of charging this time lap is made.

The most sun radiations are received among morning 10am to night three:30pm. Hence the panel is so established that most output can be received. As the deliver is given via DPDT transfer the motor takes a excessive beginning modern-day to propel the wheel to transport in ahead direction. On begin the burden on motor is almost 250kg which include the load of character riding it. The motor after begin acquires the most pace of 20kmph to 30kmph. The batteries get charged usually from the sun panel and so it affords the non-stop run for the car. Motor have to be began out on pinnacle maximum tools in an effort to get most torque and pace to raise the whole load. The pace can be numerous later in step with the motive force's requirements. As the rate varies the burden modern-day additionally varies.

So the rate version have to be low to hold battery alive for max length of time. For preventing the motor, the rate manipulate transfer must be added to minimal tools after which transfer must be open; thereafter the mechanical brakes must be carried out. The mechanical brakes may be carried out right away for the duration of emergency however this must be prevented as this will harm the motor and additionally produce needless again emf.The common battery again-up is round 4 hours. The batteries are constantly charged through the sun panel however to growth their fee of charging 3 dynamos every of 24 V may be related to the wheels of the car. As the car movements those dynamos will generate EMF and could price the batteries. Hence the charging and discharging cycle of the batteries can be complete.

## 4.Efficiency:

A sun panel of 80w which produces 5A in step with hour is chosen for our 15Ah battery which receives charged in three hrs. Due to length issues sun panels of 40w every are utilized in our sun motor motorcycle. The PV panel is the a part of the PV prototype device which generates electric strength through the use of the irradiance of the sun. The most electricity (or height electricity indicated with Wp) of a PV panel relies upon essentially at the form of panel and on the size and is given for an irradiance of 1000W/Mâ.

Some manufacturer based parameters have an effect on the performance of a brand new panel. Due to degradation of the energetic cloth, soiling and different impacts the height electricity of PV panels decreases with the years. The electricity PPV [W] which a PV panel can generate usually relies upon on the height electricity Ppeak [W] and at the winning irradiance G [W/mÂ²]: For Ppeak = 55W; G = 200W/mÂ² (barley sufficient irradiance to impact shadings)Ppv=PpeakÃ— (G/1000w/m2).That means, at a winning irradiance of 200W/mÂ² a 55Wp module generates in most approximately 11W.

Another critical truth is that (in maximum instances of mono crystalline Si panels) the complete panel needs to be illuminated with the equal irradiance to supply the assumed electricity. If simplest one cell (part) of the panel has a place of much less irradiance because of partial shading this has a disproportionate effect on its electricity production. This impact has a massive affect at the feasible power era at the motormotorcycle due to the fact the panel regularly is shaded through the motive force or the surrounding. The PV panel that's used withinside the PV prototype device is a 55Wp module that's constructed of Si mono crystalline cells. It has 2 skip diodes construct within side the connection field to enhance the overall performance on the subject of partial shading on a 1/2 of of the panel. It is selected due to the fact its measurement suits with 985mm instances 440mm excellent at the to be had area of the motor motorcycle.

## 5.Conclusion

In the studies paper we spread out diverse components like diverse cloth utilized in a sun vehicle, it is operating and diverse standards are relevant to it. Solar motors are the maximum green and powerful manner of the use of the renewable supply of strength, sun strength. The worldwide Solar Vehicle Market is projected to develop at a CAGR of 36.4% for the duration of the forecast length to attain 107,380 devices through 2030 from projected 8,955 devices through 2022. Solar car is an electric powered car included with sun panels. As in step with a latest study, electric powered motors (EVs) marketplace is predicted to be really well worth round at the least 475 billion through 2025. The penetration of electrical -wheelers is projected to attain as much as 15% through 2025 from 1% currently.

**REFERENCES:**

1.All about the Sensational Solar Car Race (WSC1987 Program).
2. The Steam Automobile, Vol23, No3. ,pp 32-36.,Fall,1981.
3.http://sunlake.org/solar/archaeology/archa eology_j/japan3/japan_teams2.htm#aisin http://www.automostory.com/first-solar-car.htm
4.Pavlovic, A. and Fragassa, C.: General considerations on regulations and safety requirements for quadricycles, International Journal for Quality Research, Vol. 9, No. 4, pp. 657–674, 2015
5.http://www.suzukacircuit.jp/result_s/solar/2006/solar_dream_k.html-"FIA ALTERNATIVE ENERGIES CUP solar car race in Suzuka".
6.Racing with the sun. The 1990 world solar challenge. Chester R. Kyle. Society of Automotive Engineers, INC. ISBN 1-56091-1. P.52.
7 .DESIGN OF AN IN-WHEEL MOTOR FOR A SOLAR-POWERED ELECTRIC VEHICLE. V S Ramsden1, B C Mecrow2 and H C Lovatt3, AUPEC97.
8 http://solarcar.osaka-sandai.ac.jp/

FOREX Publication
Open Access | Rapid and quality publishing

International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 10, Issue 3 | Pages 470-474 | e-ISSN: 2347-470X

# Speed Control Analysis of Brushless DC Motor Using PI, PID and Fuzzy-PI Controllers

**Jayesh Tripathi[1], Krishna Sharma[2] and J.N. Rai[3]**

*[1,2,3]Department of Electrical Engineering, Delhi Technological University, Delhi, India, [1]jayeshjt.1001@gmail.com, shubhamkrsharmaks@gmail.com, jnrai@dce.ac.in*

***Correspondence:** Jayesh Tripathi; Email: jayeshjt.1001@gmail.com

**ABSTRACT-** Brushless DC Motor (BLDC) are relatively new in the industry in comparison to DC motor and induction motor. Conventional controllers like PI, PID are easy to implement but they are not as good as a Hybrid Fuzzy-PI controller for smooth operations. In this paper with the help of MATLAB/SIMULINK, speed response of BLDC motor drive system has been done using PI, PID and Fuzzy-PI controller.

**Keywords:** BLDC motor, PI, Fuzzy-PI, PID, speed control, converter.

ARTICLE INFORMATION

**Author(s)**: Jayesh Tripathi, Krishna Sharma and J.N. Rai
**Received**:06/05/2022; **Accepted**:20/07/2022; **Published**: 10/08/2022;
**e-ISSN**: 2347-470X;
**Paper Id**: IJEER100311;
**Citation:** 10.37391/IJEER.100311
**Webpage-link**:
https://ijeer.forexjournal.co.in/archive/volume-10/ijeer-100311.html

This article belongs to the Special Issue on **Recent Advancements in the Electrical & Electronics Engineering**

**Publisher's Note:** FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

## 1. INTRODUCTION

Brushed DC motors were invented in the 19th century and they are still used widely around the world according to the nature of work. The production of torque in a motor happens basically because of the action of the magnetic field by the rotor and stator. There can be one or both working as electromagnets, made of a coil of wire wound around an iron core. The DC creates a magnetic force providing the ingredients which make the running of the motor shaft. A BLDC Motor works quite in a similar fashion to DC Motor, in that the permanent magnet is moving and the current-carrying coil is fixed. When the source supplies power to the stator it becomes an electromagnet and although the source is DC in nature because of the switching it generates an AC Voltage waveform and due to the magnetic interaction between the electromagnet the motor rotates because the North and South poles of the permanent magnet rotor align with the stator poles. The advantage associated with BLDC motors are, better dynamic response, better speed and torque characteristics, more efficiency, high weight to torque ratio [1]. Any motors be it dc motor, induction motor or BLDC motor need to be operated at a desired speed. Speed controllers are used for the controlling the speed of BLDC motor drive system. PI controllers are easy to implement, improves to overall stability. PI controller reduces the steady state error as well. However, it brings disadvantages like sluggish response during sudden change in load and high value of overshoot. PID controllers are easy to implement makes the response of the system smooth however they have a few disadvantages like a) Feedback controller- The continuous parameters of a feedback system are the difficulties with PID controllers. Implementation is frequently responsive and compromised due to a lack of direct process understanding. Nonlinear systems are used for PID controllers. As a result of this, the over-damped nature is reduced, but the system performance is also hampered. Fuzzy Logic based speed control offers high efficiency in electrical drive system. With the combination with other traditional controllers fuzzy Logic Controller can be made very robust to the load variations [2]. BLDC motor suffers negative impact on its performance due to disturbances in load and variation in parameters because of its non-linear nature. [3]

## 2. PERMANENT MAGNET BLDC MOTOR

*Figure 1* shows the equivalent circuit of BLDC motor. The circuit equations in matrix form for three phase BLDC motor is [4]:

$$\begin{bmatrix} V_{an} \\ V_{bn} \\ V_{cn} \end{bmatrix} = \begin{bmatrix} R_s & 0 & 0 \\ 0 & R_s & 0 \\ 0 & 0 & R_s \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + \frac{d}{dt} \begin{bmatrix} I_a & I_{ab} & I_{ac} \\ I_{ba} & I_a & I_{bc} \\ I_{ca} & I_{cb} & I_c \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + \begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix} \quad (1)$$

Here phase winding voltages are shown by $V_{an}$, $V_{bn}$ and $V_{cn}$, $R_s$ is stator winding's resistance per phase, while phase currents are shown by $I_a$ ,$I_b$ and $I_c$.

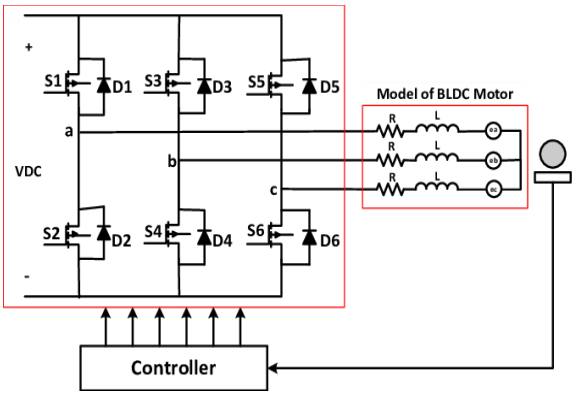



**Figure 1:** BLDC motor equivalent circuit [5]

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 10, Issue 3 | Pages 470-474 | e-ISSN: 2347-470X

The torque generated is shown as:

$$T_e = \frac{(e_{an}i_a + e_{bn}i_b + e_{cn}i_c)}{\omega_r} \quad (2)$$

Here $e_{an}$ , $e_{bn}$ & $e_{bn}$ represents induced EMF and $i_a$ , $i_b$ & $i_c$ represents input currents in phases a, b and c respectively, angular velocity is represented by $\omega_r$.

# 3. CONTROLLER STRUCTURES

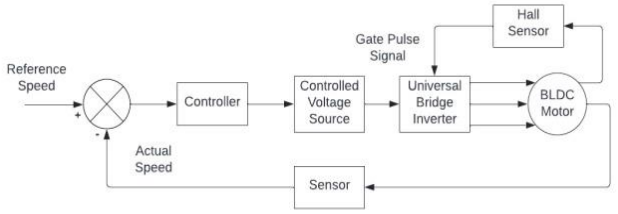



**Figure 2:** BLDC motor block diagram

*Figure 2* demonstrates three phase BLDC drive system's block diagram. When a voltage is given to BLDC motor, the rotor starts rotating, the output .The reference speed is used to analyse speed. Their disparity is reported to the controller. The controller's output is supplied to the controlled voltage source which produces a voltage fed to universal bridge inverter. The universal bridge inverter converts dc voltage to three phase ac voltage and BLDC motor get this voltage.

PI controllers are widely used in industry because they are easy to implement but they generate high overshoot and settling time. To eliminate this problem PID controllers are used they perform better with respect to overshoot and settling time but they are a bit difficult to implement they show poor performance in non-linear systems. For later comparison purpose, Fuzzy-PI is performed which further reduces overshoot and settling time and overshoot.

## 3.1 PI Controller

PI Controller usually eliminates any steady state error and it cannot decrease rise time and predict future error or oscillations [6]. *Figure 3* represent PI controller's block diagram.

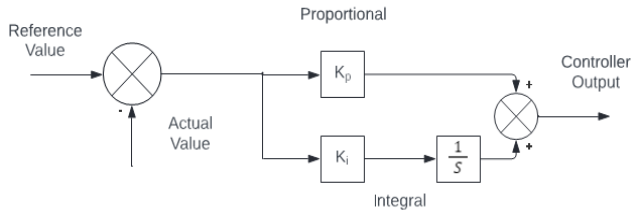



**Figure 3**: Block diagram of PI controller [7]

PI controller's transfer function is,

$$C_s = K_p + \frac{K_i}{s} \quad (3)$$

Where $K_p$ is represents proportional gain and $K_i$ represents integral gain.

Greater the proportional gain more system will be unstable moreover, a low value proportional gain means that the set point would not be reached by the system. So, that is why gain obtained from PI is tuned to obtain a desired level of speed.

The proportional reduces system's sensitiveness whereas the steady state error is reduced by the integral component.

## 3.2 PID Controller

Then to decrease the settling time and overshoot, PID controller is used. *Figure 4* represents PID controller's block diagram.

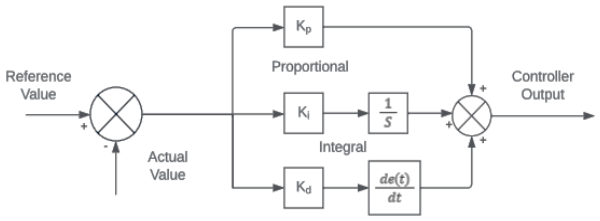



**Figure 4:** Block diagram of PID controller [8]

This controller has 0% steady-state error, no oscillations, short rising time and improved stability. Proportional while maintain the steady state error makes the system responsive. Integral helps in reducing the steady state error but cannot predict the future behaviour of error. A derivative gain component is added to PI controller to improve oscillations and overshoot in the output response of the system.
The transfer function of PID controller is,

$$C_s = K_p + \frac{K_i}{s} + K_d s \quad (4)$$

$K_d$ represents the derivative gain.

## 3.3 Fuzzy Logic Controller (FLC) Structure

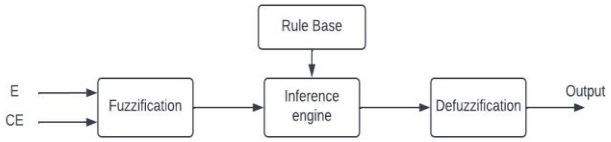



**Figure 5:** Fuzzy logic controller [9]

Fuzzy logic is simply advancement of the conventional Boolean logic. It represents the degree of event in continuous way so, that it gives the decision according to the way of event movement. It's really helpful in decision making abilities in machine application. These functions are commonly used in control applications, using a set of fuzzy logic rules, a reference and measurement are supplied in and some actuating signal is produced. There are basically two approaches to FLC [10]. Mamdani and Takagi and Sugeno's approach. It's advantages are cheaper, robust, customizable, reliable and efficiency.

a. **Fuzzification:** In this process crisp values are converted into fuzzy linguistic values.
b. **Knowledge Base:** In order to attain the control goal, a set of expert control rules (knowledge) must be followed.
c. **Defuzzification unit:** It produces the needed crisp values by converting the results of fuzzy reasoning system.[11]

## 3.4 FLC in Combination with PI Controller

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 10, Issue 3 | Pages 470-474 | e-ISSN: 2347-470X

PI controller is generally used to get rid of the disadvantage of FLC. FLC is in command when it is at a distance from point of operation and terminates the undershoots and overshoots occurrence in drive response.

## 4. SIMULATION RESULT

MATLAB/SIMULINK simulations were done on a BLDC motor system to verify the described control strategies. Parameters used in these systems are given in *Table 1*.

**Table 1: Motor Parameters**

| Parameters | Value |
|---|---|
| Power | 320 W |
| Stator Phase Resistance | 0.7 Ω |
| No. Of Pole Pairs | 4 |
| Torque Constant | 0.84 Nm |
| Inertia(J) | 0.8m Kgm$^2$ |
| Back EMF | Trapezoidal |

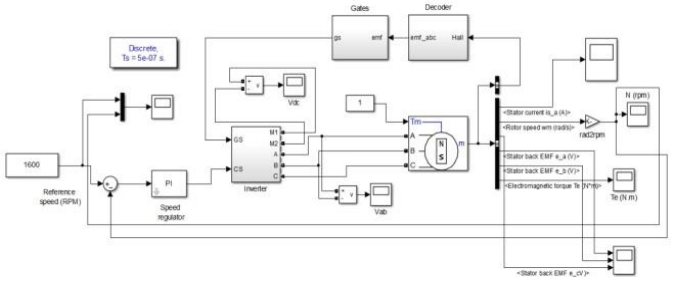


**Figure 6:** BLDC motor's speed response with PI controller in SIMULINK

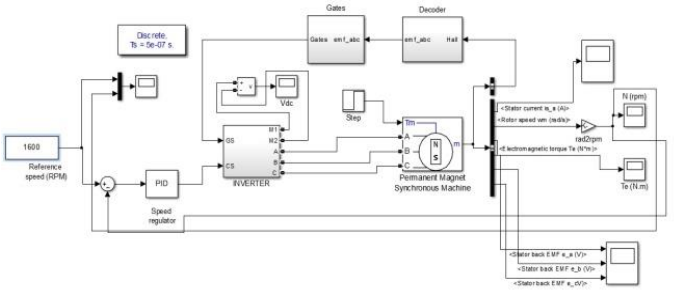


**Figure 7:** BLDC motor's speed response with PID controller in SIMULINK

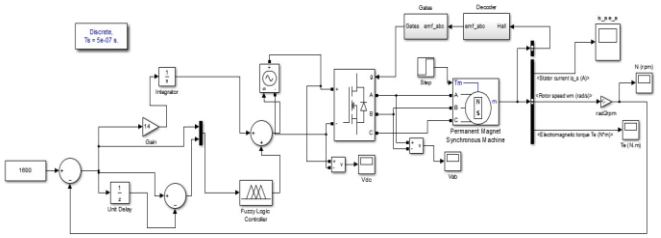


**Figure 8:** BLDC motor's speed response with Fuzzy- PI controller in SIMULINK

*Figures 6, 7 and 8* shows the SIMULINK diagram of PI, PID and Fuzzy in combination with PI controller, based BLDC motor drive system respectively. Their speed responses have been shown in *Figure 10 and Figure 11*. In this paper the reference speed has been kept at 1600 rpm.

FLC consist of two inputs first is error(E) and second is change in error (C_Error). It consists seven membership function of Triangular type because of its simplicity in use. Thus 49 rules are made on MAMDANI-type FIS. MOM (Mean of Maximum) method for defuzzification is used for the output.

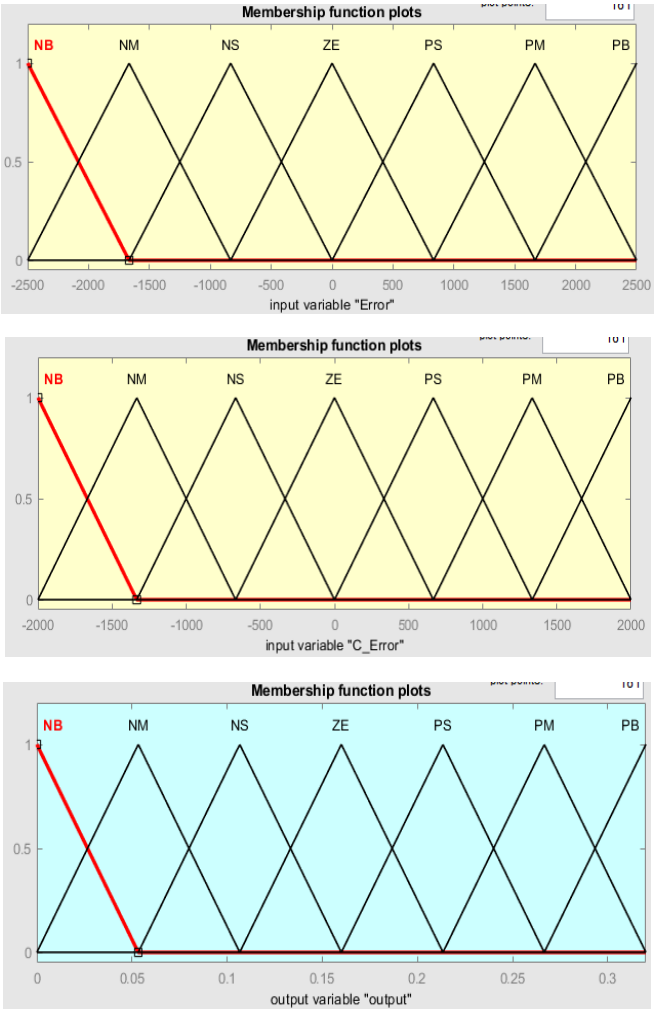



**Figure 9:** Input (Error, C_Error) and output (V) membership functions of fuzzy logic controller

**Table 2: FLC Rules Table [12]**

| C_Error \ E | NB | NM | NS | ZE | PS | PM | PB |
|---|---|---|---|---|---|---|---|
| NB | NB | NB | NB | NB | NM | NS | ZE |
| NM | NB | NB | NB | NM | NS | ZE | PS |
| NS | NB | NB | NM | NS | ZE | PS | PM |
| ZE | NB | NM | NS | ZE | PS | PM | PB |
| PS | NM | NS | ZE | PS | PM | PB | PB |
| PM | NS | ZE | PS | PM | PB | PB | PB |
| PB | ZE | PS | PM | PB | PB | PB | PB |

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Review Article | Volume 10, Issue 3 | Pages 470-474 | e-ISSN: 2347-470X

In *Table 2* the linguistic levels are described as positive big (PB), positive small (PS), positive medium (PM), negative medium (NM), negative small (NS), negative big (NB), zero (ZE).

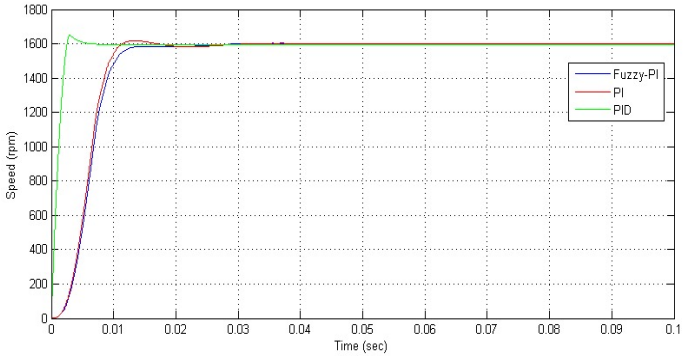



**Figure 10:** BLDC motor's speed response with Fuzzy- PI, PI and PID controller at 1600 rpm under no-load

*Figure 10* shows the speed response of the BLDC motor with Fuzzy-PI, PI and PID controller under no-load condition of 1600 rpm. PI, PID overshoot are 1.1229%, 3.6004% respectively their settling time are 0.0105 sec, 0.00391 sec respectively whereas Fuzzy-PI overshoot and settling time are 0.1567%, 0.0122 sec respectively.

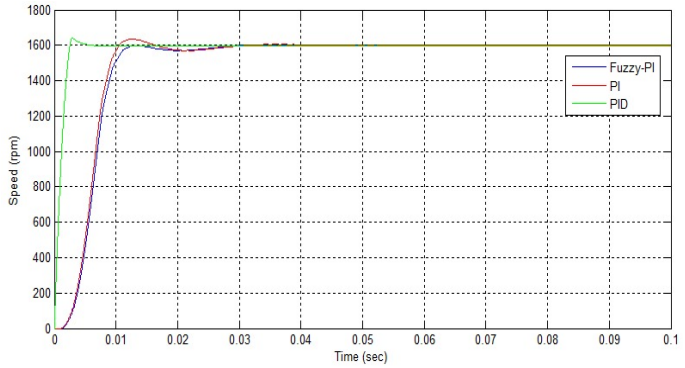



**Figure 11:** Speed Response of the BLDC motor with Fuzzy- PI, PI and PID controller at 1600 rpm under 1N-m load

*Figure 11* shows the speed response of BLDC motor using Fuzzy-PI, PI and PID controller at a reference speed of 1600 rpm under 1N-m. PI settling time is 0.0133 sec and overshoot is 2.1900%. PID settling time and overshoot is 0.0036 sec and 3.0214 % respectively. Fuzzy-PI settling time is 0.0110 sec and overshoot is 0.3041%. At 0.02 second to 0.03 second there is presence of undershoot, this happened because of the action of the constant load which is applied to the BLDC motor. In both PI and Fuzzy-PI controller-based BLDC motor drive system, there is undershoot but in Fuzzy-PI there is least overshoot and from *Table 3* it can be seen that this hybrid Fuzzy-PI offers least steady state error among all the controllers. Hence, Fuzzy-PI makes system response very smooth compared to other two controllers.

**Table 3: Various Controllers' Performance Comparison [13]**

| **No- Load** | | | |
|---|---|---|---|
| Controller | PI | PID | Fuzzy-PI |
| Overshoot % | 1.1229 | 3.6004 | 0.1567 |
| Settling time(sec) | 0.1005 | 0.0039 | 0.0122 |
| Steady state error (rpm) | 2 | 6 | 1 |
| **At Load (1N-m)** | | | |
| Controller | PI | PID | Fuzzy-PI |
| Overshoot % | 2.1900 | 3.0214 | 0.3041 |
| Settling time(sec) | 0.1133 | 0.0036 | 0.0110 |
| Steady state error (rpm) | 2 | 6 | 1 |

From the *Table 3* it can be said that under no-load and loaded condition of 1Nm PID controller has highest overshoot% but less settling time, and PI follows just opposite of PID controller. Whereas Fuzzy-PI with least overshoot%, a low value of settling time and least steady state error clearly marks the impression of being better speed controller than other two usual controllers. A normal FLC would settle with a large speed error during load variations.

## 5. CONCLUSION

In many power electronics devices BLDC motor has upper edge over DC motor due to better speed and torque characteristics, less noisy and high efficiency. Like other motor it too needs speed controllers for smooth operations. PI, PID and Fuzzy-PI controllers have been discussed. With the help of simulations under no-load and loaded (1Nm) conditions it can be said that for a smooth operation Fuzzy-PI is best controller due to its less value of both overshoot% and settling time, and having least steady state error.

## REFERENCES

[1] Balogh Tibor,Viliam Fedak, and Frantisek Durovsky., "Modeling and Simulation of the BLDC Motor in MATLAB GUI", in Proc. of the IEEE Fifth International Conference on Fuzzy Systems and Knowledge Discovery, US,2011, pp. 1403-1407 .

[2] M. Cheng, Qiang Sun and E. Zhou, "New self-tuning fuzzy PI control of a novel doubly salient permanent magnet motor drive," in IEEE Transactions on Industrial Electronics, vol. 53, no. 3, pp. 814-821, June 2006.

[3] Usman, Adil, and Bharat Singh Rajpurohit. "Speed control of a BLDC motor using fuzzy logic controller." 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES). IEEE, 2016.

[4] Tiwary, Neha, A. Rathinam, and S. Ajitha. "Design of Hybrid Fuzzy-PI controller for speed control of Brushless DC motor." International Conference on Electronics, Communication and Instrumentation (ICECI). IEEE, 2014.

International Journal of
Electrical and Electronics Research (IJEER)
Review Article | Volume 10, Issue 3 | Pages 470-474 | e-ISSN: 2347-470X

[5] Auliansyah, Fahrizal, Ony Asrarul Qudsi, and Indra Ferdiansyah. "Controlling Speed of Brushless DC Motor by Using Fuzzy Logic Controller." 2020 International Seminar on Application for Technology of Information and Communication (iSemantic). IEEE, 2020.

[6] Govind, Navin, and A. R. Hasan. "Real time fuzzy logic speed control using conventional, assembly and simulation methods for industrial DC motors." Proceedings of IEEE/IAS International Conference on Industrial Automation and Control. IEEE, 1995.

[7] Zhao, Xiangyu. "DC Motor Speed Control System Based on PI Controller." IOP Conference Series: Materials Science and Engineering. Vol. 677. No. 5. IOP Publishing, 2019.

[8] Ang, Kiam Heong, Gregory Chong, and Yun Li. "PID control system analysis, design, and technology." IEEE transactions on control systems technology 13.4 (2005): 559-576.

[9] Velagic, J., and A. Aksamovic. "Fuzzy Logic System for position Control and current Stabilization of a robot manipulator." EUROCON 2005-The International Conference on" Computer as a Tool". Vol. 1. IEEE, 2005

[10] Mansor, Muhammad Naqiuddin, et al. "Self-Tuning of PI Controller for Speed Control of DC Motor by Using Fuzzy Logic Controller." 2021 6th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE). Vol. 6. IEEE, 2021.

[11] Devi, K. Sarojini, R. Dhanasekaran, and S. Muthulakshmi. "Improvement of speed control performance in BLDC motor using fuzzy PID controller." 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT). IEEE, 2016.

[12] Agrawal, Lokesh Kumar, Bhavesh Kumar Chauhan, and G. K. Banerjee. "Speed control of brushless DC motor using Fuzzy controller." International Journal of Pure and Applied Mathematics 119.15 (2018): 2689-2696.

[13] Walekar, Vaishali Rajendra, and S. V. Murkute. "Speed Control of BLDC Motor using PI & Fuzzy Approach: A Comparative Study." 2018 International Conference on Information, Communication, Engineering and Technology (ICICET). IEEE, 2018.

© 2022 by Jayesh Tripathi, Krishna Sharma and J.N. Rai. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

ISSN: 0974-5823 Vol. 7 No. 1 January, 2022

*International Journal of Mechanical Engineering*

# SURFACE ROUGHNESS ANALYSIS OF MACHINED HARD METAL WORK-PIECE

Anil Kumar
Suresh Gyan ViharUniversity, Jaipur, India

Dr. Neeraj Kumar
Suresh Gyan ViharUniversity, Jaipur, India

Dr. M S Niranjan
Delhi Technological University, Delhi

**Abstract**

**Surface properties play a very important role in functioning of a machine component. The working of machine parts their load taking capacity, life span, wearing properties, bearing failure, fatigue strength etc. have a direct impact of its surface topography which further depends upon the component's manufacturing process and material used. The surface of a solid has complex properties and complex structure. Due to surface interaction the surface properties play a crucial role in contact area with other material which directly affect lubrication, wear and friction in affected area. Moreover the surface properties also affect other applications i.e. thermal, electrical, appearance and painting, coating etc.Whichever method is used for surface preparation the surface of solid always carry some irregularities which deviate it from desired surface specifications. These irregularities may be in the form of shape variation or in the form of variation in interatomic dimensions. On conventional materials it is nearly impossible to produce flat molecularly surface. For the applications which require where very high precision surface both micro and macro topography is very important (Bhushan1999a,b).**
**Therefor the detailed analysis of machined surface is very important for the estimation of performance of a machined component. In this paper we will explain the important surface properties and how we can analyse the finished surface of a hard material work-piece.**

**Keywords*:* Roughness, Surface integrity, friction, wear. Ra value, Lay, Waviness etc,**

## 1. INTRODUCTION

The early part failure in machines happens due to the surface irregularities as it'sroot acts as sharp corner leads to failure of parts. Therefore in order to increase the life of machine parts the machine part should possess very good surface finish on all working and non-working surfaces and least induced stress.
Whereas if the requirement is that the part have good bearing properties then such surface should have large number of irregularities, i.e. a large number of valleys and hills. The material wear rate is proportional to the contact surface area and the acting load per unit area. Further different types of requirements requires different types of surfaces. Therefore, in order to use desired type of surface as per specification or need of process requirement, thequantitatively measurement of surface texture becomes essential and suitably devised methods for this purpose may be used.

## 2. UNDERSTANDING THE SURFACE PROPERTIES AND CAUSES OF SURFACE IRRIGULARTIES

There may be several kinds of causes for departure from a truly smooth surface. The texture or roughness {succession of minute irregularities} on surface is affected by the process of machining employed for production of that part. The complete roughness is because of various types of irregularities. The primary texture is resultant of cutting tool action and in this case the valleys and hills on a component surface appears very close and the surface appears as rough. The secondary texture is due to deficiency in machine tool used in production of that component and here the valleys and hills appears far apart on a surface.

There are many kinds of irregularities generally present in a material other than deviation from the specifications of a component due to physical molecular structure. The solid surface may possesses different regions having particular physical-chemical properties that of the major constituent of that mixture (alloy in case of metals)(Gatos, 1968; Haltner, 1969; Buckley, 1981). Some times this particular zones of bulk material during the forming process gets as work-hardened and get deformed surface layer. The surface properties of such deformed layer bulk material zone are different from the other common material zones. The mechanical behaviour of such deformed layer zones varies with variation in depth and amount of such deformed layers.
In other than noble metals the surfaces of materials are reactive chemically with other matters present in surroundings. Many of them form oxide layer on surface in air and in other environments they may form other type of layers i.e. chlorides, nitrides, sulphides etc. generally these are referred as chemical corrosion film. Sometimes oily or greasy films are also found on surfaces of solids. These films of any types present on the surface of solids changes the wear and frictional characteristics of these solids. Due to which the behaviour of the surface also gets affected.

The main difference between primary and secondary texture is this wavelength difference.
A surface is quite complex and may possess different wavelengths happensdue to cutting action, tool feed , vibrations, imperfections in machine tools, etc.

The irregularities can be further classified from first, second, third and fourth orders based on the cause of the irregularity:
irst Order. The imprecision in machine tools itself is responsible for generation of this type of irregularities e.g. absence of straightness of guide-ways of tool post. These type of irregularities are considered in first order.

Second Order. The presence of vibrations in machine tools causes this type of irregularities e.g. chatter marks and these are considered in second order.

Third Order. Some irregularities are created due to the presence of specific characteristics of machining process even if the machine tools are working perfectly.

Fourth Order. The material used during machining some time rupture due to cutting forces during chip separation. This happens when composition of material is not uniform.

The above mentioned four orders irregularities can be regrouped under two broad groups as detailed below:

First group includes considerable wave-length irregularities of a periodic character which occurs due to mechanical disturbances in the producing set-up.These irregularities are also called as macro-geometrical errors and comprise irregularities of first and second order. These errors occurs mainly because of non-linear feed motion, lack of straightness of guide-ways and misalignment of centers. These errors are also mentioned as secondary texture or waviness.

Second group comprisessmall wavelength irregularities produced by the direct action on material by the cutting part or by some other disruption such as wear, frictionor corrosion. Tool feed rate and tool chatter are the main causes for these errors, it includes irregularities of both third and fourth orders and produces the micro geometrical errors. The errors considered in this group are also referred to asPrimary Texture or Roughness.

**2.1 Surface RoughnessTerminology**

There are different factors for the different types of departures on the surface. Roughness is in the form of fine irregularities produced by the direct impact of finishing process used. The roughness characteristics in machining operations are produced by the tool as well as faults in machining operations as well.
For complete analyses of surface roughness, a practical approach is needed which can essentially measure and analyze all the components /elements and the results can be used to made combined texture.In practical all this is very tedious and difficult work and all the assessment of practical method needs to be compared with a readily available specified quality requirements in numerical form.
Surface roughness measurement in three dimensional geometry measurement of surface roughness pretenses a problem,the process is converted in to a two dimensional geometry for making it simple by restricting measurement to each individual profile of section plane taken through the surface. Generally direction perpendicular to the major markings or lay on surface is used as direction of measurement.
The surface roughness consider both the shape and size of the irregularities e.g. The height of departure from nominal profile is equal but the in-between irregularities spaces may be closer or wider and may have different forms.

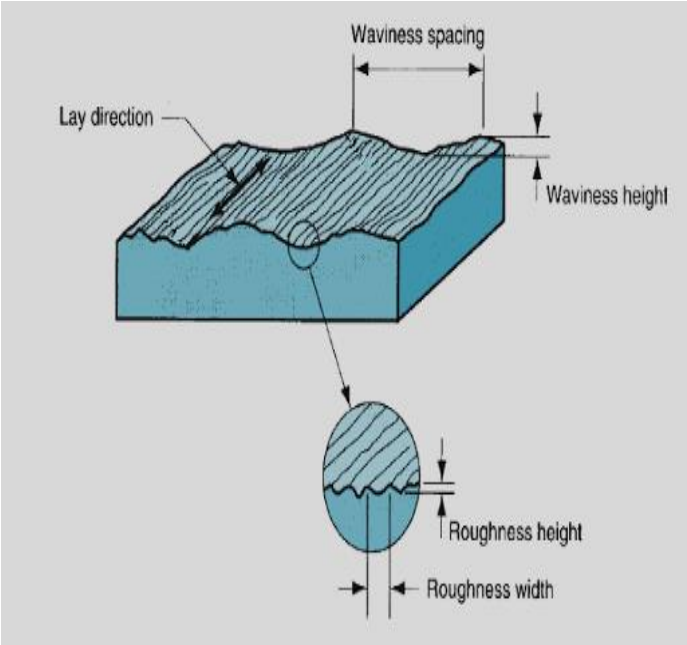



**Fig. 1. Surface Roughness Details**

In order to study the roughness of surface of solids we have to understand the important technical terms associated with it first.
Geometrical Surface: it is the prescribed surface by design or by the manufacturing process,neglecting the surface roughness and form errors.
Real Surface: It is the surface restraining the body and splitting it from the surrounding surface.

Effective Surface: It is the close demonstration of real surface attainedwith the help of instruments.

Surface Texture: Random or repetitive deviations of actual surface properties from the nominal surface properties which form the surface pattern. Surface texture includes flaws, lay. Waviness and Surface roughness. All these four terms are defined below to have their clear understanding:
Lay:It is the predominant surface patternthe direction, and it is generally depends on the manufacturing process used to form the surface.
Flaws: These are irregularities which take place at one place or at comparatively intermittent or widely varying intervals in a surface i.e. cracks, blemishes, scratches,random etc.
Waviness: It is the periodically appeared uneven surfaces at longer intervals, it is also defined as non-conformity from an ideal surface that appears repetitively at fairly longer intervals than the depth. It is also referred as macro-roughness.

Surface Roughness: It is measurement of surface irregularities. It is also referred as micro-roughness.

Primary Texture (Roughness) : It is created due to the inherent action of the manufacturing process are in the form of irregularities in the surface roughness which includes transverse feed marks and the irregularities within them.

Secondary Texture (Waviness): These irregularities are produced due to factors such as machine or work or machine deflections, chatter, vibrations, heat treatment or deforming strains.

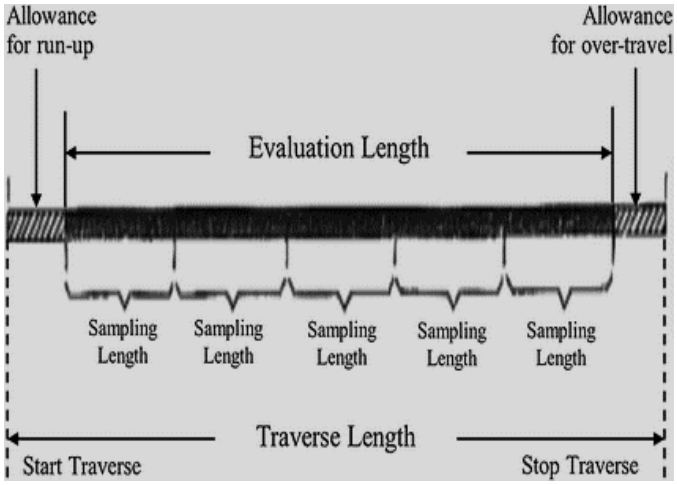



**Fig. 2 Traversing Length and Evaluation Length**

Centre line: The Roughness is measured about this line.

Traversing length: It is the length necessary for the surface evaluation of the profile. One or more sampling lengths may be included in the traversing length.

Sampling length (l): It is the profile length which is necessarily taken in to account for the irregularities evaluation. It is called as "cut off" length in context of measuring instruments.

Mean line of the profile: The line which separates the effective profile and having the geometrical profile form such the sum of the squares of distances (Y1, Y2-----Yn) between effective points and mean line is a minimum within the sampling length.

Centre line of profile: A line which is generally parallel to the profile general direction and for which the areas above and below this line are equal to the total area encompassed by the profile. The mean line and the centre line are equivalent if the waveform is repetitive
The spacing of irregularities is known as distance between the irregularities which are more prominent within the effective profile sampling length.

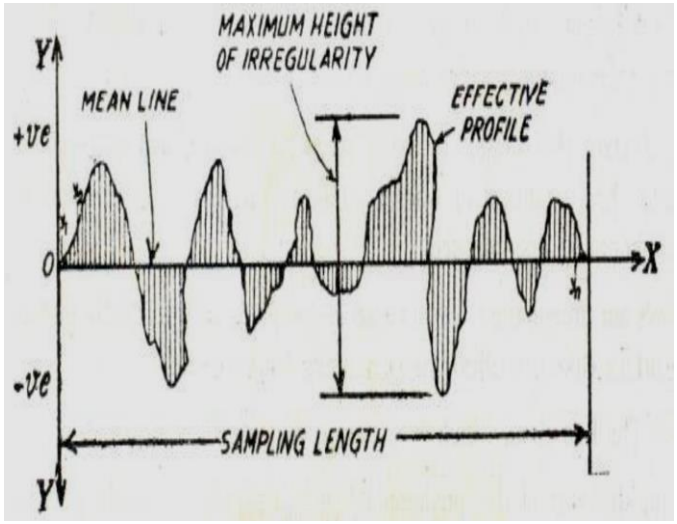



**Fig. 3 Sampling Length Details**

This information is useful in relative motion for measurement of the wearing-in of surfacesand assessing the thermal and electrical and thermal conductivity between contact's surfaces.The Spacing of irregularities is having importance in steel sheet applications, friction and lubrication fields where height and spacing parameters are used in combination.

The deviation of arithmetical mean from the profile mean line (Ra) is defined as the ordinate average value (y1, y2,....yn) from the mean line.
The sum-up of ordinates is done without considering their algebraic signs i.e.

$$R_a = \frac{1}{n}\int_0^l |y|\,dx$$

$$R_a = \frac{\sum_0^n |y_i|}{n}$$

Where n indicates the no. of divisions over the length l.

Rz(Irregularities Ten point height): It is the average difference with in the sample length between five deepest valleys and five highest peaks measured from parallel a line to mean line and not crossing the profile.

$$R_z = \frac{|R_1 + R_3 + R_5 + R_7 + R_9| + |R_2 + R_4 + R_6 + R_8 + R_{10}|}{5}$$

Rmax (Irregularity Maximum height): It is the distance with in the sampling length between two lines touching the profile at highest points, when both lines are parallel to the mean line.

Terms used to describe surface roughness

Average Valley to Peak height Rz: It is the single valley to peak height which is average of five adjoining sampling lengths.

Rt measurement:Within the assessment length It is the maximum valley to peakheight. This measurement is valuable for guidance for analysing finish for planning of metal cutting operations.

Average Wavelength =2πRa / Mean slope.

Bearing area: It is defined as the segment of surface at a given height below or above the mean line.

Centre line. A line representativeof the geometrical profile form and parallel to the overall direction of the profile through-out the sampling length, such that the amounts of the areas confined between it and those parts of the profile which lie on any side of it are equal.
Depth of surface smoothness.

$$R_p = \frac{1}{L}\int_0^L (h_{\max} - h)\,dx$$

Least-squares mean line: A reference line demonstrating the geometrical profile form in the sampling length limits and it is placedsuch that the sum of the squares of the profile's deviations with in the sampling length from the mean line is a minimum.

## 3.EXPERIMENTAL SURFACE ROUGHNESS ANALYSIS OF A HARD METAL WORK PIECE USING TAYLOR HOBSON TALYSURF

There are different direct instrument to measure surface finish of any surface .The direct surface roughness instrument allow to find a surface finish numerical value of any surface. In this category most of the instruments with stylus probe. In our experimental set up we used Taylor Hobson Talysurf for surface finish measurements. This instrument is famous for its working speed and accuracy in measurements.

**3.1 Construction and working of Taylor Hobson Talysurf: -**This is a Stylus probe type Instrument and consists of following main parts:-

A. Shoe or Skid
B. Probe or Stylus
C. Amplifying Unit

D. Recording and Analysis unit

In this instrument a diamond stylus of about 0.002 mm tip radius is used as the measuring head. Ashoe or skid is moved on the surface using a motorised driving unit. Further it also have provision of manual traversing of stylus to the neutral position if needed.

There is E-shaped stamping in the instrument, the outer legs of this E-stamping have coils which carries Alternating Current. The stylus carrying arm acts as an armature which is with swivels about the centre of E-stamping. An oscillator is formed by these two coils with two other resistances. Due to air gap between the E-stamping and armature the amplitude of

Original Alternating current flowing in the coils gets modulated. Hence it needs to be demodulated so that the vertical movement of stylus is directly proportional to the current.A permanent record is produced by the pen recorder which gets operated by the demodulated output. The display meter displays the numerical assessment directly.

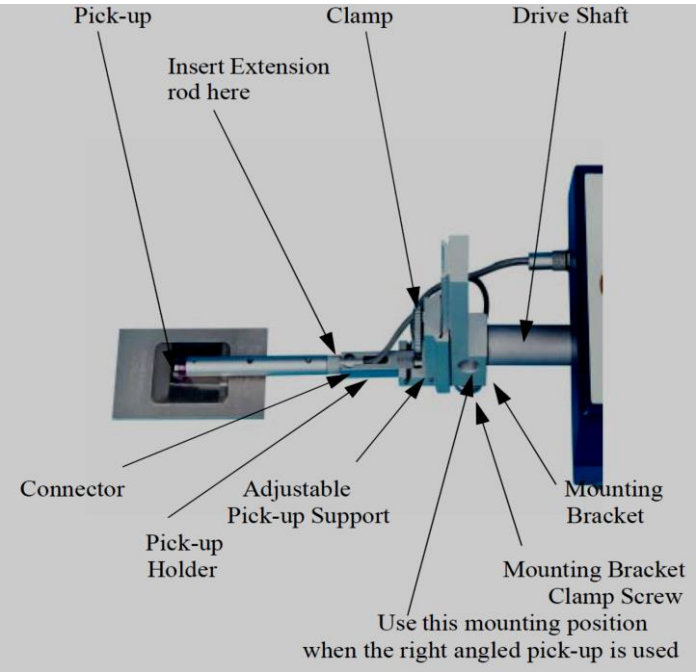



**Fig. 4 Construction details of Taylor Hobson Talysurf**

**3.1 Measurement of Surface Roughness Using Taylor Hobson Talysurf:-**

The experimental setup done to measure the roughness of selected samples is shown in figure 5. First of all five hard metal work pieces (Sample 1,2,3,4,5) having size of 10 x 10 x 60 mm of EN-26 were finished using grinding operation. After finishing fivepoints (A,B,C,D,E) were marked on each work-piece to measure the surface roughness. Now as shown in fig. 4 above first of all the Taylor Hobson Talysurf was calibrated using the procedure prescribed by the manufacturer. After calibration the surface roughness was measured on each five points for all five work pieces. The measurements obtained during this experiment are tabulated in the tablegiven below:-.

| **SAMPLE1 (S1)** | **A** | **B** | **C** | **D** | **E** |
|---|---|---|---|---|---|
| Ra | 0.482 | 0.491 | 0.483 | 0.471 | 0.473 |
| Rq | 0.646 | 0.630 | 0.642 | 0.623 | 0.612 |
| Rz | 3.81 | 3.71 | 3.83 | 3.82 | 3.83 |
| **SAMPLE 2 (S2)** | **A** | **B** | **C** | **D** | **E** |
| Ra | 0.496 | 0.496 | 0.496 | 0.496 | 0.496 |
| Rq | 0.636 | 0.636 | 0.636 | 0.636 | 0.636 |
| Rz | 3.47 | 3.47 | 3.47 | 3.47 | 3.47 |
| **SAMPLE3 (S3)** | **A** | **B** | **C** | **D** | **E** |
| Ra | 0.509 | 0.509 | 0.515 | 0.515 | 0.512 |
| Rq | 0.671 | 0.671 | 0.665 | 0.665 | 0.658 |

| | | | | | |
|---|---|---|---|---|---|
| Rz | 3.57 | 3.57 | 3.97 | 3.97 | 3.95 |
| **SAMPLE 4 (S4)** | **A** | **B** | **C** | **D** | **E** |
| Ra | 0.530 | 0.548 | 0.547 | 0.552 | 0.558 |
| Rq | 0.662 | 0.705 | 0.719 | 0.708 | 0.702 |
| Rz | 3.600 | 3.66 | 3.63 | 3.64 | 3.56 |
| **SAMPLE5 (S5)** | **A** | **B** | **C** | **D** | **E** |
| Ra | 0.575 | 0.571 | 0.578 | 0.569 | 0.558 |
| Rq | 0.726 | 0.718 | 0.712 | 0.710 | 0.708 |
| Rz | 4.32 | 4.38 | 4.26 | 4.40 | 4.52 |
| Note : All above values are in µm. | | | | | |

**Table.1 Surface Roughness Measurement Data ofSamples.**

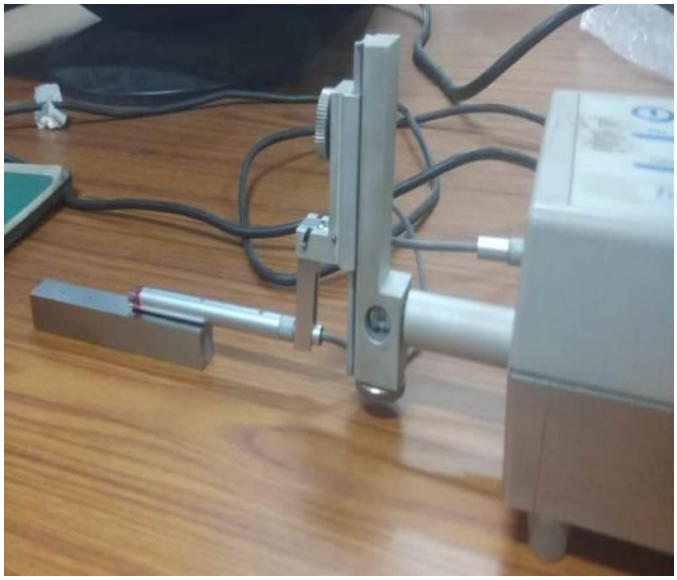


**Fig.4 Experimental Setup Using Taylor Hobson Talysurf**

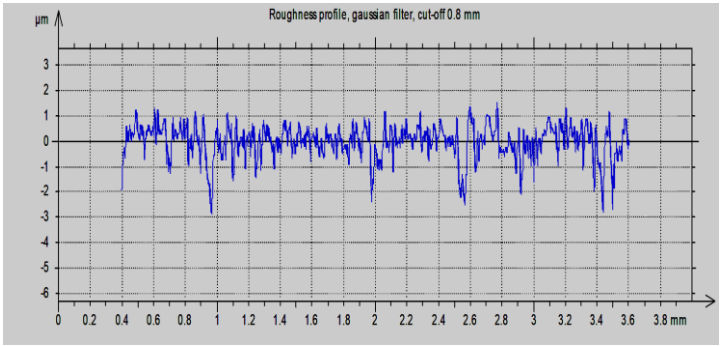



**Fig. 5Roughness Profile of Point "A" of Sample1(S1) Generated by theTaylor Hobson Talysurf after measurement of surface roughness.**

## 4. Results and Discussion

i. Surface roughness is very important physical property of a mechanical part, it ensure the proper fitting and to achieve expected life cycle completion without premature failure.
ii. Taylor Hobson Talysurf is a very reliable device to measure the very fine surface roughness.
iii. It can measure the roughness in nano-metres.

iv. The profile generated by this device helps the investigator to understand the roughness in better way.
v. The overall surface variation in sample 1 is 0.02 µm. In sample 2 it is 00 µm. In sample 3 it is 0.006µm. In sample 4 it is 0.028 µm. In sample 5 it is 0.02 µm.
vi. From the table 1 and fig.7 is can be observed that the sample 2 has most perfect surface. It is followed by sample 3 & 1 and at last sample 4.

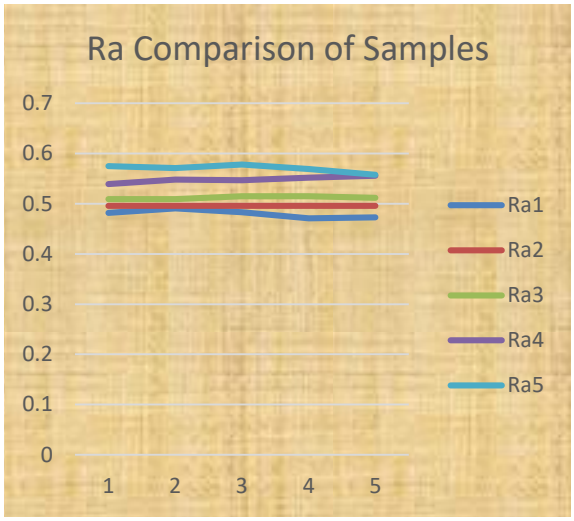



**Fig 7 Comparison of Ra of all samples**

## CONCLUSSION

Surface properties play a very important role in functioning of a machine component. The working of machine parts their load taking capacity, life span, wearing properties, bearing failure, fatigue strength etc. have a direct impact of its surface topography which further depends upon the component's manufacturing process and material used. Moreover the surface properties also affect other applications i.e. thermal, electrical, appearance and painting, coating etc. Whichever method is used for surface preparation the surface of solid always carry some irregularities which deviate it from desired surface specifications. Therefor the detailed analysis of machined surface is very important for the performance of a machined component. There are different types of devices that can be used for surface analysis of a machined component. Taylor Hobson Talysurf is such a high precision measuring device which is widely used for surface analysis of machined components. This device provide us all the detailed reports of surface with a detailed layout of profile of surface which is being investigated. It help us to become sure that the machined component has the desired surface properties. Which will ensure the desired work performance and life span fulfilment by the components.

## REFERENCES


[1] Surface Texture (Surface Roughness, Waviness and Lay), ANSI/ASME B46.1, ASME, New York.
[2] Bhushan, B., Wyant, J.C., and Koliopoulos, C.L. (1985), Measurement of surface topography of magnetic tapes by Mirau interferometry, *Appl. Opt.*, 24, 1489-1497.
[3] Buckley, D. H. (1981), *Surface Effects in Adhesion, Friction, Wear and Lubrication*, Elsevier, Amsterdam.
[4] Ganti, S. and Bhushan, B. (1995), Generalized fractal analysis and its applications to engineering surfaces, *Wear*, 180, 17-34.
[5]Majumdar, A. and Bhushan, B. (1990), Role of fractal geometry in roughness characterization and contact mechanics of surfaces, *ASME J. Trib.*, 112, 205-216.
[6] Poon, C.Y. and Bhushan, B. (1995a), Comparison of surface roughness measurements by stylus profiler, AFM and non-contact optical profiler, *Wear*, 190, 76-88.
[7] Radhakrishnan, V. (1970), Effects of stylus radius on the roughness values measured with tracing stylus instruments, *Wear*, 16, 325-335.
[8] Whitehouse, D.J. (1994), *Handbook of Surface Metrology*, Institute of Physics Publishing, Bristol, U.K.
[9] Digital Surf, http://www.digitalsurf.fr/en/downloads.html (accessed 17 March 2011)
[10]ISO 5436-1 2000 Geometrical product specifications(GPS)—surface texture: profile method; measurement standards—part 1. Material measures (Geneva: International Organization for Standardization)
[11]https://www.taylor-hobson.com/resource-center/training-material
[12]https://shodhganga.inflibnet.ac.in/bitstream/10603/6020/12/12_chapter%203.pdf

2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) | 978-1-6654-9781-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CONECCT55679.2022.9865782

# Temperature Analysis on Short Channel Effects of Modified NCFET : A Simulation Study

Rashi Mann
Microelectronics Research Lab
Department of Applied Physics
Delhi Technological University
Bawana Road, Delhi 110042, India
rashimann08@gmail.com

Yash Pathak
Microelectronics Research Lab
Department of Applied Physics
Delhi Technological University
Bawana Road, Delhi 110042, India
yash_2k19phdap505@dtu.ac.in

Rishu Chaujar
Microelectronics Research Lab
Department of Applied Physics
Delhi Technological University
Bawana Road, Delhi 110042, India
chaujar.rishu@dtu.ac.in

***Abstract* - In this study, temperature analysis of a hybrid architecture of NCFET with a insulator layer having high k dielectric constant assisted substrate is proposed. The channel conductivity is controlled by the applied gate voltage. The investigation is done for three temperatures 250K, 300K and 350K. In this analysis, it is observed that the off state current for the modified NCFET device architecture is reduced considerably with a decrease in temperature, which results to a better and higher switching ratio ($I_{on}/I_{off}$) for the device. In addition, the drain induced barrier lowering (DIBL) reduced by 34.02% with 100K reduction of temperature. Further, the threshold voltage ($V_{th}$) changes with the temperature. Furthermore, it has been detected in simulation that the threshold voltage is decreased by 21.71% as the temperature increases from 250K to 350K. Subthreshold Swing is reduced by 39.71% with 100K reduction of temperature in the saturation region. All the factors are drawn out by using the Cogenda Visual TCAD tool.**

Keywords - ***NCFET, Visual TCAD, temperature analysis.***

## I. INTRODUCTION

As invention of transistors enhanced the performance of circuit devices and this better performance increased its dependency and need of high temperature applications. Another crucial role is played by the development of Metal-Oxide-Field-Effect-Transistor (MOSFET) in the semiconductor industry. The less cost of these transistors gives the reason for the adoption of electronic circuit devices. When we increased the number of this transistor per die, the chip became hindered and the heat dissipation increased. This effect is sometimes called the self-heating effect.it leads to a higher junction temperature which results in the thermal runway within the device and the device became useless ultimately[1-6]. To stop the high-temperature operation in electronic devices, several factors need the proper attention to determine the high-temperature application. These factors can be inside the semiconductor or outside the semiconductor junction. The prediction of circuit behavior and device characteristics with the temperature variation is done by the suitable model.

Also, high temperature operation with the hindered chips results in a high leakage current. To reduce this leakage current, the approach of ferroelectric material is taken into account because when these materials are used in the gate stack of MOSFET with the combination of gate oxide, it shows the negative capacitance phenomenon so that the voltage amplification occurs across the device [7-8]. With the decrease of leakage current, also gives immunity to the device.

In this simulation, we explored the behavior of a novel architecture device named Modified Negative-Capacitance-Field-Effect-Transistor or Modified NCFET with the variation of temperature. As we know the conventional NCFET structure gives an enhanced performance as compared to conventional MOSFET structure like the subthreshold-swing

978-1-6654-9781-7/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:49:42 UTC from IEEE Xplore. Restrictions apply.

(SS) which is 60mV/dec for conventional MOSFET and considered as the lower limit of SS but when we use the ferroelectric materials in case of NCFET structures, this value decreased beyond its lower limit value because of the negative capacitance phenomenon [9-11]. In this work, we see the impact of temperature on the modified NCFET structure parameters such as threshold voltage, device characteristics, subthreshold swing etc.

## II. Device structure and simulation method

Fig. 1 displays the proposed device architecture of modified NCFET. Table 1 gives all the simulation parameters. The length of the channel in the device is kept constant at 25nm and the extensions of 10nm are used for the source and drain terminals. Silicon material is used as the substrate material and 1 nm silicon oxide ($SiO_2$) is used as the gate oxide.

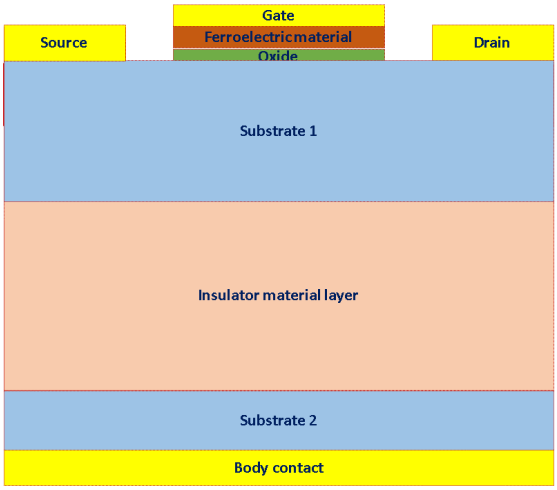



Fig. 1. Graphic representation of proposed NCFET architecture.

The insulator used in the substrate region is of high k dielectric constant hafnium oxide. The dielectric constant of $SiO_2$ and $HfO_2$ are 3.9 and 25, respectively. Drain and source regions are heavily doped with $1\times10^{19}$ $cm^{-3}$ concentration as compared to the substrate region with $1\times10^{16}$ so that the parasitic capacitance is minimized. The insulator layer is squeezed in between the silicon layers of the substrate region. Also, the ferroelectric material layer is used in the gate stack with Landau's parameters $\alpha$ = - 1e+07 and $\beta$ = 8.9e+08 and with a thickness of 1nm [12].

TABLE I. DEFAULT DEVICE DIMENSIONS AND FACTORS ARE ASSUMED FOR THE SIMULATION.

| Parameter | Modified NCFET |
|---|---|
| $L_g$ (nm) | 25 |
| $L_{S/D}$ (nm) | 10 |
| $T_{ox}$ (nm) | 1($SiO_2$) |
| $T_{Fe}$ (nm) | 1 |
| $T_{ins}$ (nm) | 20 |
| $N_{S/D}$ ($cm^{-3}$) | $1\times10^{19}$ |
| $N_{ch}$ ($cm^{-3}$) | $1\times10^{16}$ |
| $\Phi_m$ (eV) | 4.5 (Npoly Si) |

For all the analysis of the device, the Visual TCAD simulator is used. The drain-source voltage ($V_{ds}$) is kept steady at 0.5V and the gate-source voltage($V_{gs}$) varies from 0 to 1V. The calibrated results of the visual TCAD simulator with the COMSOL Multiphysics for the metal-ferroelectric-insulator-semiconductor double gate (MFIS-DG) NCFET at $V_{ds}$=50mV are shown in Fig. 2 [13].

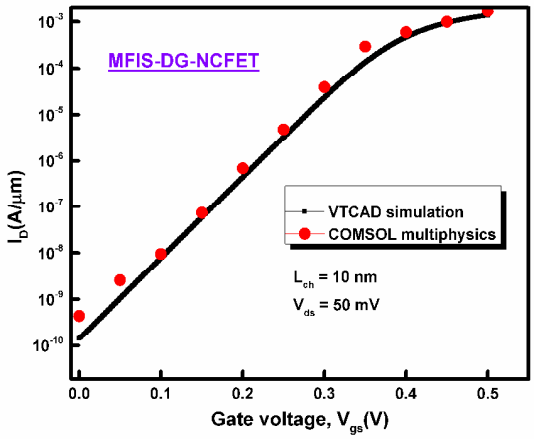



Fig. 2. Calibration of simulated results with experimental work of metal-ferroelectric-insulator-semiconductor double gate (MFIS-DG) NCFET at $V_{ds}$=50mV.

## III. Results and discussion

In this section of the study, we present the temperature analysis of the modified NCFET structure. We pay the attention to the electric field and lower temperature analysis. Fig. 3 shows the current-voltage characteristics of the modified NCFET structure on the linear scale.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:49:42 UTC from IEEE Xplore. Restrictions apply.

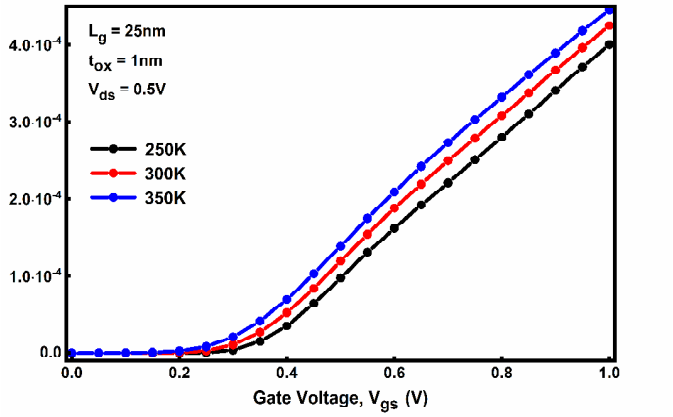



Fig. 3. The transfer (IDVG) characteristic of modified NCFET at different temperatures with $V_{ds}$=0.5V.

This variation is drawn out at different temperatures and as shown in the graph, the leakage current is decreasing with the reduction of temperature. That is why, the switching ratio ($I_{on}/I_{off}$) increases at a lower temperature as shown in Fig. 4

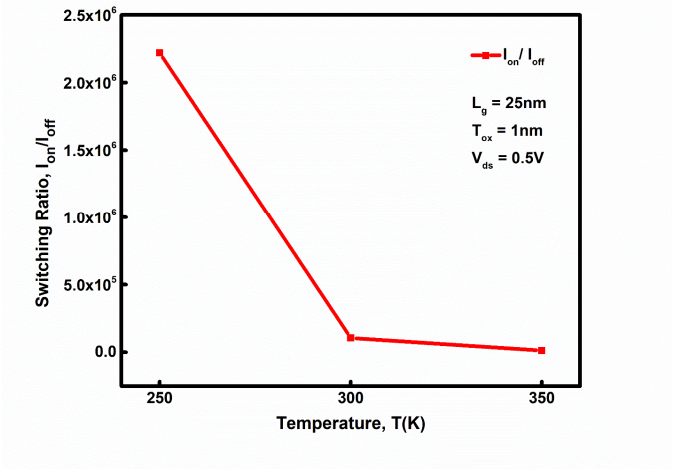



Fig. 4. The switching ratio variation for modified NCFET at a temperature range from 250K to 350K.

Also, from Fig. 5, we can see that the threshold voltage for the device is also decreasing with the increase in temperature [14-16]. Threshold voltage can be defined as the voltage that is mandatory to turn on the device. And with an increase in the operating temperature, the mobile electron gets free to set the channel firstly and thus the threshold voltage reduces at a higher temperature.

With the decrease in the channel length of the device, the short channel effects start dominating. Two of them are drain induced barrier lowering (DIBL) and subthreshold swing (SS). The first factor is drain induced barrier lowering which defines the channel modulation effect and with the increase in the temperature, the DIBL value starts increasing which is shown in the same Fig. 5

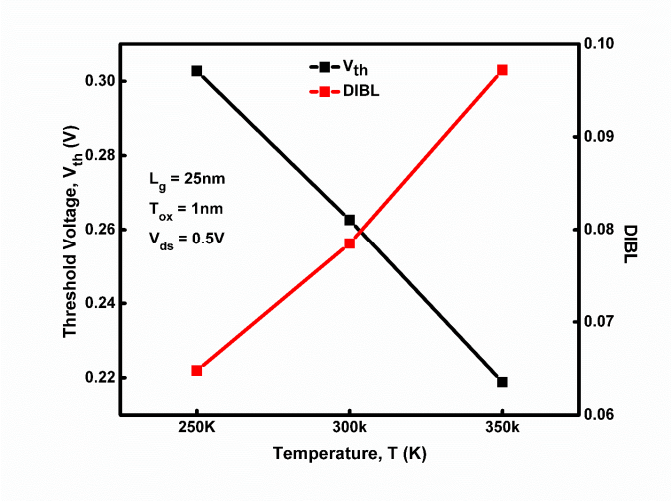



Fig. 5. The threshold voltage and DIBL behaviour with temperature variation for modified NCFET.

This is because the increasing temperature decreases the barrier near the drain side. The value of DIBL is calculated by using the formula given in (1) and is done for $V_{th}$ at $V_{ds}$ = 0.5 V.

$$\text{DIBL} = \frac{\partial V_{th}}{\partial V_{ds}} \quad (1)$$

Secondly, the minimum change of channel potential to conduct 10 order current magnitude for the conventional MOSFET device is 60mV which is called subthreshold tyranny or subthreshold swing slope [17]. With the decrease in the temperature, the SS value approaches the standard value which is shown in Fig. 6 in the form of a bar graph. So, it can be said that with an increase in the operating temperature, the minimum change of channel potential to conduct the 10 order current magnitude is increased.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:49:42 UTC from IEEE Xplore. Restrictions apply.

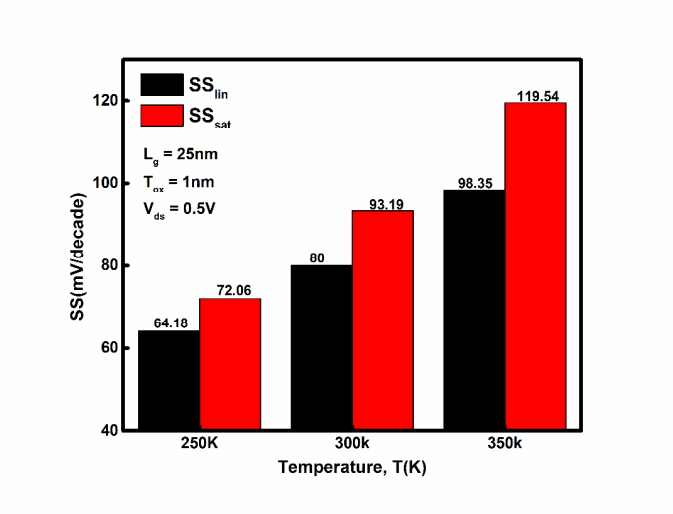



Fig. 6. The Subthreshold swing changes with temperature variation for modified NCFET structure.

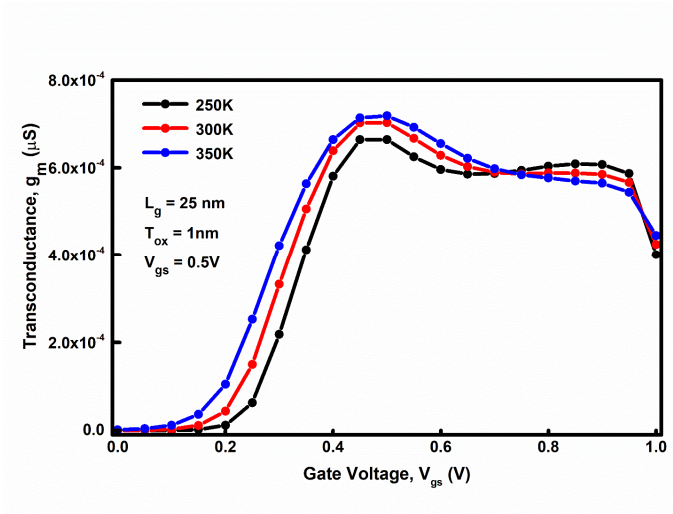



Fig. 7. The transconductance behaviour with temperature variation for modified NCFET structure.

TABLE II. SS AND DIBL COMPARISON FOR MODIFIED NCFET STRUCTURE AT DIFFERENT TEMPERATURES.

| Temperature | Wg [um] | Lg [um] | Vth [V] | DIBL | SSlin [mV/dec] | SSsat [mV/dec] |
|---|---|---|---|---|---|---|
| 250K | 0.05 | 0.025 | 0.257 | 0.0647 | 64.18 | 72.07 |
| 300K | 0.05 | 0.025 | 0.207 | 0.0784 | 80.00 | 93.19 |
| 350K | 0.05 | 0.025 | 0.150 | 0.0971 | 98.35 | 119.54 |

Further, the transconductance variation of the device with the $V_{gs}$ for the temperature range 250K to 350K at the constant drain-source voltage is given in Fig. 7. Transconductance or mutual conductance defines the electrical characteristic that gives the various output current of a device with the variation of input voltage of a device is given by (2).

$$g_m = \frac{\partial I_D}{\partial V_{gs}} \quad (2)$$

From the graph, we can see that with the increase in temperature, the peak value of transconductance increases. Thus, we can say that the mutual conductance increases with an increase in the temperature.

## IV. Conclusion

In this paper, we established the role of temperature change on an insulator layer of hafnium oxide ($HfO_2$) based NCFET in terms of some basic factors using the Visual TCAD simulator. The modified NCFET structure illustrates improved results with the lower leakage current and better switching ratio with the temperature change. The switching ratio increased by two decimal factors when we move from 350k to 250k. Also, the SS value is reduced by 34.7% and 39.71% in the linear and saturation region, respectively with a decrease in temperature from 350K to 250K and a threshold voltage decrease of 27.71% with a temperature change from 250K to 350K. The DIBL is increased by 21.18% as the temperature increases from 250K to 350K. The analog parameter also changed with the temperature change like the transconductance is increased by 7.45% as we move from 250K to 350K. So, we conclude that with an increase in temperature, the $V_{th}$ and subthreshold swing reduced significantly and the DIBL value, transconductance increase which have significant applications in various areas.

**References**

[1] G. E. Moore, Cramming more components onto integrated circuits, Proc. IEEE **2002**, 86, 82.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:49:42 UTC from IEEE Xplore. Restrictions apply.

[2] P. K. Bondyopadhyay, Moore's law governs the silicon revolution, Proc. IEEE **1998**, 86, 78.

[3] C. A. Mack, Fifty years of Moore's law, IEEE Trans. Semicond. Manuf. **2011**, 24, 202.

[4] G. E. Moore, Progress in digital integrated electronics, Tech. Dig. - Int. Electron Devices Meet. **1975**, 11.

[5] M. M. Waldrop, The chips are down for Moore's law, Nature, **2016**, 530, 144.

[6] E. Mollick, Establishing Moore's Law, IEEE Ann. Hist. Comput. **2006**, 28, 62.

[7] R. Aitken, V. Chandra, J. Myers, B. Sandhu, L. Shifren, G. Yeric, On device design for steep-slope negative-capacitance field-effect-transistor operating at sub-0.2V supply voltage with ferroelectric $HfO_2$ thin film, Symp. VLSI Technol., Dig. Tech. Pap., **2014**, 1.

[8] K. J. Kuhn, In 13th int. workshop computational electronics, IEEE, Piscataway, NJ, **2009**, p. 1.

[9] K. Boucart, A. M. Ionescu, Double-gate tunnel FET with high-gate dielectric, IEEE Trans. Electron Devices, **2007**, 54, 1725.

[10] Y. Khatami, K. Banerjee, Steep subthreshold slope n-and p-type tunnel-FET devices for low-power and energy-efficient digital circuits, IEEE Trans. Electron Devices, **2009**, 56, 2752.

[11] G. Fiori, G. Iannaccone, Ultralow-voltage bilayer graphene tunnel FET, IEEE, Electron Device Lett., **2009**, 30, 1096.

[12] S. Salahuddin, S. Datta, Use of negative capacitance to provide voltage amplification for low power nanoscale devices, Nano Lett., **2008**, 8, 405.

[13] G. Pahwa, A. Agarwal, Y. S. Chauhan, Numerical Investigation of Short Channel Effects in Negative Capacitance MFIS and MFMIS Transistors: Subthreshold Behavior, IEEE Trans. Electron Devices , **2018,** 65, 5130-5136.

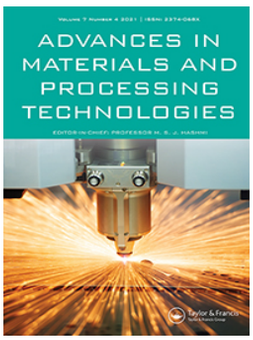
[14] A Kumar, MM Tripathi, R Chaujar, Reliability Issues of $In_2O_5Sn$ Gate Electrode Recessed Channel MOSFET: Impact of Interface Trap Charges and Temperature, IEEE Transactions on Electron Devices, **2018**, 65, 860-866.

[15] J Madan, R Chaujar, Temperature associated reliability issues of heterogeneous gate dielectric—Gate all around—Tunnel FET IEEE Transactions on nanotechnology, **2017**, 17, 41-48.

[16] N Gupta, R Chaujar, Investigation of temperature variations on analog/RF and linearity performance of stacked gate GEWE-SiNW MOSFET for improved device reliability, Microelectronics Reliability, **2016**, 64,235-241.

[17] B Kumar, R Chaujar, TCAD temperature analysis of gate stack gate all around (GS-GAA) FinFET for improved RF and wireless performance, Silicon, **2021**, 13, 3741-3753.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 08,2022 at 04:49:42 UTC from IEEE Xplore. Restrictions apply.



Taylor & Francis
Taylor & Francis Group

Advances in Materials and Processing Technologies

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tmpt20

# Wear behaviour on a Pin-on-Disc setup for thermal spray of coating using ABAQUS: study and computational analysis

R. S. Walia, Ravi Pratap Singh, Ravi Butola & Mohit Tyagi

**To cite this article:** R. S. Walia, Ravi Pratap Singh, Ravi Butola & Mohit Tyagi (2022): Wear behaviour on a Pin-on-Disc setup for thermal spray of coating using ABAQUS: study and computational analysis, Advances in Materials and Processing Technologies, DOI: 10.1080/2374068X.2022.2126926

**To link to this article:** https://doi.org/10.1080/2374068X.2022.2126926

Published online: 22 Sep 2022.

Submit your article to this journal

Article views: 31

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tmpt20

ADVANCES IN MATERIALS AND PROCESSING TECHNOLOGIES
https://doi.org/10.1080/2374068X.2022.2126926

Taylor & Francis
Taylor & Francis Group

Check for updates

# Wear behaviour on a Pin-on-Disc setup for thermal spray of coating using ABAQUS: study and computational analysis

R. S. Walia[a], Ravi Pratap Singh[b], Ravi Butola[c] and Mohit Tyagi[b]

[a]Department Production & Industrial Engineering, Punjab Engineering College, Chandigarh, India;
[b]Department of Industrial & Production Engineering, National Institute of Technology, Jalandhar, India;
[c]Department of Mechanical Engineering, Delhi Technological University, Delhi, India

**ABSTRACT**

The tribological simulation can be seen as a new study problem developed through modern information technology and computer simulation technology in combination with a variety of mathematical approaches in the field of tribology. There is a need to seek new conducts and means of in-depth study of the wear quantitative law, or else, it will become difficult for a fundamental breakthrough in the prevention and control of wear to be achieved. The tribological simulation research of the typical Pin-on-Disc sliding wear test system has been carried out by the adoption of computer technology and mathematical analysis methods to provide a new wear-analysis method for the simulation of the basic tribological contact problems. In the present research conduct, an attempt has been made to establish the wear simulation model considering multiple parameters based on experimental and theoretical facts. The parameters of the computational model have been determined through the experimental test data. The simulation model has been observed to reflect the basic relationship between input and output of the pin on the dry disc sliding wear test. Experiments performed have confirmed that, under the conditions analysed, the load effect on the wear rate is more critical than the other conditions. The wear rate varies linearly with the considered load conditions. The simulations have been conducted at the different angular velocities of 50, 60 and 70 rpm. These simulations have also revealed the effect of pin diameter on the depth of wear, and the deep insight into the increase of same with an increase in surface area.

**ARTICLE HISTORY**
Accepted 14 September 2022

**KEYWORDS**
Thermal spray of coating; wear; finite element method; ABAQUS; tribology; Pin-On-Disk; dry sliding

## 1. Introduction

The subject of Tribology requires knowledge of many different scientific specialities and as it requires substantial involvement of experimental science, the growth in this field has always been noticeably followed by experiments and significant observations [1]. A tribosystem is dependent on various variables such as contact stress, surface roughness, sliding speed, chemical environment, temperature, humidity, material type etc. and even a slight variation in the values of any two at a time yields a completely different set of results. This interdependency thus continues to be the major reason behind the lack of

**CONTACT** Ravi Pratap Singh singhrp@nitj.ac.in Department of Industrial & Production Engineering, National Institute of Technology, Jalandhar, India

© 2022 Informa UK Limited, trading as Taylor & Francis Group

any concrete or finite results in its element modelling. Friction and wear characteristics are of utmost importance to engineers and material scientists for designing and evaluating an engineering system [2,3]. Thus, friction and wear happen to be the most widely used data in evaluating the performance of a tribosystem. From experiments conducted in the past, it is clear that friction and wear are chaotic processes with complex, nonlinear nature varying between tribosystems. However, it is still possible to study it and accurately model the contact behaviour between the two materials [2].

Tribology refers to the study of the combined interaction or synergy of engineered systems having two pairs of liquid or solid in contact with each other. The relative motion between the two surfaces makes it impossible to impulsively predict the behaviour of contact. This area holds tremendous challenges and unexplored regions for further research [3]. The objective of most tribological testing is to emphasise the frictional power values and to measure the wear values for a wear system. With a specific end goal to completely comprehend the rubbing and wear values, the parameters used to portray grinding and wear must be studied on a nodal level. Most tribological applications deal with non-conformal surfaces with very small contact areas. Such systems like bearings are subjected to very high pressures [1–3].

One of the first wear simulation approaches was developed by Podra and Andersson [1], who analysed the wear behaviour with commercial finite element software ANSYS, where the linear wear law was used. The Lim and Ashby wear map was used to observe the wear mechanism of a spherical Pin-on-Disc. FEA wear simulation output was then treated based on wear coefficient-sliding distance change equivalence and was evaluated on a relative scale of comparing design options. Kim et al. [2] prescribed a numerical approach to study the wear characteristics. The approach is based on the progressive accumulation of wear material when a metal oscillates over another metal contact. The methodology adopted uses a reciprocating Pin-on-Disc tribometer to measure a wear rate for the pair which is then used as an input to an FEA that simulates a block-on-ring experiment. Bosman et al. [3] used a mild wear map for boundary lubricated contacts in their experimental study. In this, the transitions from mild to severe wear and vice versa are studied. A wear map of grey cast iron is reported by Riahi et al. [4] to summarise the wear rate and its mechanism on A30 grey cast iron. The criterion for the transition from mild to severe wear is a thermal one while the mild wear model is based on the hypothesis for an additive sacrificial chemical layer [5,6].

Concerning ourselves with protective coatings, Broszeit et al. [7] compared the deposition and application of CrN, the coating deposited by physical vapour deposition (PVD) to that of chromium (Cr) electroplating and behaviour of titanium nitride (TiN). CrN solves the problems observed because the coating thickness of CrN slightly decreases by increasing the temperature while the thickness of TiN increases at elevated temperatures. Electroplated chromium (Cr) shows less deposition rate than TiN and CrN. Costa et al. [8] proposed that the high-velocity oxygen fuel (HVOF) WC–10%Co–4%Cr thermal sprayed and TiN, CrN, and DLC physical vapour deposition (PVD) coatings were applied to increase titanium substrate wear and friction resistance. Graphite carbon structures for Ti–6Al–4 V alloy DLC coated and aluminium – bronze 630 tribological pair work as a solid lubricant to prevent the wear. According to the need for industrial application; Navinsek et al. [9] studied single layer, low and high-temperature CrN coatings, and double TiN + CrN coatings. CrN PVD coatings deposited at 450°C and

250°C are a new field of application. The thickness of 5–10 mm of CrN coating works more corrosion resistant than a traditional coating of TiN, thermally stable up to 700°C. Double TiN + CrN coatings were used as a very high abrasive resistant coating in the production of rotors (electromotors), and for cold forming and forging in the mass manufacturing of screws. Zhang and Meng [10] suggested a linear sliding wear model and a simplified mathematical method for it. The effects of geometrical parameters, material properties, and applied operating conditions on the progress of dimensional and volumetric wear rates were looked into for a rotating rotor bushing that slides on the ground plane.

Unal et al. [11] studied the friction and wear behaviour of pure polytetrafluoroethylene, glass fibre reinforced and bronze and carbon (C) filled polytetrafluoroethylene polymers. Experiments were performed at sliding speeds of 0.32, 0.64, 0.96, and 1.28 m/s under a load of 5, 10, 20, and 30 N. The results yielded that, for pure PTFE, the friction coefficient decreased with the increase in load. It was found that glass fibre, bronze, and carbon fillers when used as additives to PTFE effectively reduced the overall wear rate of the composite [12]. In Der-Liang et al.'s experiments [13], WC – Ni metal is doped with zirconium carbide or titanium carbide which improves its abrasive and erosive wear resistance. TiC has been identified to make platelet grain structure for strengthening leading to improved abrasion resistance. ZrC helps in suppressing grain coarsening and abnormal growth of WC grains as well as increasing the hardness of the material.

Onuoha et al. [14] researched the work in which TiC-stainless steel (grade 316 L) cermets have been made, with the TiC grain size varied through heat-treatment, and steel binder contents ranging between 10 and 30 vol%. Microstructural analysis of this work presented mean TiC grain sizes of 4 and 10 μm, respectively, for fine- and coarse-grained cermets with the grain size quite consistent as a function of binder content. Hovsepian et al. [15] studied a molybdenum and tungsten doped carbon-based coating (Mo – W–C) material, developed to decrease friction in boundary lubricated sliding conditions at ambient and at high temperatures. Mo – W–C coating presented the lowest friction coefficient among several commercially available DLC coatings at ambient temperature. Bolelli et al. [16] presented the tribological properties of High-Velocity Air-Fuel sprayed Fe–31Cr–12Ni–3.6B–0.6C (wt%) coatings studied as a function of the deposition parameters. At room temperature, ball-on-disk geometry sliding against Al2O3 material is controlled by abrasive grooving and interlamellar cracking having some tribe-oxidation. But at 700°C, the sliding wear rates are evened out, and they are one order of magnitude higher than that at room temperature, because of severe abrasive grooving. Culliton et al. [17] investigated the tribological properties of as-sprayed NiCrBSi and WC/12Co Flame Spray coatings applied to two cast aluminium alloys: high-copper LM24 (AlSi8Cu3Fe), and low-copper LM25 (AlSi7 Mg). The coefficient of friction decreased from 0.69–0.72 to 0.12–0.35. Wear (Pin-on-Disc) decreased by a factor of 103–104, which was due to the high surface roughness of the coatings. Abderrahman et al. [18] investigated defect depth in the material using solving the heat equation by finite element method and simulation has been carried out to detect internal defect depth in materials. Wear of materials can be defined as progressive damage to a surface caused by relative motion concerning another substance [19]. An experimental investigation by Mahamani [20] shows that normal load, sliding velocity, and reinforcement ratio offer a significant influence on wear

performance. The fine-grained cermets exhibited better wear resistance and hardness compared to the result of the other. The specific wear rate increased with applied load and/or binder content, for both fine- and coarse-grained materials [21,22]. Ashwini M. et al. [23] have reported a development approach for rapid manufacturing by CNC process and find out machining parameters for intricate parts and calculate the effect of different machining parameters by using finite element analysis.

In view of the above discussion, an attempt has been made to establish the wear simulation model considering multiple parameters based on experimental and theoretical facts. All three classifications have been used to describe the type of wear present to understand wear mechanisms between solids as it often influences friction behaviour and induces failure. The parameters of the computational model have also been determined through the experimental test data. The objectives of this research study are to model the Pin-on-Disc setup on ABAQUS and analyse the wear behaviour of the thermal spray coating, and consequently, obtain the relationship between the angular velocity of the disc and wear depth.

## 2. Material and methods

### 2.1. Material properties for POD setup

In the property module, the material properties of hard chrome (Disc) and cylinder liner (Pin) are specified, and body sections are assigned to specified regions of the pin and the disc. The disc and pin consist of one material and one section. In Table 1, the different material properties are presented.

### 2.2. Dimensional design of POD setup

The considered dimensions for the pin and the disc are represented in Table 2 and the same has also been reflected in Figure 1(a,b) .

### 2.3. Finite element analysis

The finite modelling simulation has been designed to model the macroscopic 2D profile of the contacting surface of the pin generated due to wear as a function of the contact pressure and sliding distance for a certain sliding speed. Finite element analysis (FEA) is the modelling of products and systems in a virtual environment, to find and solve potential (or existing)

Table 1. Material properties of pin and disc.

| Section Name | Material Behaviour | Type | Young's Modulus [N/mm$^2$] | Poisson's Ratio | Section Type |
|---|---|---|---|---|---|
| Pin | Mechanical, Elastic | Isotropic | 150000 | 0.3 | Solid/Homogenous |
| Disc | Mechanical, Elastic | Isotropic | 103420 | 0.29 | Solid/Homogenous |

Table 2. Dimensions of the pin and disc.

| The Disc | | The Pin | |
|---|---|---|---|
| Outer radius | 30 mm | Radius | 2.5 mm |
| Inner radius | 5 mm | Thickness (depth) | 15 mm |
| Thickness (height) | 5 mm | | |

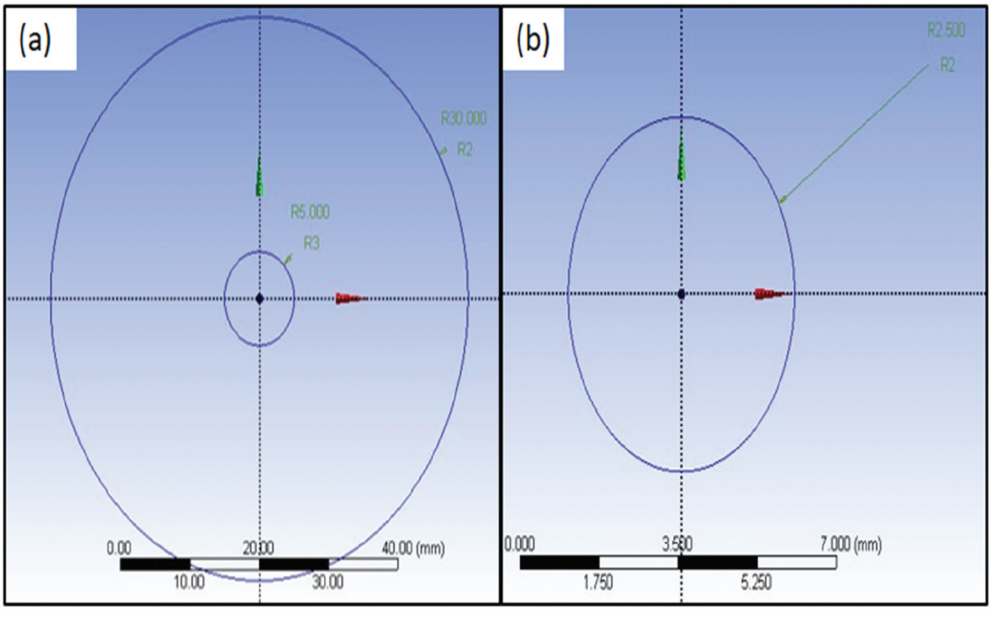



**Figure 1.** (A, b) Disc and pin dimensions.

structural or performance issues. The computational power and low cost of modern-day computers have made Finite Element Analysis feasible for many industries and research groups alike. FEA is the practical application of the finite element method (FEM) and involves the calculation of deflection, stress, strain, buckling, and other phenomena. The finite element analysis implemented in this work develops a numerical model for calculating depth wear. Subroutines have been developed and used in conjunction with Abaqus and adaptive meshes to study the changes at the surfaces in contact, considering 3D wear effects coupled with mechanical behaviour, in opposition to some models reported in the literature that consider 2D geometries and/or only compute wear at the post-processing.

A finite element (FE) model comprises a system of points, called 'nodes', which form the shape of the design. Connected to these nodes are the finite elements themselves which form the finite element mesh and contain the material and structural properties of the model, defining how it will react to certain conditions. The density of the finite element mesh varies throughout the material, depending on the anticipated change in stress levels of a particular area. Regions that experience high changes in stress are provided with a higher mesh density than those that experience little or no stress variation. Points of interest may include fracture points of previously tested material, fillets, corners, complex detail, and high-stress areas.

### 2.4. Solid modeling

Solid modelling is the first step for doing any 3D analysis and testing and it gives 3D physical outline for new products. In the present work, as the pin and disc are simple configurations, the modelling has been carried out in ABAQUS itself. Figure 2 is reflecting the assemble of pin and disc model.

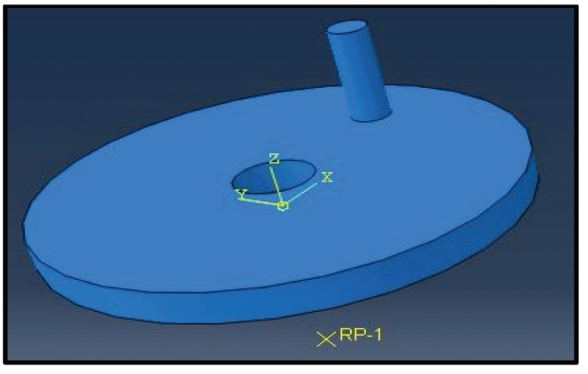



Figure 2. Assembly of pin and disc model.

### 2.5. Assumptions for analysis

- The interaction between the pin and disc is of the frictional type with a penalty of 0.3 [4,7].
- The material of both pin and disc is isentropic [4,7].
- Specific wear rate (K = 6E-5) is constant [4,7].

## 3. Static analysis

This analysis aims to study Pin-on-Disc setup and verification of the results within the desirable limits. ABAQUS software is used to analyse the contact pressure by performing static analysis for the given load and angular velocity to determine the wear depth. This analysis involves discretisation in the form of meshing, assignment of boundary conditions and loading of forces.

### 3.1. Meshing design

This means that the coordinate location of every node and the size, shape and connectivity of every element has been determined manually. This technique, in contrast to creating solid models using computer-aided design (CAD) components of ANSYS, automatically picks and connects the nodes, creating elements based on a few user-defined restrictions. The elements used in meshing are of the type C3D8. It is a continuous solid element and has 8 nodes, one at each corner. Upon mesh independence, this was found to be the most suitable setup, as the model is made in such a way that the elements are almost square in the contact region and the computational time is the least.

### 3.2. Boundary conditions for FE model and load

It is important to choose the boundary conditions in a proper way to simulate the clamping of the clamp holder. The boundary conditions should not generate extra stresses in the pin and the setup should be able to apply a force on the top.

Coupling for the disc cannot be used as the material can't move after this due to its Poisson's ratio. Therefore, the outer surface of the pin is fixed in the x and y-direction and the movement is free in the z-direction. Figure 3 is exploring the different boundary conditions of pin and disc. All these boundary conditions on the surface make sure that the surface on the underside of the pin can move towards the disc in the negative-z-direction and that the surfaces of the pin and disc stay aligned.

There are two ways in which a force can be given on the top surface of the disc. The first one is to make a force vector to act at a point. The second one is to let pressure act on a surface. The second one was deemed best since the first option results in a non-homogenous force distribution in the interface region. Figure 4 is reflecting the load depicting on the pin portion.

## 4. Results and discussion

For evaluating the conducted experimental results, required parameters have been selected and a visualisation module has further been created on the screen. The proper selection of the boundary conditions of the pin can be seen through the simulation results. On the top surface of the pin, there has not been observed any extra stress generated. Stresses transpired owing to the undertaken boundary conditions can also be realised. In addition to this, no meaningful displacement has been revealed to occur under the effect of considered parameters to plot the nodes at the lower surface of the disc part in the z-direction. This further confirms the appropriate selection of the boundary conditions of the disc.

In the first step, the pressure is equally distributed in the pin as being reflected by a single colour. Thus, the von Mises stresses are of the same magnitude at every point in the pin. Whereas in the second step, at the inside of the pin the stresses are higher as compared to the stresses generated on the outer side. Because of the circular movement

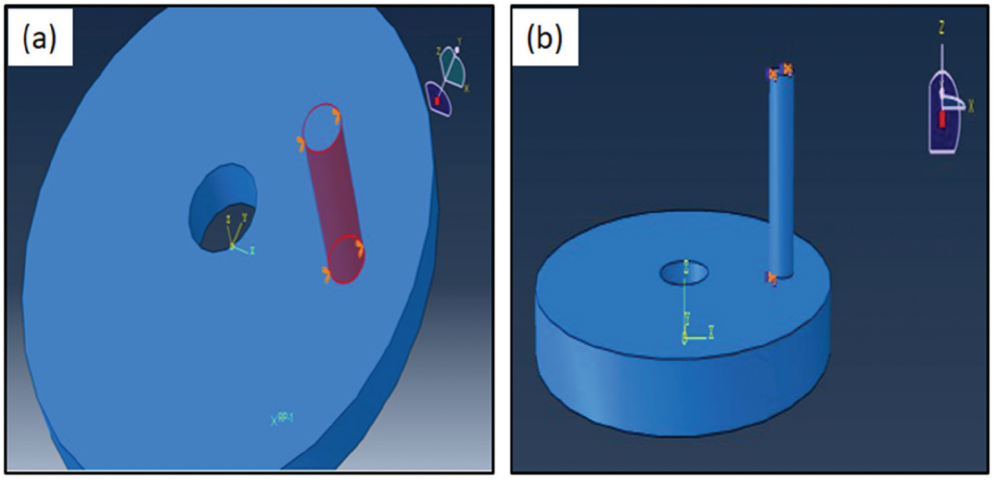



Figure 3. Boundary conditions of pin and disc.

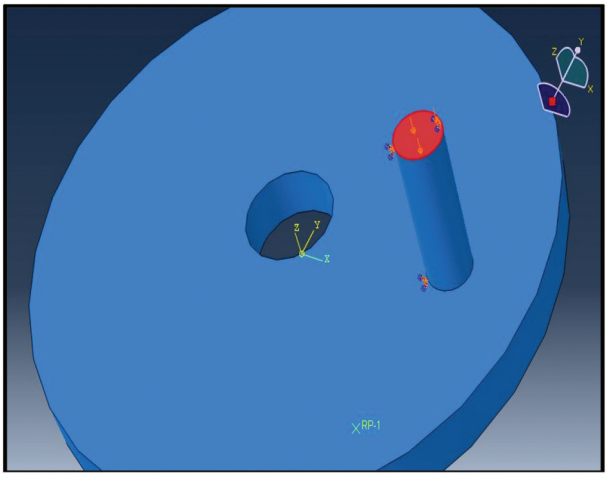



Figure 4. Plot depicting load on the pin.

of the disc, the disc stresses in the second step float a little bit. Due to the rough mesh, the stresses at the front of the pin are higher than at the end of the pin and the stresses near the centre of the pin are higher than those on the outer radius. For evaluating model robustness, different values of the quantities have to be checked. The basis on which these values are decided is calculated with the equations presented by Podra and Andersson [1] The prescribed values should have a certain fixed value at a certain point in the simulation. The contact pressure in the last increment of the first step should be equal to the applied pressure. The equation used for the calculation of the wear depth is the most popularly accepted Archard equation which states that:

$$W = K \times S \times P$$

Where,

W denotes the worn volume,

K denotes the wear per unit load per sliding distance and

S denotes the sliding distance,

P denotes the contact pressure between the sliding surfaces.

The key outcome expected from a Finite-Element wear analysis is Contact Pressure (CPRESS). The contact pressure magnitude is recorded for all the sets of inputs; a specific wear-rate value is taken from the experimental analysis, and the sliding distance is calculated from the input parameters. And finally, by putting all the values in the Archard equation, the wear depth is calculated for all sets of inputs. Table 3 is reflecting the various pressure values for the respective load on pin surface. However, Table 4 is exploring the contact pressure between the pin and the disc.

### 4.1. Experimental validation

For any simulation work, validation of results is the most important step to verify the accuracy of the simulation work. For experiments, a hard chrome disc and cylinder liner pin are tested in Pin-on-Disc apparatus with the same parameters as used previously in

**Table 3.** Pressure values for the respective load on pin surface.

| Load (N) | Pressure (MPa) |
|---|---|
| 50 | 2.564 |
| 60 | 3.055 |
| 70 | 3.565 |
| 80 | 4.074 |
| 90 | 4.583 |
| 100 | 5.093 |

**Table 4.** Contact pressure between the pin and the disc.

| S. No. | Load(N) | Contact Pressure (MPa) at 50 rpm | Contact Pressure (MPa) at 60 rpm | Contact Pressure (MPa) at 70 rpm |
|---|---|---|---|---|
| 01 | 50 | 4.162 | 3.588 | 4.247 |
| 02 | 60 | 4.912 | 4.279 | 5.056 |
| 03 | 70 | 5.732 | 4.993 | 5.906 |
| 04 | 80 | 6.55 | 5.699 | 6.747 |
| 05 | 90 | 7.369 | 6.415 | 7.597 |
| 06 | 100 | 8.189 | 7.128 | 8.433 |

By substituting the values of contact pressure and taking the wear coefficient K=6E−5 constant for coating material; the wear depth is calculated as shown in Table 5.

the simulation work i.e. sliding speed, load, and running time. Then, the values of wear depth found have been noted down and mentioned in Table 6.

Comparative graphs are then plotted between the experimental data and analytical data for a better understanding of result accuracy. Figures 5–7 have been plotted to reflect and understand the behaviour of wear concerning the different loading conditions, under the effect of three rpm values namely; 50 rpm, 60 rpm, and 70 rpm, correspondingly. The obtained experimental results for the wear have been revealed to be in fine agreement with the simulation-based results. The values for wear exemplified in Figure 8 (at 70 rpm) show the variation the experimental data under the conditions of higher rpm values which further can be observed as the main cause for such attained variation with the

**Table 5.** Wear depth at different rpm of the disc.

| S. No. | Load(N) | Wear (μm) at 50 rpm | Wear (μm) at 60 rpm | Wear (μm) at 70 rpm |
|---|---|---|---|---|
| 01 | 50 | 0.1307 | 0.13526 | 0.18686 |
| 02 | 60 | 0.1543 | 0.1613 | 0.22237 |
| 03 | 70 | 0.18 | 0.18822 | 0.25975 |
| 04 | 80 | 0.20577 | 0.21484 | 0.29674 |
| 05 | 90 | 0.2315 | 0.24183 | 0.3341 |
| 06 | 100 | 0.25726 | 0.2687 | 0.3709 |

**Table 6.** Experimental values of wear at different rpm of the disc.

| S. No. | Load(N) | Wear (μm) at 50 rpm | Wear (μm) at 60 rpm | Wear (μm) at 70 rpm |
|---|---|---|---|---|
| 01 | 50 | 0.1286 | 0.1432 | 0.18422 |
| 02 | 60 | 0.1524 | 0.18711 | 0.23245 |
| 03 | 70 | 0.18232 | 0.20351 | 0.26241 |
| 04 | 80 | 0.21478 | 0.22841 | 0.3113 |
| 05 | 90 | 0.2426 | 0.2517 | 0.3569 |
| 06 | 100 | 0.26289 | 0.28459 | 0.39287 |

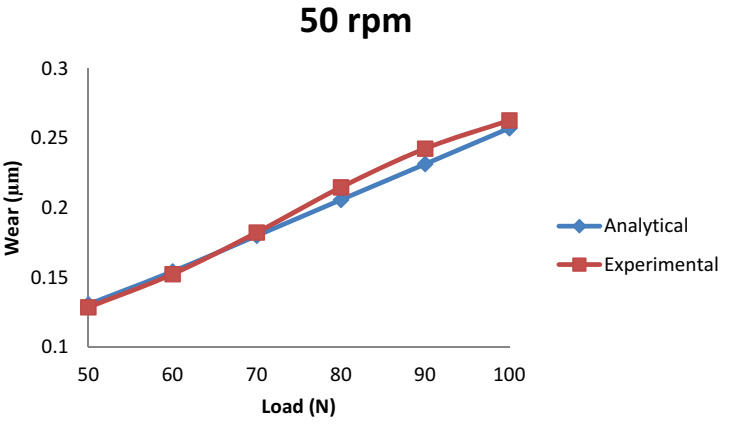



Figure 5. Graph showing comparative wear versus load graph at 50 rpm.

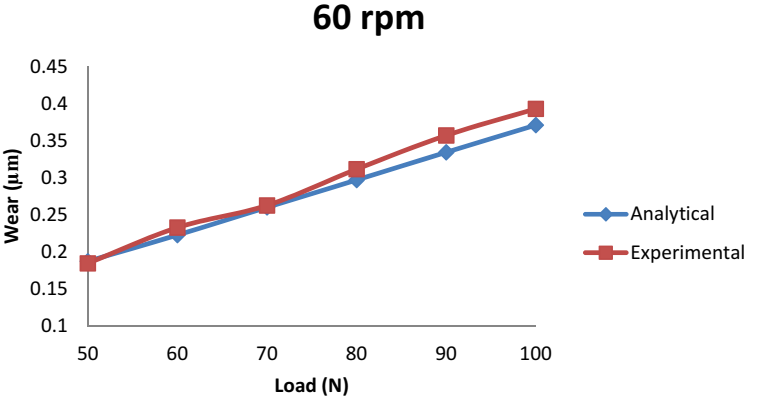



Figure 6. Graph showing comparative wear versus load graph at 60 rpm.

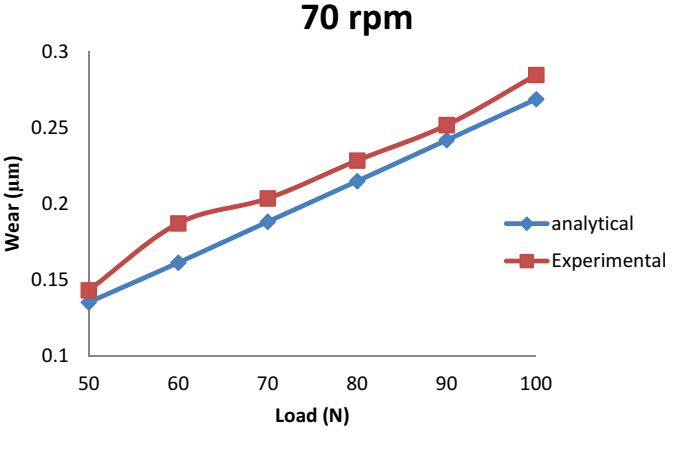



Figure 7. Plot showing comparative wear versus load graph at 70 rpm.

simulation results. In all the three considered operating conditions, the wear behaviour has been revealed to get varied linearly with the varied values of loads.

Based on the above graphs and results it can be concluded that the results obtained from experimental and analytical methods are identical as; with an increase in load and angular velocity, wear depth also increases. The two important outcomes from a wear model from an engineering standpoint are: change in dimensions and localised effects of wear. The analysis conducted through the employed ABAQUS software tool has also been represented in Figures 8(a–f) and 9(a–f). Figure 8(a–f) is showing contact pressure behaviour on the ABAQUS tool under the conditions of 50 rpm with a load of 50 N, 60 N,

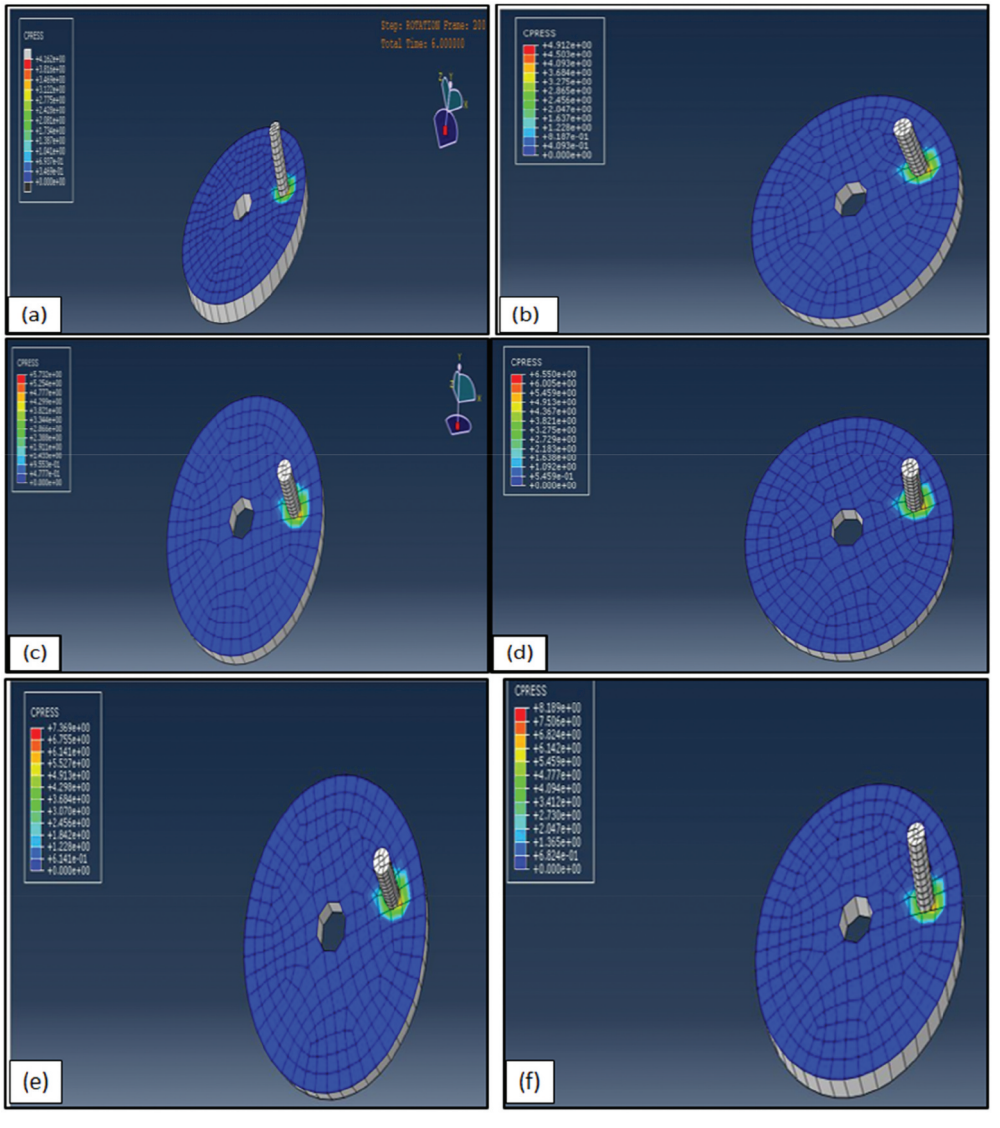



**Figure 8.** Plots showing contact pressure snapshots of ABAQUS at 50 rpm with a load of 50N, 60N, 70N, 80N, 90N and 100N as shown in a, b, c, d, e, f, respectively.

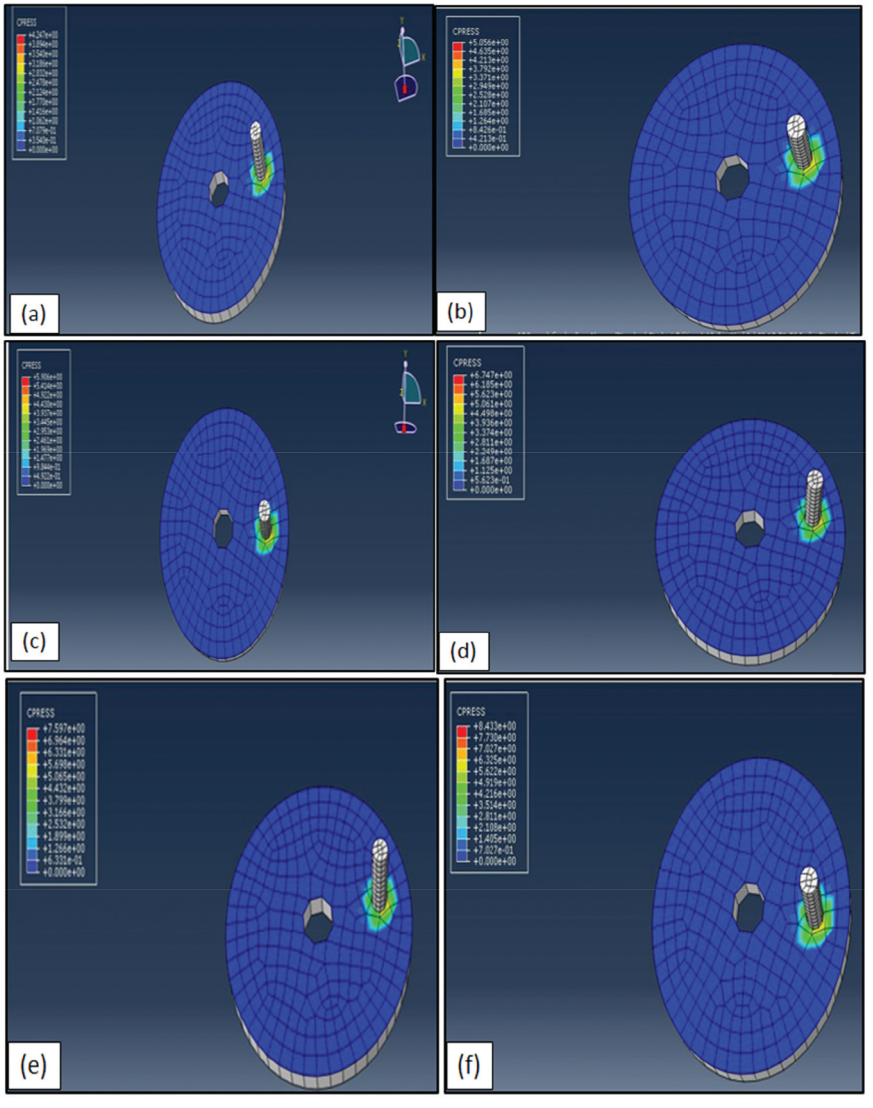



**Figure 9.** Plots showing contact pressure snapshots of ABAQUS at 70 rpm with a load of 50N, 60N, 70N, 80N, 90N, and 100N as shown in a, b, c, d, e, f, respectively.

70 N, 80 N, 90 N, and 100 N. Whereas, Figure 9(a–f) is exemplifying the contact pressure behaviour on the ABAQUS tool under the conditions of 70 rpm with a load of 50 N, 60 N, 70 N, 80 N, 90N, and 100 N.

Dimensional changes resulting from wear are rarely uniformly distributed; therefore, any wear prediction model averages the changes in dimension across the contact in consideration. The FEA model can give wear results at a nodal level, which can be averaged across the contact.

## 5. Conclusions

In the present investigation, the analysis of a Pin-on-Disc system has been performed using a software tool called - *ABAQUS CAE 6.13*. Based on the analytical Pin-on-Disc test results, the finite element model has also been updated, which has further been revealed with decent and reliable outcomes. This developed model has also been substantiated through a comparative analysis with the experimental results and has been observed to possess a fine degree of accuracy. The following inferences can be made;

- The pressure distribution that has been calculated by the finite element program is found to be an accurate representation of the pressure distribution in real-life practical problems.
- Through the introspection of the simulation results, it has been found that as the load on the pin increases, depth wear also gets increased while angular velocity remains the same.
- Experiments performed have confirmed that, under the conditions analysed, the load effect on the wear rate is more critical than the other conditions. The wear rate varies linearly with the considered load conditions.
- The simulations have been conducted at the different angular velocities of 50, 60, and 70 rpm. These simulations have also revealed the effect of pin diameter on the depth of wear and the deep insight into the increase of same with an increase in surface area.

## Disclosure statement

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

## ORCID

Ravi Pratap Singh http://orcid.org/0000-0001-5827-3560

## References

[1] Podra P, Andersson S. Simulating sliding wear with finite element method. Tribol Int. 1999;32:71–81.
[2] Kim NH, Won D, Burris D, et al. Finite element analysis and experiments of metal/metal wear in oscillatory contacts. Wear. 2005;258(11–12):1787–1793.
[3] Bosman R, Schipper DJ, "Mild wear maps for boundary lubricated contacts", Wear 2012, 280–281, 54–62.
[4] Riahi AR, Alpas AT. Wear map for grey cast iron. Wear. 2003;255(1–6):401–409.
[5] Pearson SR, Shipway PH, Abere JO, et al. The effect of temperature on wear and friction of a high strength steel in fretting. Wear. 2013;303(1–2):622–631.
[6] Friedrich C, Berg G, Broszeit E, et al. PVD Cr N coatings for tribological application on piston rings. Surf Coat Technol. 1997;97(1–3):661–668.
[7] Broszeit E, Friedrich C, Berg G. Deposition properties and applications of PVD CrN coatings. Surf Coat Technol. 1999;115(1):9–16.
[8] Costa MYP, Cioffi MOH, Voorwald HJC, et al. An investigation on sliding wear behavior of PVD coatings. Tribol Int. 2010;43(11):2196–2202.

[9] Navinšek B, Panjan P, Milošev I. Industrial applications of CrN (PVD) coatings, deposited at high and low temperatures. Surf Coat Technol. 1997;97(1–3):182–191.
[10] Zhang M, Meng G. "Numerical simulation of sliding wear between the rotor bushing and ground plane in micromotors. Sens Actuators a. 2006;126(1):15–24.
[11] Unal H, Mimaroglu A, Kadıoglu U, et al. Sliding Friction and Wear Behaviour of Polytetrafluoroethylene and its Composites under Dry Conditions. Mater Des. 2004;25 (3):239–245.
[12] Cocks M. Interaction of Sliding Metal Surfaces. J Appl Phys. 1962;33(7):2152–2161.
[13] Yung D.L, Antonov M, Veinthal R . Wear behaviour of doped WC–NI based hard metals tested by four methods, Wear. 2016;352:171–179.
[14] Onuoha CC, Jin C, Farhat ZN, Kipouros, GJ, Plucknett, KP . The effects of TiC grain size and steel binder content on the reciprocating wear behaviour of TiC-316L stainless steel cermets, Wear. 2016;350:116–129.
[15] Hovsepiana PE, Mandala P, Ehiasariana AP, Sáfrán, G, Tietema, R, Doerwald, D . Friction and wear behaviour of Mo–W doped carbon-based coating during boundary lubricated sliding, Appl. Surf. Sci. 2016;366:260–274.
[16] Bolelli G, Milanti A, Lusvarghi L, Trombi, L, Koivuluoto, H, Vuoristo, P, .Wear and impact behaviour of high velocity air-fuel sprayed Fe–Cr–Ni–B–C alloy coatings, Tribol. Int. 2016;95:372–390.
[17] Culliton D, Betts A, Carvalho S, Kennedy, D . Improving tribological properties of cast Al-Si Alloys through application of wear-resistant thermal spray coatings, J. Therm. Spray Technol. 2013;22(4):491–501.
[18] Eddazi A, Belattar S. Determination of the defect depth by infrared thermography using finite element method. J Eng Technol. 2017;6(2):215–219.
[19] Hegadekatte V, Huber N, Kraft O. Modeling and simulation of wear in a pin on disc tribometer. Tribology Lett. 2006;24:51–60.
[20] Mahamani, Mahamani A. Dry sliding wear and friction characteristics of AA6061 - TiB2/zrb2 in-situ metal matrix composites at room temperature. Int J Surf Sci Eng. 2017;11(2):148–173.
[21] Kanavalli. Application of User Defined Subroutine UMESHMOTION in Abaqus for Simulating Dry Rolling/Sliding Wear [Master Thesis]. Stockholm (Sweden): Royal Institute of Technology (KTH); 2006.
[22] Chowdhury MA, Helali M. The effect of relative humidity and roughness on the friction coefficient under horizontal vibration. Open Mech Eng J. 2008;2:128–135.
[23] Ramteke AM, Ashtankar KM. Finite element analysis study for machining parameters using CNC-RP machining of hip joint part with innovative sacrificial support structure as a fixture. J Eng Technol. 2018;7:337–359.