

SCHOLARLY PUBLICATIONS

*A CURRENT AWARENESS BULLETIN
OF RESEARCH OUTPUT*

@DTU

(101st Edition)

MAY 2021

BY: CENTRAL LIBRARY

DELHI TECHNOLOGICAL UNIVERSITY

(FORMERLY *DELHI COLLEGE OF ENGINEERING*)

GOVT. OF N.C.T. OF DELHI

SHAHBAD DAULATPUR, MAIN BAWANA ROAD

DELHI 110042

PREFACE

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

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

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

This issue contains the information published during **May, 2021**. The arrangement of the contents is alphabetical. The full text of the article which is either subscribed by the university or available in the web is provided in this bulletin.

Central Library

CONTENTS

1. A complete consumer behaviour learning model for real-time demand response implementation in smart grid, *6.Swati Sharda, 3.Mukhtiar Singh* and *3.Kapil Sharma*, IT and Electrical, DTU
2. A Hybrid Jaya–Powell’s Pattern Search Algorithm for Multi-Objective Optimal Power Flow Incorporating Distributed Generation, *6.Saket Gupta, 3.Narendra Kumar*, Laxmi Srivastava, Hasmat Malik, Alberto Pliego Marugán and Fausto Pedro García Márquez, Electrical, DTU
3. A Hybrid Model for Combining Neural Image Caption and k-Nearest Neighbor Approach for Image Captioning, *8.Kartik Arora, 8.Ajul Raj, 8.Arun Goel* and *3.Seba Susan*, IT, DTU
4. A Language-Independent Speech Sentiment Analysis Using Prosodic Features, *8.Monil Bansal, 8.Sampriti Yadav* and *3.Dinesh K. Vishwakarma*, IT, DTU
5. A Review on Computation Methods Used in Photoplethysmography Signal Analysis for Heart Rate Estimation, Pankaj, Ashish Kumar, Rama Komaragiri and *3.Manjeet Kumar*, Electronics, DTU
6. Actionable strategy framework for digital transformation in AECO industry, *6.Sanjay Bhattachary* and K.S. Momaya, DTU
7. Advancements in steam distillation system for oil extraction from peppermint leaves, *8.1.Ravi Kant* and *3.Anil Kumar*, Mechanical and Environmental, DTU
8. An Adaptive Master-Slave Technique using Converter Current Modulation in VSC-based MTDC System, *3.Radheshyam Saha, 3.Madhusudan Singh* and *6.Ashima Taneja*, Electrical, DTU
9. Analysis of COVID-19 Tweets During Lockdown Phases, *7.Prince Tyagi, 8.Naman Goyal* and *3.Trasha Gupta*, Mechanical and Mathematics, DTU

10. Are exports and imports of India's trading partners cointegrated? Evidence from Fourier bootstrap ARDL procedure, **6.Khyati Kathuria** and **3.Nand Kumar**, Humanities, DTU
11. Comparative Study of Different Image Captioning Models, **8.Sahil Takkar**, **8.Anshul Jain** and **8.Piyush Adlakha**, Computer Science, DTU
12. Comparative study of wind induced mutual interference effects on square and fish-plan shape tall buildings, **8.SUPRIYA PAL**, **3.RITU RAJ** and **3.S ANBUKUMAR**, Civil, DTU
13. Comparison of Genetic Algorithm and Taguchi Optimization Techniques for Surface Roughness of Natural Fiber-Reinforced Polymer Composites, **6.Susheem Kanwar**, **3.Ranganath M. Singari** and **3.Ravi Butola**, DTU
14. Comparison of response of building against wind load as per wind codes [IS 875 – (Part 3) – 1987] and [IS 875 – (Part 3) – 2015], **7.Naveen Suthar** and **3.Pradeep K. Goyal**, Civil, DTU
15. Data Preprocessing based Connecting Suicidal and Help-Seeking Behaviours, **8.Aayush Mittal**, **8.Abhishek Goyal** and **8.Mohit Mittal**, Computer Science, DTU
16. Deep and Shallow Covariance Feature Quantization for 3D Facial Expression Recognition, Walid Hariri, Nadir Farah and **3.Dinesh Kumar Vishwakarma**, IT, DTU
17. Degraded Document Image Binarization using Novel Background Estimation Technique, **8.Harshit Jindal**, **3.Manoj Kumar**, **8.Akhil Tomar** and **8.Ayush Malik**, Computer Science, DTU
18. Demystifying deepfakes using deep learning, **8.Raj Kumar Singh**, **8.Prachi Vinod Sarda**, **8.Shruti Aggarwal** and **3.Dinesh Kumar Vishwakarma**, IT, DTU
19. Design and Analysis of a Bandpass Filter Using Dual Composite Right/Left Handed (D-CRLH) Transmission Line Showing Bandwidth Enhancement, **6.Priyanka Garg** and **3.Priyanka Jain**, Electronics, DTU

20. Design and evaluation of stand-alone solar-hydrogen energy storage system for academic institute: A case study, **8.Alfred John**, **8.Srijit Basu**, **8.Akshay** and **3.Anil Kumar**, Mechanical and Environment, DTU
21. Design of Compact Circular Microstrip Patch Antenna using Parasitic Patch, **3.Richa Sharma**, **3.N.S.Raghava** and **3.Asok De**, ECE, DTU
22. Design of photonic crystal OR gate with multi-input processing capability on a single structure, **6.Chandan Kumar**, **6.Punit**, **6.Praveen Kumar**, Preeti Rani and **3.Yogita Kalra**, Applied Physics, DTU
23. Designing and Analyzing the Brake Master Cylinder for an ATV vehicle, Shubham Upadhyaya, Divyam Raj, Kaushal Gupta, Rakesh Chander Saini, Ramakant Rana and **3.Roop Lal**, Mechanical, DTU
24. Detection of Cyberbullying on Social Media Using Machine learning, **8.Varun Jain**, **8.Vishant Kumar**, **8.Vivek Pal** and **3.Dinesh Kumar Vishwakarma**, IT, DTU
25. Detection of Malicious Transactions using Frequent Closed Sequential Pattern Mining and Modified Particle Swarm Optimization Clustering, **3.Rajni Jindal** and **3.Indu Singh**, Computer Science, DTU
26. Determinants Of Job Opportunities In Skill Development Institutions: Indian Perspective, **6.Manoj Kumar**, **3.Suresh Kumar Garg** and Shraddha Mishra, DSM, DTU
27. Development of Efficient Antimicrobial Zinc Oxide Modified Montmorillonite Incorporated Polyacrylonitrile Nanofibers for Particulate Matter Filtration, **6.Priya Bansal** and **3.Roli Purwar**, Chemistry, DTU
28. DEVELOPMENT OF PREDICTIVE MODEL FOR SURFACE ROUGHNESS USING ARTIFICIAL NEURAL NETWORKS, **8.Nikhil Rai**, **3.M. S. Niranjan**, **8.Prateek Verma** and **7.Prince Tyagi**, Mechanical, DTU
29. Dielectric Modulated Junctionless Biotube FET (DM-JL-BT-FET) Bio-Sensor, Anubha Goel, **3.Sonam Rewari**, Seema Verma, S.S. Deswal and R.S. Gupta, Electronics, DTU

30. Effect of fly ash and graphite addition on the tribological behavior of aluminium composites, Vipin Kumar Sharma, *3.Ramesh Chandra Singh* and *3.Rajiv Chaudhary*, Mechanical, DTU
31. E-FUCA: enhancement in fuzzy unequal clustering and routing for sustainable wireless sensor network, *3.Pawan Singh Mehra*, Computer Science, DTU
32. Energetic and exergetic study of dual slope solar distiller coupled with evacuated tube collector under force mode, *6.Aseem Dubey*, *3.Samsher* and *3.Anil Kumar*, Mechanical, DTU
33. Enhancements in mechanical properties of dissimilar materials using friction stir welding (FSW) - A review, *3.R.S Mishra* and *6.Shivani Jha*, Mechanical, DTU
34. Evaluating Deep Neural Network Ensembles by Majority Voting cum Meta-Learning scheme, *8.Anmol Jain*, *8.Aishwary Kumar* and *3.Seba Susan*, IT, DTU
35. Evaluating the Effect of Process Parameters on FSP of Al5083 Alloy Using ANSYS, *Shourya Sahdev*, *6.Himanshu Kumar*, *3.Ravi Butola* and *3.Ranganath M. Singari*, Mechanical, DTU
36. Evaluation of Moth-Flame Optimization, Genetic and Simulated Annealing tuned PID controller for Steering Control of Autonomous Underwater Vehicle, *3.Sudarshan K. Valluru*, *8.Karan Sehgal* and *8.Hitesh Thareja*, Electrical, DTU
37. EVOLUTION IN MANUFACTURING OF GRID STIFFENED STRUCTURES THROUGH CAM AND ADDITIVE TECHNIQUES, *8.Kshitij Tripathi*, *8.Kunal Kukreja* and *3.Dr. AK Madan*, Mechanical, DTU
38. Expectation maximization clustering and sequential pattern mining based approach for detecting intrusions in transactions in databases, *3.Indu Singh* and *3.Rajni Jindal*, Computer Science, DTU
39. Fast Under Water Image Enhancement for Real Time Applications, *3.Aruna Bhat*, *8.Aadhar Tyagi*, *8.Aarsh Verdhan* and *8.Vaibhav Verma*, Computer Science and Software Engineering, DTU

40. Generation of COVID-19 Chest CT Scan Images using Generative Adversarial Networks, *8.Prerak Mann, 8.Sahaj Jain, 8.Saurabh Mittal* and *3.Aruna Bhat*, Computer Science, DTU
41. Handwriting Recognition for Medical Prescriptions using a CNN-Bi-LSTM Model, *8.Tavish Jain, 8.Rohan Sharma* and *3.Ruchika Malhotra*, Computer Science, DTU
42. Hindi-English Code Mixed Hate Speech Detection using Character Level Embeddings, *8.Rahul, 8.Vasu Gupta, 8.Vibhu Sehra* and *8.Yashaswi Raj Vardhan*, CSE, DTU
43. Hyper-Parameter analysis of Deep Auto Encoder for Flow Prediction, *8.Prakrit Tyagi, 8.Pranav Bahl* and *3.BB Arora*, Mechanical, DTU
44. Impact of multi threshold transistor in positive feedback source coupled logic (PFSCCL) fundamental cell, *6.Ranjana Sivaram*, Kirti Gupta and *3.Neeta Pandey*, Electronics, DTU
45. Investigation of machining performance in die-sinking electrical discharge machining of pentagonal micro-cavities using cylindrical electrode, *6.Shrikant Vidya, 3.Reeta Wattal* and P Venkateswara Rao, Mechanical, DTU
46. IP Traffic Classification of 4G Network using Machine Learning Techniques, *8.Rahul, 8.Amit Gupta, 8.Anupam Raj* and *8.Mayank Arora*, CSE, DTU
47. Justifying Biofield (Aura) Studies as Complementary and Alternative Medicine (Cam), *8.Ankit Dutta, 8.Subnear Kour* and *3.Dr. Priyanka Jain*, Electronics, DTU
48. Leakage reduction in dual mode logic through gated leakage transistors, *6.Neetika Yadav, 3.Neeta Pandey* and *3.Deva Nand*, Electronics, DTU
49. Mining Tourists' Opinions on Popular Indian Tourism Hotspots using Sentiment Analysis and Topic Modeling, *8.Shefali Singh, 8.Tureen Chauhan, 8.Vibhas Wahi* and *3.Priyanka Meel*, IT, DTU

50. Modeling and Analysis of High-Performance Triple Hole Block Layer Organic LED Based Light Sensor for Detection of Ovarian Cancer, Shubham Negi, **3.Poornima Mittal**, and Brijesh Kumar, ECE, DTU
51. MODELING FOR THE ENERGY POTENTIAL OF BIOGAS POWER PLANTS IN NATIONAL CAPITAL TERRITORY, **6.Rohit Agrawal** and **3.S.K. Singh**, Environmental, DTU
52. Multi Domain Fake News Analysis using Transfer Learning, **8.Pratyush Goel**, **8.Samarth Singhal**, **8.Snehil Aggarwal** and **3.Minni Jain**, CSE, DTU
53. Multimedia Data Summarization Using Joint Integer Linear Programming, **8.Sidhant Allawadi**, **8.Ritika**, **8.Vivek Rana** and **3.Minni Jain**, CSE, DTU
54. Multi-modal biometric recognition system based on FLSL fusion method and MDLNN classifier, **3.Ajai Kumar Gautam** and **3.Rajiv Kapoor**, Electronics, DTU
55. Novel Algorithm For Optimal PMU Placement For Wide Ranging Power System Observability, **6.Nitish Arora** and **3.S.T.Nagarajan**, Electrical, DTU
56. Optical Flow-Based Weighted Magnitude and Direction Histograms for the Detection of Abnormal Visual Events Using Combined Classifier, Gajendra Singh, **3.Rajiv Kapoor** and Arun Khosla, Electronics, DTU
57. Optimal design of FOPID Controller for the control of CSTR by using a novel hybrid metaheuristic algorithm, **6.NEHA KHANDUJA** and **3.BHARAT BHUSHAN**, Electrical, DTU
58. Performance analysis of solar driven combined recompression main compressor intercooling supercritical CO₂ cycle and organic Rankine cycle using low GWP fluids, **6.Yunis Khan** and **3.Radhey Shyam Mishra**, Mechanical, DTU
59. Physicochemical Studies on Interaction Behavior of Potato Starch Filled Low Density Polyethylene Grafted Maleic Anhydride and Low Density Polyethylene Biodegradable Composite Sheets, **3.A. P. Gupta**, Vijai Kumar, **6.Manjari Sharma** and **6.S. K. Shukla**, Applied Chemistry, DTU

60. Plasmon assisted tunnelling through silver nanodisk dimer optical properties and quantum effects, Venus Dillu, Preeti Rani, **3.Yogita Kalra** and **3.Ravindra Kumar Sinha**, Applied Physics, DTU
61. Point-of-Care PCR Assays for COVID-19 Detection, **6.Niharika Gupta**, **6.Shine Augustine**, Tarun Narayan, Alan O'Riordan, **3.Asmi Das**, D. Kumar, John H. T. Luong and **3.Bansi D. Malhotra**, Biotechnology, DTU
62. Prevalence and risk analysis of fluoride in groundwater around sandstone mine in Haryana, India, **6.Saurav Kumar Ambastha** and **3.A. K. Haritash**, Environmental, DTU
63. Realistic face generation using a textual description, **8.Anukriti Kumar**, **8.Anurag Mudgil**, **8.Nakul Dodeja** and **3.Dinesh Kumar Vishwakarma**, IT, DTU
64. Recall-based Machine Learning approach for early detection of Cervical Cancer, **6.Apoorva Gupta**, **8.Ashutosh Anand** and **3.Yasha Hasija**, Biotechnology, DTU
65. Recent advancements of PCM based indirect type solar drying systems: A state of art, Mukul Sharma, Deepali Atheaya and **3.Anil Kumar**, Mechanical and Environment, DTU
66. Review of electrospun hydrogel nanofiber system: Synthesis, Properties and Applications, **8.Tanushree Ghosh**, **8.Trisha Das** and **3.Roli Purwar**, Applied Chemistry, DTU
67. Review on Spray Powder Process of Additive Manufacturing, **8.Md Zia Arzoo**, **8.Mozammil Hassan** and **3.A.K.Madan**, Mechanical, DTU
68. Sensory Vision Substitution using Tactile Stimulation, **7.Pavitra Gandhi** and **3.Anamika Chauhan**, IT, DTU
69. Simulation and Comparative Study of Various Maximum Power Point Tracking Techniques, **7.Abhishek Singh**, **8.Sambhav Khatri** and **8.Sumit Kumar Gola**, DTU

70. Solving Community Detection in Social Networks: A comprehensive study, *8.Prashant Kumar, 8.Raghav Jain, 8.Shivam Chaudhary* and *3.Sanjay Kumar*, CSE, DTU
71. STOCK PRICE ESTIMATION BASED ON HISTORICAL INFORMATION AND TEXTUAL SENTIMENT ANALYSIS, *8.Abhay Gupta, 8.Aman Gupta, 8.Anshul Chaudhary* and *3.Rajesh Kumar Yadav*, CSE, DTU
72. Strongly coupled plasma effect on excitation energies of O-like ions and photoionization of F-like ions, R Sharma and *3.A Goyal*, Applied Physics, DTU
73. Testing normality in the time series of EMP indices: an application and power-comparison of alternative tests, Sanjay Kumara and *3.Nand Kumar*, Humanities, DTU
74. Therapeutic Targeting of Repurposed Anticancer Drugs in Alzheimer's Disease: Using the Multiomics Approach, *6.Dia Advani* and *3.Pravir Kumar*, Biotechnology, DTU
75. Thermal and Electrical Behaviour of the Persistent Current Switch for a Whole-Body Superconducting MRI Magnet, *6.Ajit Nandawadekar*, V. Soni, N. Suman, Sankar Ram T, R. Kumar, S.K. Saini, R G Sharma, *3.Mukhtiar Singh*, and Soumen Kar, Electrical, DTU
76. Time Efficient IOS Application For CardioVascular Disease Prediction Using Machine Learning, *8.Vansh Kedia, 8.Swesh Raj Regmi, 8.Khushi Jha, 8.Aman Bhatia, 8.Siddhant Dugar* and *8.Bickey Kumar Shah*, CSE and IT, DTU
77. Toxic Speech Detection using Traditional Machine Learning Models and BERT and fastText Embedding with Deep Neural Networks, *Pranav Malik, 8.Aditi Aggrawal* and *3.Dinesh K. Vishwakarma*, IT, DTU
78. Transfer Learning for Detection of COVID-19 Infection using Chest X-Ray Images, Nikhil Bhatia and *3.Geetanjali Bhola*, IT, DTU

79. Two-Stage Stochastic Programming Model for Optimal Scheduling of RES-Based Virtual Power Plants in Electricity Markets, **6.Meegada Indeevar Reddy**, **3.Radheshyam Saha** and **3.Sudarshan K. Valluru**, Electrical, DTU

80. VLSI implementation of transcendental function hyperbolic tangent for deep neural network accelerators, **7.1.Gunjan Rajput**, Gopal Raut, Mahesh Chandra and Santosh Kumar Vishvakarma, Electronics, DTU

81. WASTE MANAGEMENT BY “WASTE TO ENERGY” INITIATIVES IN INDIA, M. Kumar, **6.S. Kumar** and **3.S.K. Singh**, Environmental, 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 complete consumer behaviour learning model for real-time demand response implementation in smart grid

Swati Sharda¹ · Mukhtiar Singh² · Kapil Sharma¹

Accepted: 3 May 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Accurate and optimal implementation of Demand Response (DR) programs essentially requires knowledge of occupants' behavioral aspects regarding power usage. Maintaining consumers' comfort has become an imperative component along with cost reduction; there is utmost need to understand their power consumption trends completely. In this paper, a complete solution regarding consumer behavior learning has been presented for designing efficient demand response algorithms. Firstly, appliance-level power forecasting has been performed using deep learning ensemble model: CNN-LSTM and XG-boost; Secondly, dynamic itemset counting (DIC), a variant of the Apriori algorithm, has been utilized to generate association rules which determine appliance-appliance association and discovery. In this way, all the aspects of the dynamic and non-stationary nature of appliances' power time series have been addressed for DR implementation. Using two benchmark datasets, it has been demonstrated that the proposed approach exhibits enhanced performance in comparison to other competitive models in terms of RMSE and MAE.

Keywords Power forecasting · Deep learning · Ensemble model · Association mining

1 Introduction

In the smart grid era, demand response (DR) programs are considered an increasingly valuable resource option for the energy demand imbalance problem, which is always a concern for power grid operators [7]. Load shifting regulates load flow and reduces energy cost by rescheduling consumers' energy consumption patterns during peak hours in response to dynamic prices or financial incentives. In this regard, appliance level power forecasting can assist residential consumers in responding effectively to DR programs. The accurate load forecasts of individual consumers will determine flexibility in

demand and make them aware of their energy usage, allowing them to manage their usage costs better. Moreover, it can help utilities identify promising consumers for participation in DR programs in the power shortage scenario.

Until now, power forecasting has been performed at the house and submetering level. With the advent of IoT, fine-grained load data for domestic appliances are readily available, allowing predictions at the appliance level. Appliance-by-appliance consumption information will enable consumers to improve their energy efficiency. It also provides home energy automation systems to either directly control appliances or give the consumers recommendations about the period resulting in lower energy costs for the usage of appliances based on the learned user's behavioral habits from historical data. Hence, it determines appliances' flexibility for participating in DR programs. It showed that appliance-level energy usage information could help residents save up to 12% in energy costs instead of receiving conventional monthly details at the whole building level [6].

Brown et al. [1] incorporated individual energy profiles to implement automated energy management based on consumers' occupancy and behavior. Since appliance-level energy requirements depend upon the number of residents, their occupation, action, outside weather, location, etc., determining a single algorithm that forecasts appliance power while

✉ Swati Sharda
swatisharda2807@gmail.com

Mukhtiar Singh
mukhtiarsingh@dce.ac.in

Kapil Sharma
kapilsharma@ieee.org

¹ Department of Information Technology, Delhi Technological University, Delhi, India

² Department of Electrical Engineering, Delhi Technological University, Delhi, India

capturing different consumers' behavioral patterns is challenging. However, the approaches and discussions on this subject are still in the primitive stage and not mature enough because of the high volatility of the residential load profiles.

Most of the previous studies are based on short-term load forecasting at the building level, or aggregate level [5, 20]. Recently, different approaches have been suggested for the accurate load forecasting of individual residential customers.

Artificial Intelligence (AI) techniques support demand-side flexibility (DSF), which helps consumers to play an active role in demand shifting programs [4]. These days, deep learning has become one of the most popular techniques for time-series forecasting [29]. Unlike shallow learning, deep learning typically involves stacking multiple layers of the neural network and relying on stochastic optimization to solve complex problems [26]. The Long Short-Term Memory network (LSTM) and Gated Recurrent Unit network (GRU) are usually more potent than traditional RNN as reported in some load forecasting tasks [15].

The LSTM deep learning model is used for short-term load forecasting at individual customer levels in the smart grid [14, 24]. Simple backpropagation neural network has been utilized compared to LSTM for short-term household power prediction [14]. A pooling based LSTM strategy is used in [24] which utilizes multiple household load profile data from smart meters for forecasting purposes. A combination of convolutional neural network (CNN) and bi-directional LSTM (Bi-LSTM) has been utilized for household electric energy consumption prediction (EECP) [17, 27]. A hybrid CNN-GRU model to predict short-term electricity consumption in residential buildings by learning both spatial and temporal features of multi-variate time-series [21]. A concept of transfer-learning and a cluster-based strategy has been utilized for training an electricity forecasting model based on LSTM [16]. However, these models are not suitable for real-time implementation. A two-dimensional (2D) CNN using recurrence plots has been implemented for load forecasting of individual residential customers [23]. However, this technique works specifically for time-series that repeat their states.

Many researchers focus on probabilistic forecasting models for short-term and household level power forecasting. A probabilistic density short-term power forecasting model based on deep learning and quantile regression has been proposed in [10]. However, the multi-layer perceptron (MLP) based deep learning technique is not suitable for large and complex data sets. A probabilistic household level load forecasting has been reported using LSTM deep neural network [28] and hidden Markov model [12]. Most of the forecasting approaches used in these literary works focus on individual household-level smart meter data. In contrast, the power forecasting at the appliance level in real-time is much more sparse.

It is widely acknowledged that an ensemble of multiple deep learning models can boost prediction efficiency and

has higher generalization skills than individual models [8, 2]. The goal of combining multiple models is to obtain a more accurate estimate than the one obtained by a single model as the errors in aggregating diverse model predictions can be easily compensated. Various successful methods have been put forward to boost load prediction accuracy by integrating several models.

It is well known that a boosting-based ensemble is more effective in handling the time-series data set based on long-range dependencies. An ensemble method for short-term load forecasts based on the hybrid LGBM-XGB-MLP model has been proposed in [18]. However, the extreme learning machine (ELM) based architecture is a two-layer neural network that cannot handle the long-term dependencies and volatility of the appliance's power series. A hybrid deep neural network with a fuzzy clustering approach has been utilized in [25] for hourly load forecasting. Here, the fuzzy approach is used to cluster data into multiple subsets, further taken as input to the deep neural network model.

The CNN-LSTM model has been used in [25] for predicting energy consumption at the residential level. Results indicate that CNN-LSTM performs better than LSTM for individual household load forecasting. A deep ensemble model for probabilistic load forecasting has been developed for individual customers [26, 27]. The quantile strategy combined with the LASSO technique has been utilized [26], whereas the deep residual network (ResNet) for day-ahead forecasting has been reported in [27]. However, the neural network-based deep learning model is still a better choice. But probabilistic forecast being non-linear and non-convex, is not suitable for real-time DR programs.

The proposed work will significantly help in the development of a scheduling optimization algorithm for real-time demand response. The consumer behavioral aspects can be determined only by their appliances' power usage pattern and related association. The scheduling algorithm using load shifting shifts controllable appliances to later intervals for electricity cost minimization without violating consumer comfort. Learning consumer behavior helps maintain his comfort, which means determining the earliest start time and finish time of various appliances and their power requirement at different time slots.

The association mining further enhances the behavioral learning by providing information to the algorithm about which appliance should preferably run after the currently running appliance (in case many appliances are ready to run). Association mining helps to provide supervised information to the scheduling algorithm to preferably run the next appliance associated with the currently running appliance. In this way, the proposed work is essential for developing consumer-oriented scheduling appliances for demand response implementation.

Since the data set of household appliance power consumption is both noisy and real, and no individual forecasting

model may be generalized for all consumers. The energy forecasting domain demands more robustness, higher prediction accuracy, and generalization ability for real-world implementation. Therefore, an ensemble model combining RNNs and gradient boosting tree capabilities for superior prediction performance has been developed and applied for appliance-level forecasting. Hybrid Convolutional LSTM (CNN-LSTM) deep learning models are used as base learners for the XG-Boost algorithm.

The main innovations and contributions of this paper include:

1. Development of a multi-stage ensemble deep learning model with powerful learning ability for appliance level power forecasting.
2. A 5-minute prediction time horizon has been considered for the appliance level power forecasting, which is more suitable for real-time demand response programs.
3. Appliance-Appliance association mining using Dynamic itemset counting (DIC) algorithm to determine which appliances frequently occurred together.
4. The proposed model's performance has been rigorously evaluated on publicly available datasets, namely GREEND and UK-DALE.

The rest of this paper contains the following discussions: our proposed solution's methodology is introduced in Section II. The training and testing phases of the deep learning ensemble forecasting model are discussed in this section. The experiment details are discussed in Section III. The numerical results have been presented in Section IV. Finally, the paper has been concluded with some short remarks in Section V.

2 Forecasting model architecture

A deep learning-based ensemble model has been developed to capture appliances' stochastic power usage at 5 min intervals. The multi-stage boosting ensemble technique elevates the base model's predictive strength by covering large data space and minimizing the error than those obtained by individual models.

The hybrid deep learning Convolutional LSTM (CNN-LSTM) has been utilized as the first stage of our Ensemble model. The three CNN-LSTM model outputs are stacked together and fed to boosting stage for the final forecasted value. XGBoost, namely eXtreme Gradient Boosting, combines trees in a boosting manner and currently provides state-of-the-art performance amongst several prediction challenges. XGBoost allows parallel programming without significant loss of accuracy.

The Ensemble model prediction step can be divided into two phases: 1) Model Training and 2) Testing. The training

and testing phases have been illustrated in Fig. 1 and described as follows:

2.1 Training phase

In this phase, the ensemble model is initialized with random weights, and these weights get updated with each training cycle.

Input Data - Training data is divided into batches with a sequence length (k) of 128 samples. These batches ($n = 128$) are fed as input to the convolution-1D layer. The sequence of training examples can be represented as $(x_1, y_1), (x_2, y_2), \dots, (x_{128}, y_{128})$ with $x_t \in \mathbb{R}^{n \times k}$ and $y_t \in \mathbb{R}^n$ for $1 \leq t \leq 128$. x_t denotes input for univariate time-series and y_t denotes output.

Convolution layers - These layers can learn from the raw time-series data directly without scaling or differencing and deriving interesting features from the shorter (fixed-length) sequence of the total time-series dataset. Two layers of Conv-1D have been utilized to give the model a fair chance of learning features from the long noisy input data. The first layer with 64 parallel feature maps and a kernel size of 3 takes the input shape 128×1 and produces the output shape of 126×64 . This layer is used to learn basic features. The second Conv-1D layer with same configuration has been utilized to learn more complex features. The output shape of this layer is 124×64 . The output of the convolution layer can be expressed as [30]:

$$C_r(t) = f\left(\sum_{i=1}^l \sum_{j=1}^k x(i + s(t-1), j) \omega_r(i, j) + b(r)\right) \quad (1)$$

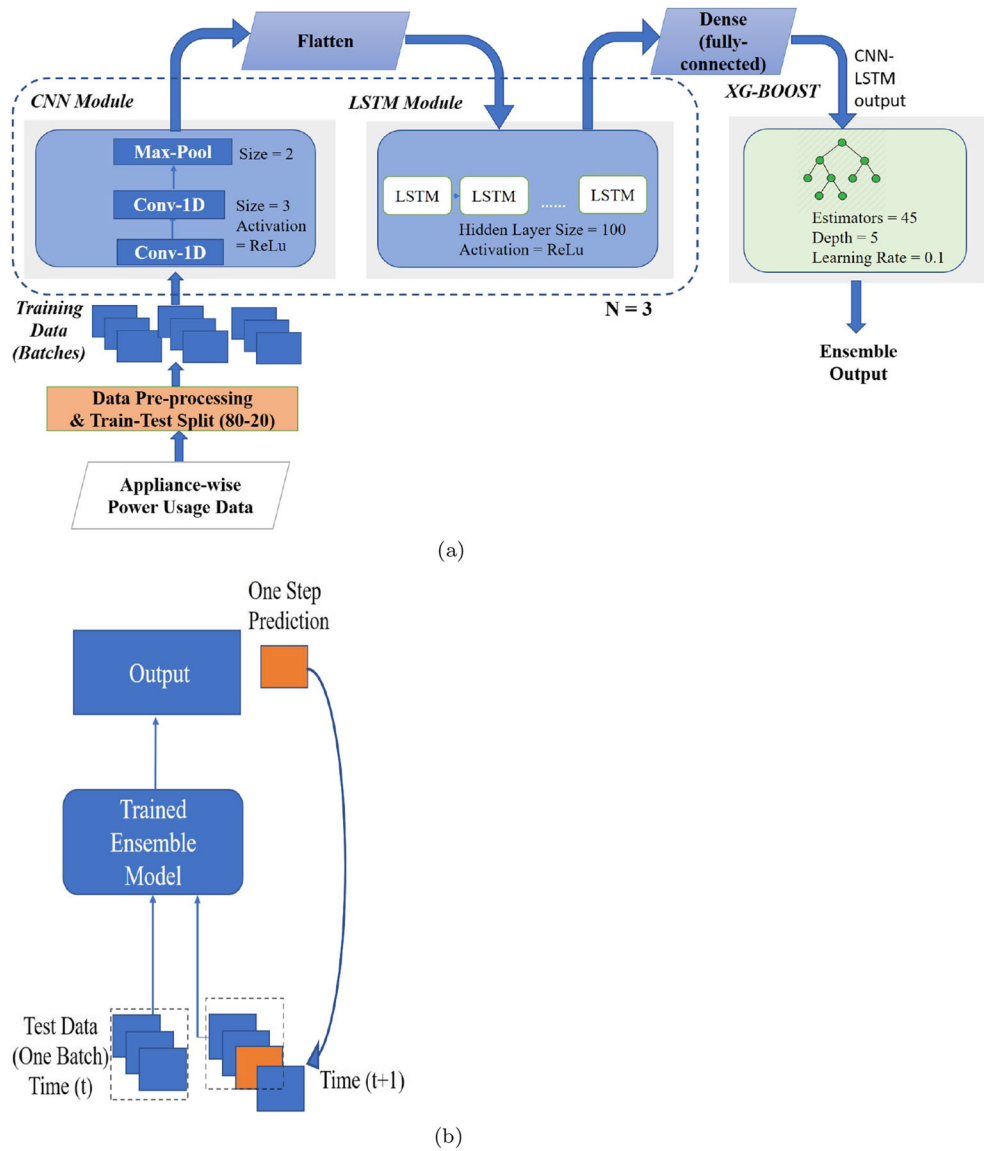
where $x \in \mathbb{R}^{n \times k}$ denotes the input time series or the output of preceding layer, s denotes the convolution stride, $C_r(t)$ refers to the t^{th} component of r^{th} feature map, $\omega_r \in \mathbb{R}^{l \times k}$ and $b(r)$ refers to the weights and bias of the r^{th} convolution filter. This filter connects the j^{th} feature map of layer $l-1$ with i^{th} feature map of layer l .

MaxPool Layer - The pooling layer reduces the learned characteristics to 1/4 of their size and consolidates them into the critical elements. It prevents the overfitting of learned features by taking the maximum value within the window region. With pool size set to 2, the output shape from this layer is 62×64 . This layer output can be expressed as:

$$P_r(t) = \max(C_r((t-1)l+1), C_r((t-1)l+2), \dots, C_r(tl)) \quad (2)$$

Flatten Layer - It reduces the feature maps into a single one-dimensional vector. The output shape after this layer is a vector containing 3968 (62×64) values. This layer is followed by repeat vector layer which converts current 2D vector of shape (none, 3968) into 3D vector with shape (none, 1, 3968) to make it suitable to input to next LSTM layer.

Fig. 1 Ensemble forecasting model architecture. **a** Training, **b** Testing



LSTM Layer - The LSTM gating architecture is computationally efficient than traditional RNNs. The selective read, write and forget procedure followed in LSTM avoids explosive and vanishing gradient problems [11].

At each timestep t , for input x_t , each memory cell c_t is updated and a hidden state h_t is generated as output according to the following equations [9]:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}x_{t-1} + b_i), \\ f_t &= \sigma(W_{xf}x_t + W_{hf}x_{t-1} + b_f), \\ o_t &= \sigma(W_{xo}x_t + W_{ho}x_{t-1} + b_o), \\ c_t &= f_t \oplus c_{t-1} + i_t \oplus \phi(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \\ h_t &= o_t \oplus \phi c_t, \end{aligned} \quad (3)$$

In the proposed model, one LSTM layer with 100 neurons has been utilized. The output vector from this layer is of shape (none, 1, 200).

Dense Layer - It is a fully connected layer used to reduce the vector's size. The proposed model uses two dense layers. The first dense layer reduces the vector size of (none, 1, 200) to (none, 1, 100). The second dense layer is the output layer, producing a single output of the CNN-LSTM forecast with output shape (none, 1, 1).

XGBoost - XGBoost, namely eXtreme Gradient Boosting, is an integrated learning method that uses decision trees and random forests to make predictions. It uses boosted decision trees to obtain final predictions using base learners. However, a gradient decent algorithm is used to reduce the errors effectively. If $F = \{F_1, F_2, F_3, \dots, F_m\}$ are the set of base learners. The final prediction can be given by:

$$\hat{y}_i = \sum_{t=1}^m F_t(x_i) \quad (4)$$

where $\{x_1, x_2, x_3, \dots, x_m\}$ are data points. The loss function of XGBoost on t^{th} iteration [3]:

$$L^t = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \lambda f_t \quad (5)$$

where L^t denotes the loss at t^{th} iteration. l is based on the loss function of the former $t - 1$ tree, y_i is the label of x_i , λ denotes regularization parameter, f_t represents t^{th} tree output, $l(y_i, \hat{y}_i)$ is the training loss of x_i , $\hat{y}_i^{(t-1)}$ represents the prediction of the combination of all the tree models.

The second-order expansion of Taylor is performed on the above equation to obtain the final loss and may be represented as:

$$LOSS = -\frac{1}{2} \sum_{j=1}^T w_j + \gamma T \quad (6)$$

where w_j is the prediction for node j which can be expressed by following equation:

$$w_j = -\frac{G_i}{H_i + \lambda} \quad (7)$$

where G_i is $\sum_{i \in I_j} g_i$ and H_i is $\sum_{i \in I_j} h_i$. Here, g_i and h_i are the first order and second order derivative loss of predictions at previous iterations, respectively. Also, I_j denotes the set of instances belonging to node j . The smaller value of $LOSS$ denotes the better structure of tree.

For XGBoost module, the proposed ensemble model utilizes 45 estimators with learning rate set to be 0.1. The depth of tree is taken as 5. The fraction of columns to be randomly sampled for each tree, denoted by parameter *col_sample* by tree is set as 0.3. The objective function of regressor is set to be linear.

In the proposed ensemble forecasting method, the hybrid structure of CNN-LSTM handles the dynamics and non-stationarities of real-world time series accurately. The output of three discrete CNN-LSTM models is stacked together and fed to the XGBoost model. In this way, the prediction performance gets further boosted, and better prediction results have been achieved.

For m number of CNN-LSTM models in an ensemble, the forecast result for time series with n observations, (y^1, y^2, \dots, y^n) , is given by

$$\hat{y}^t = \sum_{i=1}^m w_m \hat{y}_m^t \quad \text{for } t = 1, \dots, n. \quad (8)$$

where \hat{y}_m^t denotes the forecast output (at the t^{th} time stamp) obtained using the m^{th} CNN-LSTM model and w_m is the associated weight. Each weight w_m is assigned to a corresponding CNN-LSTM models's forecast output. Also, $0 \leq w_m \leq 1$ and $\sum_{i=1}^m w_m = 1$.

2.2 Testing phase

Testing samples in the batches of sequence length 128 are fed to a trained ensemble model to predict power at timestep t . To provide a robust estimation of modeling parameters, walk-forward validation has been performed. This methodology involves moving along the time series one step by applying a moving window to available time-series data. The forecasted value of the trained ensemble model is evaluated against the actual value. The next time step $t + 1$ includes this actual expected value from the test set for the forecast on the next time step $t + 2$ by further moving the window next step. The procedure is repeated until the end of test data is reached. The testing process is illustrated in Fig. 1(b).

2.3 Appliances' association rules extraction

The frequently associated appliances are extracted using dynamic itemset counting (DIC), a variant of the Apriori algorithm. This algorithm incorporates the dynamic change (addition and deletion) of appliances power us-age in the database. It means it can incorporate the changing behavioral aspect of occupants well. With this approach, a small portion of the database is mined at each iteration, which reduces the memory overhead and improves efficiency compared to Apriori.

This algorithm uses a support-confidence framework to extract association rules and generating frequent itemsets of appliances' i.e. the set of appliances that often run together. The correlation rule can be expressed as-

$$X \Rightarrow Y [\text{Support}, \text{Confidence}, \text{Correlation}] \quad (9)$$

where the correlation can be measured using Lift metric that provides more insight into support-confidence relationship. where

$$\text{Lift}(X, Y) = \frac{\text{Confidence}(X \Rightarrow Y)}{\text{Support}(Y)} \quad (10)$$

3 Experiment details

3.1 Data set

The experimental study has been carried out using two popular open-access data sets for evaluating the performance of the proposed ensemble model, namely, GREEND and UK-Dale. The GREEND dataset contains appliance-wise data of 8 different houses in Italy and Austria. We utilize the data of house number 2, which is an apartment with one floor in Klagenfurt (AT). The residents are a young couple, spending most of the daylight time at work during weekdays, mostly being home in the evenings and weekends. The data collection module's

plugs kit consists of a Zigbee network having nine sensing outlets, each collecting active power measurements from the connected load.

The UK-DALE (UK Domestic Appliance Level Electricity) data set records the power consumption of five UK houses, appliance-wise and the whole house as well. We have used 8 appliances of house number 1 for our experiment. The detailed description of both data sets is described in Table 1.

3.2 Data preprocessing

The GREEND data set's null values are replaced with the most frequent power value for each appliance. Then, the 1-second data is resampled to 5 minutes by taking the average of 300 samples for each appliance. Similarly, for the UK-DALE dataset, the 6-second data samples per appliance are resampled into 5 minutes by taking an average of 50 samples. The duration of 5 minutes is chosen as it is appropriate for load shifting under a real-time environment using real-time pricing (RTP) schemes. Also, the chosen horizon incorporates the minimum operating duration of smart household appliances. Then, the resampled data is divided into training and testing as an 80 to 20 ratio. Out of 80% training data, again, 20% is used for validation purposes and select appropriate hyperparameters. Table 1 contains the details of the number of data samples in training, testing, and validation for both data sets. The training data is divided into batches of sequence length 128 to be used as input to the ensemble model. Similarly, testing data is also divided into batches with a sequence length of 128 to predict power output from the trained ensemble model.

Table 1 Data sets information

	GREEND [19]	UK-DALE [13]
Location	Austria, Italy	UK
Duration	1 year and 4 months	5 year and 5 months
House	2	1
Resolution	1 Hz	6 sec
Total Samples (5 min)	1,34,933	4,68,836
Training Samples (5 min)	1,00,481	2,81,327
Validation Samples (5 min)	8,612	37,501
Testing Samples (5 min)	34,452	1,87,509
Training Set	15-02-2014 to 21-03-2015	09-11-2012 to 15-07-2015
Testing Set	21-03-2015 to 29-06-2015	15-07-2015 to 26-04-2017

3.3 Performance criteria

Two standard evaluation metrics measure the ensemble model's forecasting performance: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). These are described as follows:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N e_i^2} \quad (11)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |e_i| \quad (12)$$

where

$$e_i = Power_{forecast,i} - Power_{actual,i}$$

is known as forecast error. $Power_{forecast,i}$ is the forecasted power of i^{th} sample and $Power_{actual,i}$ is the actual power of i^{th} sample. N is the number of samples used for measuring accuracy.

3.4 Reference forecast methods and model tuning

1) **Feed-forward neural network (FFNN)** This is the basic form of the neural network used for regression problems. It uses two hidden layers with 64 neurons each and the ReLU activation function. The sequence length is set to be 128. Mean Square Error (MSE) is used as a loss function where RMSprop optimizer has been incorporated. The model has been run for 20 epochs.

2) **Long-short term memory (LSTM)** - This belongs to the family of recurrent neural networks (RNNs) that has recently gained attention for time series forecasting. The architecture of LSTM consists of 1 input layer with ten hidden neurons used with the ReLU activation function. This layer is followed by one dense layer. Adam optimizer has been utilized for updating the weights and reducing errors in the model. The model has been run for 20 epochs with batch size and sequence length of 128 samples.

3) **Convolutional-LSTM (CNN-LSTM)** - is a hybrid model combining convolutional and LSTM model. The architecture of the CNN-LSTM algorithm consists of two convolutional 1D (Conv1D) layers with kernel size 3 and 64 filters. This layer is followed by the Maxpool layer with pool size 2 and LSTM layer with 200 hidden neurons. The output layer with a linear activation function consists of one output neuron. The model is trained batch-wise with a sequence length of 128, maximum epochs 20 and learning rate 0.01 with Adam optimizer.

4) **Convolutional-XGBoost (CNN-XGBoost)** - It is a multi-stage ensemble model having three CNN models

and XGBoost. The outputs from all CNN models are stacked together and fed to the XGBoost regressor. Each CNN model has two convolutional layers having 64 filters, kernel-size is 3 with ReLu activation. Two dense layers at the end to change the output size to 512 and 1, respectively. Then, the outputs are stacked, and the XG-Boost regressor boosts the output to generate the final power prediction. The configuration of XG-Boost has been taken the same as the proposed model for a fair comparison.

All the reference forecast methods have been tuned to the best possible hyperparameters for one step ahead forecasts. The grid search method has been utilized to tune the hyperparameters of the proposed model. The table corresponding to hyperparameter tuning is presented in appendix 8. These deep learning models have been trained with TITAN RTX GPU using Python 3.6 with Keras 2.2.4 library on a computer system with IntelCore-i7 CPU.

4 Results and analysis

The results in Tables 2 and 3 demonstrate the ensemble deep learning model's effectiveness in improving forecasting performance (in terms of both RMSE and MAE) against all reference models on the GREEND dataset. We can categorize GREEND appliances as fixed appliances (microwave, water kettle, radio, dryer, kitchenware, and bedside light), controllable appliances (dishwasher and washing machine), thermostatically controllable (TCL) (Fridge). For fixed appliances, the average RMSE of FFNN, LSTM, CNN-LSTM, CNN-XGBoost, and the proposed model is 15.93, 15.28, 14.03, 12.41, and 9.032, respectively. For controllable appliances, on average, the RMSE of FFNN, LSTM, CNN-LSTM, CNN-XGBoost, and Ours is 30.085, 48.015, 35.65, 33.13,

Table 2 Comparison of the proposed Ensemble model with reference forecasting models on GREEND data set with respect to RMSE (W/m^2)

	FFNN	LSTM	CNN-LSTM	CNN-XG	OURS
Fridge	56.62	66.88	47.36	45.78	43.49
Dishwasher	6.23	7.45	6.22	5.95	4.15
Microwave	17.16	16.67	15.07	14.86	12.98
Water-kettle	61.68	59.76	56.27	48.76	35.55
Washing Machine	53.94	88.58	65.08	60.32	51.62
Radio	4.31	6.83	4.25	3.91	2.61
Dryer	5.56	1.16	1.14	1.11	1.09
Kitchenware	4.90	5.32	3.75	2.75	0.77
Bedside light	1.98	1.94	3.73	3.11	1.19

Table 3 Comparison of the proposed Ensemble model with reference forecasting models on GREEND data set with respect to MAE (W/m^2)

	FFNN	LSTM	CNN-LSTM	CNN-XG	OURS
Fridge	24.79	51.06	28.75	26.32	22.33
Dishwasher	3.18	3.66	4.95	3.06	2.04
Microwave	8.09	7.23	4.89	3.98	2.86
Water-kettle	22.00	29.68	11.67	11.02	10.72
Washing Machine	14.35	23.93	15.94	13.56	11.97
Radio	0.96	2.19	1.33	0.93	0.66
Dryer	1.33	0.83	0.82	0.73	0.60
Kitchenware	3.23	3.10	2.76	1.85	0.62
Bedside light	0.16	0.19	0.15	0.14	0.13

and 27.885, respectively. There is only one TCL appliance in GREEND whose RMSE and MAE can be seen from Tables 2 and 3, respectively.

For the UK-Dale dataset, all the appliances come under the fixed category except boiler which is a TCL appliance. Tables 4 and 5 show the outstanding performance of ensemble model over all other reference models. Here, the average RMSE of fixed appliances of FFNN, LSTM, CNN-LSTM, CNN-XGBoost, and Ours is 5.34, 4.45, 4.42, 3.80, and 2.63, respectively. The average MAE of fixed appliances of FFNN, LSTM, CNN-LSTM, CNN-XGBoost, and Ours is 1.91, 2.44, 1.50, 1.28, and 0.82, respectively.

For visualization, the prediction results of the Ensemble model and other comparative models on all GREEND appliances are shown in Fig. 2. Similarly, the prediction results of all appliances in the UK-Dale dataset are shown in Fig. 3. On analyzing these results, the ensemble model generally fits the actual data is much better than other comparable models, which validates the fact that the proposed ensemble model has better prediction performance.

Results indicate that the multi-stage CNN-LSTM XGBoost ensemble performs slightly better on the UK-Dale data set

Table 4 Comparison of the proposed Ensemble model with reference forecasting models on UK-Dale data set with respect to RMSE (W/m^2)

	FFNN	LSTM	CNN-LSTM	CNN-XG	OURS
Boiler	35.19	48.80	32.24	31.06	29.69
Thermal pump	0.78	1.43	0.86	0.80	0.73
Laptop	3.29	2.64	0.96	0.90	0.88
TV	4.50	5.67	4.41	3.79	2.88
LED Lamp	0.42	0.77	0.43	0.38	0.34
Kitchen Light	13.63	13.48	14.18	12.65	10.39
Kettle	6.10	6.21	6.12	4.61	0.55
Toaster	8.70	15.55	4.00	3.53	2.69

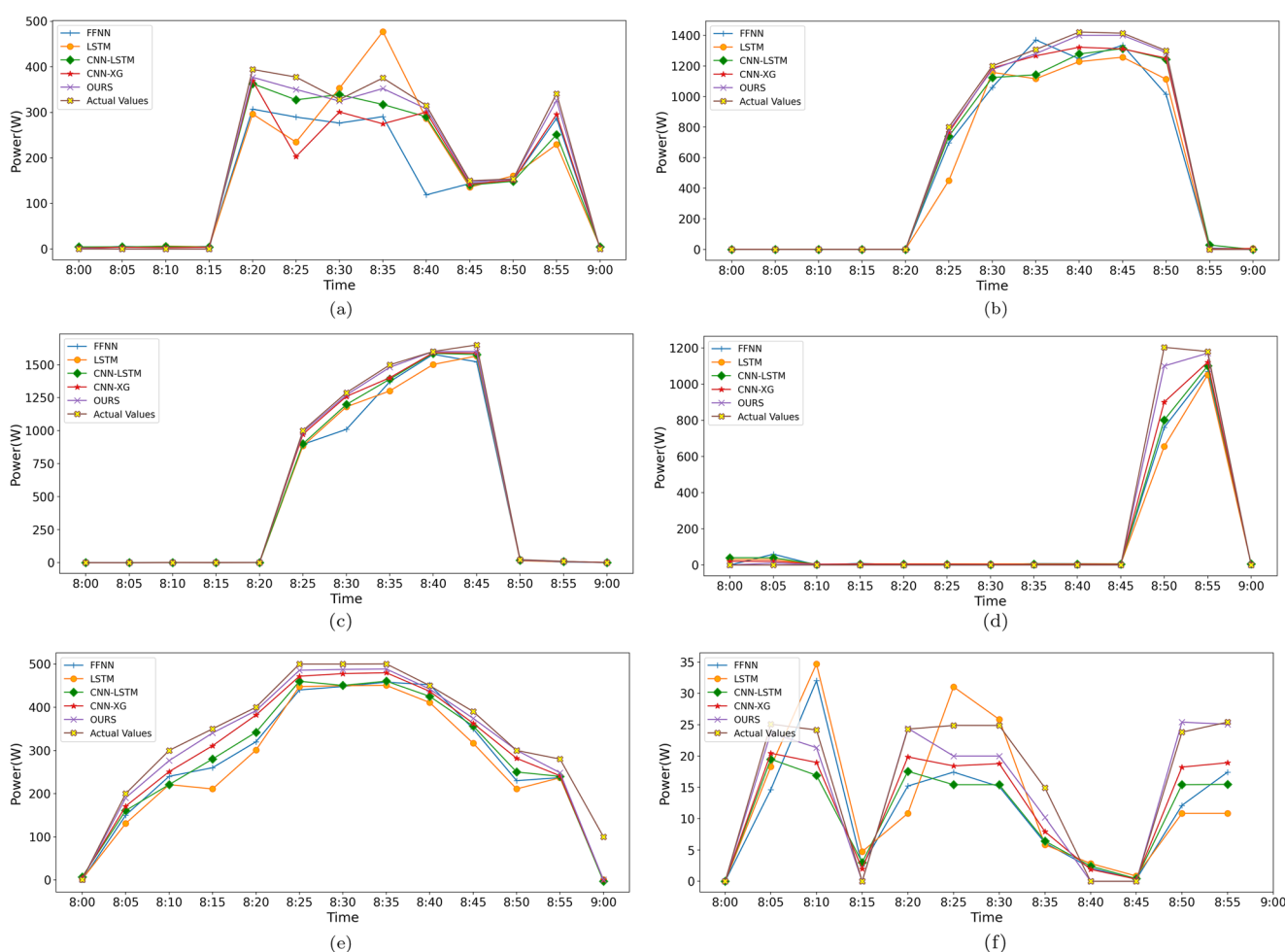
Table 5 Comparison of the proposed Ensemble model with reference forecasting models on UK-Dale data set with respect to MAE (W/m^2)

	FFNN	LSTM	CNN-LSTM	CNN-XG	OURS
Boiler	14.62	21.54	12.26	12.00	10.94
Thermal pump	0.29	0.70	0.49	0.38	0.25
Laptop	1.73	0.45	1.39	1.00	0.40
TV	2.12	2.19	2.20	2.04	1.57
LED Lamp	0.21	0.37	0.13	0.11	0.01
Kitchen Light	3.04	3.36	3.06	2.69	1.86
Kettle	0.39	0.39	0.87	0.74	0.35
Toaster	5.64	9.67	2.37	2.05	1.35

than the GREEND dataset in terms of both RMSE and MAE. It is due to periodicity observed in appliance usage in the UK-Dale data set. For the GREEND data set, in terms of RMSE, the proposed model performance for controllable appliances is 7.5%, 53.05%, 24.46%, 17.21% better than FFNN, LSTM,

CNN-LSTM, and CNN-XGBoost, respectively. Similarly, for fixed appliances, the proposed model is 55.28%, 51.41%, 43.36%, 31.52% superior to FFNN, LSTM, CNN-LSTM, CNN-XGBoost, respectively. For the UK-Dale dataset, on fixed appliances, the proposed model performance beats FFNN, LSTM, CNN-LSTM, and CNN-XGBoost by 68%, 51.41%, 50.78%, 36.39%, respectively. The TCL appliance's proposed model shows a lesser RMSE of 29.69 on UK-Dale than 43.49 on GREEND. The working code of the proposed work can be found here [22].

The proposed model performs better than all other comparative models because the combination of CNN and LSTM allows the LSTM layer to extract patterns and sequential dependencies in the time-series. In contrast, the CNN layer, through dilated convolutions methods and filters, further improves this process. This approach mainly helps in granular level forecasting (5 min in our case). The benefit of this model is that the model can support very long input sequences that can be read as blocks or subsequences by the CNN model, then pieced together by the LSTM model. Further, the

**Fig. 2** Appliance-wise power forecasting using Ensemble model on GREEND appliances. **a** Fridge, **b** Dishwasher, **c** Microwave, **d** Water-kettle, **e** Washing machine, **f** Radio

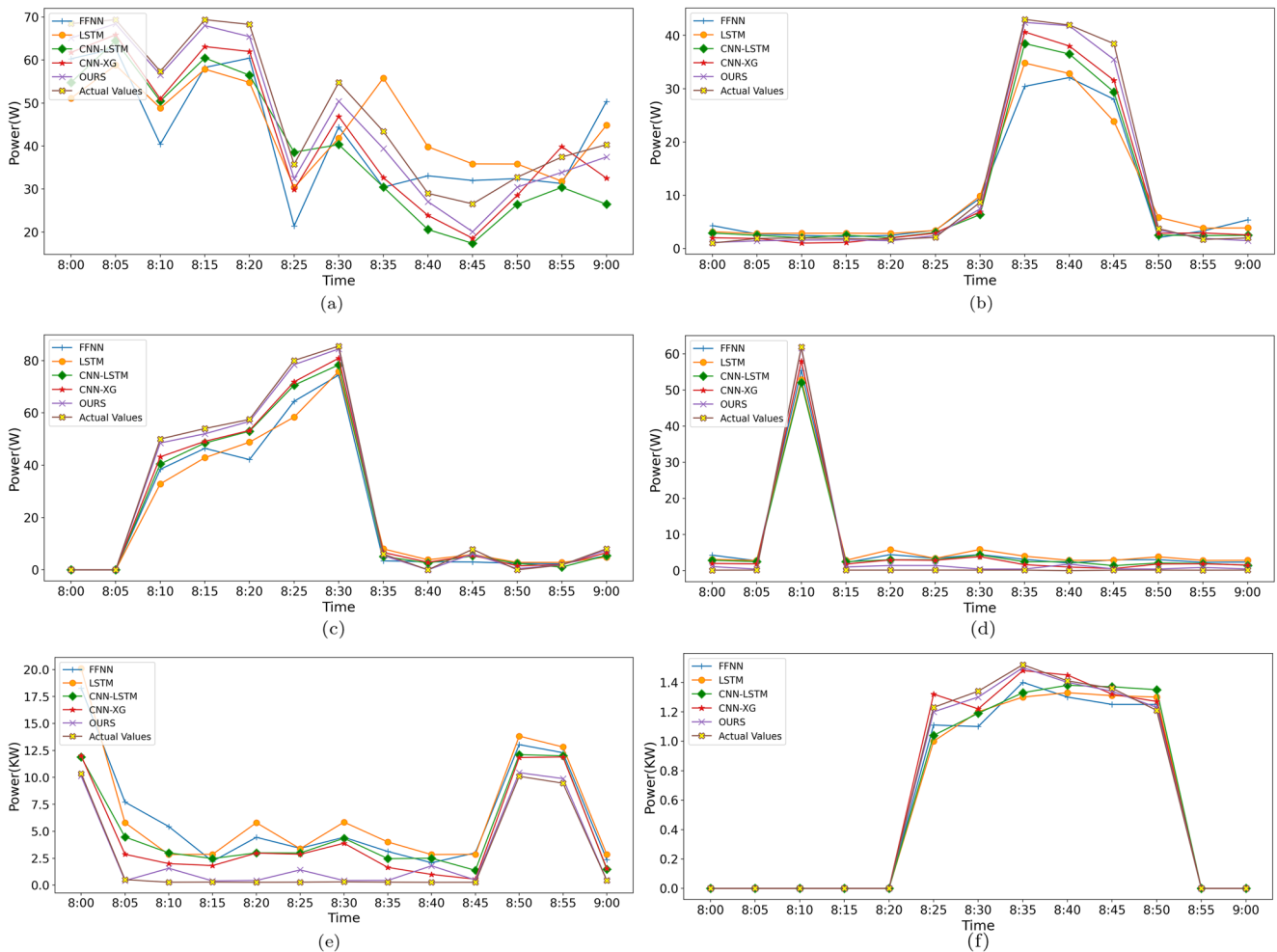


Fig. 3 Appliance-wise power forecasting using Ensemble model on UK-Dale appliances. **a** Boiler, **b** Solar pump, **c** Laptop, **d** Kitchen light, **e** Kettle, **f** Toaster

performance has been enhanced by using the XG-Boost tree, which boosts the performance of its base models by providing high preference to poorly estimated samples over well-estimated samples.

4.1 Appliances association analysis

For GREEND dataset, the strong association rules are exhibit by four appliances: radio, bedside light, dishwasher, and

Table 6 Association rules on GREEND equipments

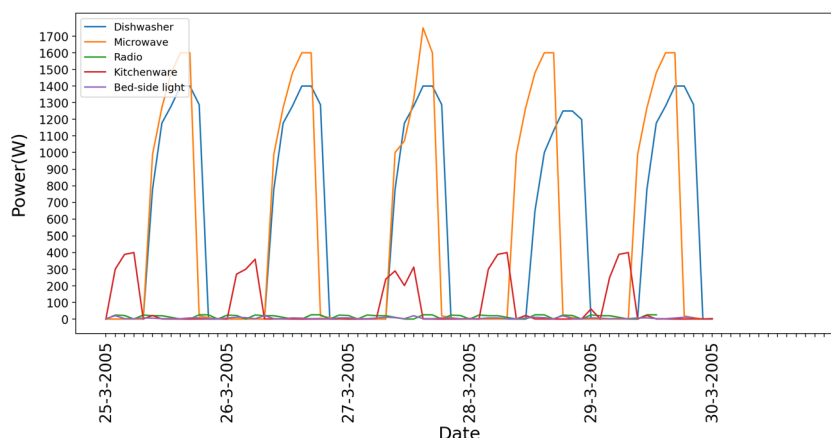
S.no.	Association Rule	Support	Confidence	Lift
1.	Radio Bedside light	0.16	0.89	1.2
2.	Dishwasher Microwave	0.21	0.93	1.5
3.	Microwave Dishwasher	0.24	0.92	1.4
4.	Bedside light Radio	0.12	0.80	1.3

microwave. Table 6 shows the association rules, with support, confidence and lift parameters with $min_sup \geq 0.2$. Further, the energy consumption curves of these appliances as represented in Fig. 4, compliments these association rules discovered and proves their simultaneous usage by the consumer. Similarly, the associations rules are generated for UK-Dale appliances as well, as presented in Table 7. These associations results determine occupants' behavioral traits. For example, a radio is used at the bedside light, depicting the occupant listens to the radio with the bedside light switched on. Moreover, for other occupants of the UK-Dale house, there is a strong association found in the usage of kettle, toaster, and kitchen light. It means the occupant likes to use a kettle and toaster while in the kitchen.

4.2 Training time analysis

In terms of training time, the FFNN model takes 7 seconds per epoch, the LSTM model takes 600 seconds per epoch, CNN-

Fig. 4 Associations represented by Energy curves



LSTM takes 401 seconds per epoch, CNN-XGBoost takes 700 seconds per epoch, and the proposed model takes 800 seconds per epoch. Its better performance can compensate for the more significant training time of the proposed model. The 800 seconds per epoch are taken during training the model. Once the model is trained, it gives comparable performance to other reference models for real-time prediction.

5 Conclusion

Appliance level power prediction is quite challenging due to the volatility and behavioral habits of individual consumers. An Ensemble deep learning model has been developed to capture the stochastic dynamics of appliances' power time series data. It utilizes two powerful algorithms' inherent advantages: a deep learning-based CNN-LSTM and tree-based Xtreme Gradient Boosting (XG-Boost) for performance enhancement. The prediction is carried out at a 5-minute time-horizon which is best suited for load shifting under real-time pricing schemes. Moreover, the dynamic itemset counting (DIC) algorithm has been utilized for determining which appliances are often used together. Rigorous experimental analysis on two open-source data sets (GREEND and UK-Dale) verifies the Ensemble model's outstanding performance in terms of RMSE and MAE accuracy metrics. The percentage decrease in RMSE of the proposed ensemble model on

GREEND data set is 32.18%, 49.54%, 27.73%, and 19.43% in compared to FFNN, LSTM, CNN-LSTM, and CNN-XGBoost, respectively. For the UK-Dale data set, the RMSE of the proposed Ensemble model is 40.58%, 65.09%, 27.17%, and 18.15%, lesser than FFNN, LSTM, CNN-LSTM, and CNN-XGBoost, respectively.

Appendix

Table 8 Optimal hyperparameters of proposed model

Parameters	Values	Search Range
Training Steps	50	{40, 50, 100}
Kernel size	3	{2, 3, 4}
Pool Size	2	{2, 3, 4}
LSTM layer size	100	{50, 100, 120}
Learning Rate	0.01	{0.0001, 0.001, 0.01 }
Estimators Size	45	{30, 45, 55 }
Tree depth	5	{3, 5, 6}

Table 7 Association rules on UK_DALE equipments

S.no.	Association Rule	Support	Confidence	Lift
1.	Kettle, Toaster Kitchen Light	0.20	0.90	1.6
2.	Kitchen Light Toaster	0.18	0.80	1.1
3.	Solar Pump Boiler	0.13	0.75	1.1
4.	Toaster Laptop	0.15	0.78	1.2
5.	Kitchen Light Kettle	0.19	0.92	1.5

References



1. Brown R, Ghavami N, Siddiqui H-U-R, Adjrad M, Ghavami M, Dudley S (2017) Occupancy based household energy disaggregation using ultra wideband radar and electrical signature profiles. *Energy and Buildings* 141:134–141
2. Cao Z, Wan C, Zhang Z, Li F, Song Y (2020) Hybrid ensemble deep learning for deterministic and probabilistic low-voltage load forecasting. *IEEE Transactions on Power Systems* 35(3):1881–1897
3. Chen, T., and Guestrin, C. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD -16* (2016)
4. Ciabattoni L, Comodi G, Ferracuti F, Foresi G (2020) Ai-powered home electrical appliances as enabler of demand-side flexibility. *IEEE Consumer Electronics Magazine* 9(3):72–78

5. Elattar, E.E., Sabiha, N.A., and Alsharef, M. Short term electric load forecasting using hybrid algorithm for smart cities. *Appl Intell* 50 (2020), 3379–3399
6. Errapotu SM, Wang J, Gong Y, Cho J, Pan M, Han Z (2018) Safe: Secure appliance scheduling for flexible and efficient energy consumption for smart home iot. *IEEE Internet of Things Journal* 5(6): 4380–4391
7. Good N, Ellis KA, Mancarella P (2017) Review and classification of barriers and enablers of demand response in the smart grid. *Renewable and Sustainable Energy Reviews* 72:57–72
8. Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>
9. Greff, K., Srivastava, R. K., Koutnk, J., Steunebrink, B. R., and Schmidhuber, J. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems* 28, 10 (2017), 2222–2232
10. Guo Z, Zhou K, Zhang X, Yang S (2018) A deep learning model for short-term power load and probability density forecasting. *Energy* 160:1186–1200
11. Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. *Neural Computation*. 9(8):17351780
12. Ji Y, Buechler E, Rajagopal R (2019) Data-driven load modeling and forecasting of residential appliances. *IEEE Transactions on Smart Grid* 1–1
13. Kelly J, Knottenbelt W (2015) The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. *Scientific Data* 2:150007
14. Kong W, Dong ZY, Jia Y, Hill DJ, Xu Y, Zhang Y (2019) Short-term residential load forecasting based on lstm recurrent neural network. *IEEE Transactions on Smart Grid* 10(1):841–851
15. Kumar, S., Hussain, L., Banarjee, S., and Reza, M. Energy load forecasting using deep learning approach-lstm and gru in spark cluster. In *2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT)* (2018), pp. 1–4
16. Le T, Vo MT, Kieu T, Hwang E, Rho S, Baik SW (2020) Multiple electric energy consumption forecasting using a cluster-based strategy for transfer learning in smart building. *Sensors* 20:9
17. Le T, Vo MT, Vo B, Hwang E, Rho S, Baik SW (2019) Improving electric energy consumption prediction using cnn and bi-lstm. *Applied Sciences* 9:20
18. Massaoudi M, Refaat SS, Chihi I, Trabelsi M, Oueslati FS, Abu-Rub H (2021) A novel stacked generalization ensemble-based hybrid lgbm-xgb-mlp model for short-term load forecasting. *Energy* 214:118874
19. Monacchi, A., Egarter, D., Elmenreich, W., D'Alessandro, S., and Tonello, A. M. Greend: An energy consumption dataset of households in italy and austria. In *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)* (Nov 2014), pp. 511–516
20. Ponocko, J., and Milanovic, J. Forecasting demand flexibility of aggregated residential load using smart meter data. *IEEE Transactions on Power Systems PP* (01 2018), 1–1
21. Sajjad M, Khan ZA, Ullah A, Hussain T, Ullah W, Lee MY, Baik SW (2020) A novel cnn-gru-based hybrid approach for short-term residential load forecasting. *IEEE Access* 8:143759–143768
22. Sharda, S. link. <https://github.com/SwatiSharda/Appliance-power-forecasting>
23. Sheng Z, Wang H, Chen G (2020) Convolutional residual network to short-term load forecasting. *Appl Intell* 6(2):911–918
24. Shi H, Xu M, Li R (2018) Deep learning for household load forecasting a novel pooling deep rnn. *IEEE Transactions on Smart Grid* 9(5):5271–5280
25. Sideratos G, Ikonomopoulos A, Hatziaargyriou ND (2020) A novel fuzzy-based ensemble model for load forecasting using hybrid deep neural networks. *Electric Power Systems Research* 178:106025
26. Ullah A, Haydarov K, Ul Haq I, Muhammad K, Rho S, Lee M, Baik SW (2020) Deep learning assisted buildings energy consumption profiling using smart meter data. *Sensors* 20:3
27. Ullah FUM, Ullah A, Haq IU, Rho S, Baik SW (2020) Short-term prediction of residential power energy consumption via cnn and multi-layer bi-directional lstm networks. *IEEE Access* 8:123369–123380
28. Wang Y, Gan D, Sun M, Zhang N, Lu Z, Kang C (2019) Probabilistic individual load forecasting using pinball loss guided lstm. *Applied Energy* 235:10–20
29. Xu, W., Peng, H., and X., Z. A hybrid modelling method for time series forecasting based on a linear regression model and deep learning. *Appl Intell* 49 (2019), 3002–3015
30. Zhao B, Xiao S, Lu H, Liu J (2017) Waveforms, classification based on convolutional neural networks. In, (2017) *IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. 162–165

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

Article

A Hybrid Jaya–Powell’s Pattern Search Algorithm for Multi-Objective Optimal Power Flow Incorporating Distributed Generation

Saket Gupta ¹, Narendra Kumar ¹, Laxmi Srivastava ², Hasmat Malik ³ , Alberto Pliego Marugán ^{4,*} and Fausto Pedro García Márquez ⁵ 

¹ Electrical Engineering Department, Delhi Technological University, Delhi 110042, India; sguptamits@gmail.com (S.G.); dnk_1963@yahoo.com (N.K.)

² Electrical Engineering Department, Madhav Institute of Technology & Science, Gwalior 474005, India; srivastaval@hotmail.com

³ BEARS, CREATE Tower, NUS Campus, Singapore 138602, Singapore; hasmat.malik@gmail.com

⁴ Department of Quantitative Methods, CUNEF University, 28040 Madrid, Spain

⁵ Ingenium Research Group, Universidad Castilla-La Mancha, 13071 Ciudad Real, Spain; FaustoPedro.Garcia@uclm.es

* Correspondence: alberto.pliego@cunef.edu



Citation: Gupta, S.; Kumar, N.; Srivastava, L.; Malik, H.; Pliego Marugán, A.; García Márquez, F.P. A Hybrid Jaya–Powell’s Pattern Search Algorithm for Multi-Objective Optimal Power Flow Incorporating Distributed Generation. *Energies* **2021**, *14*, 2831. <https://doi.org/10.3390/en14102831>

Academic Editor:
Luis Hernández-Callejo

Received: 9 April 2021
Accepted: 11 May 2021
Published: 14 May 2021

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: A new hybrid meta-heuristic approach Jaya–PPS, which is the combination of the Jaya algorithm and Powell’s Pattern Search method, is proposed in this paper to solve the optimal power flow (OPF) problem for minimization of fuel cost, emission and real power losses and total voltage deviation simultaneously. The recently developed Jaya algorithm has been applied for the exploration of search space, while the excellent local search capability of the PPS (Powell’s Pattern Search) method has been used for exploitation purposes. Integration of the local search procedure into the classical Jaya algorithm was carried out in three different ways, which resulted in three versions, namely, J-PPS1, J-PPS2 and J-PPS3. These three versions of the proposed hybrid Jaya–PPS approach were developed and implemented to solve the OPF problem in the standard IEEE 30-bus and IEEE 57-bus systems integrated with distributed generating units optimizing four objective functions simultaneously and IEEE 118-bus system for fuel cost minimization. The obtained results of the three versions are compared to the Dragonfly Algorithm, Grey Wolf Optimization Algorithm, Jaya Algorithm and already published results using other methods. A comparison of the results clearly demonstrates the superiority of the proposed J-PPS3 algorithm over different algorithms/versions and the reported methods.

Keywords: distributed generation; hybrid Jaya–PPS algorithm; meta-heuristic; OPF; PPS

1. Introduction

With the increasing trend of penetration of renewable distributed generating (DG) units in the present day inter-connected restructured power system, the importance of solving optimal power flow problems has increased many folds. Optimal power flow results are crucial for planning, economic operation and control of an existing electrical power system, as well as for its future expansion planning. At the beginning of the 1960s, Carpentier addressed the OPF problem as an extension of economic load dispatch for the first time in history [1]. Since then, researchers have contributed significantly to this crucial issue. In a given electrical network, the OPF solution is required to regulate the control or decision variables set in the feasible region that optimizes some pre-defined objective functions. For the OPF problem, the control variables used are: V_g (generator bus voltages), P_g (generators’ active power outputs excluding slack bus), phase shifters, Tr (tap-settings of regulating transformer) and Q_c (injected reactive power using capacitor banks, FACTS devices etc.). Some of these variables are discrete, e.g., Tr , injected reactive power source

output Q_c , phase shifters, while others are continuous (e.g., P_g and V_g). The discrete nature of the control variable poses a challenge for the optimization technique and makes OPF a non-convex problem [2,3].

Integration of DGs seems to be quite appealing, but it is important to analyze their impact in a power network [4]. Optimal location and size of the DG unit have a significant effect on the reliability of power supply, operational cost, voltage profile, power loss and environmental pollution and voltage stability in a power system. Therefore, it has become a crucial task for researchers and industry personnel to determine the optimal location for DG and the size of the DG [5]. With the increase of the power injection from DGs into a power network, it is equally important to find out the optimum power generation and optimal setting of other control parameters to minimize fuel cost, emission cost and real power loss with an improved voltage profile [6].

OPF is a complex optimization problem, which associates several constraints and decision or control variables. The common objectives of the OPF problem are fuel cost minimization, emission minimization, real power loss minimization, voltage profile improvement and/or a combination of two or more of these objectives. The conventional algorithms depend on convexity to find the global best solution and are required to simplify relationships to achieve convexity. However, since the OPF problem is non-convex in general, several local minima can exist. If the valve point loading effects of thermal generators are taken into account, the non-convexity increases even further. Furthermore, traditional optimization approaches often use initial starting points (except for linear programming and convex optimization) and often converge or diverge to locally optimal solutions. These approaches are normally limited to particular cases of OPF and do not have much flexibility in terms of different kinds of objective functions or constraints that could be employed [7,8]. Except for linear programming and convex optimization, most of the conventional optimization algorithms cannot be guaranteed to be globally optimal because traditional algorithms are mainly local search. As a result, the final solution is always often dependent on the initial starting points.

Nowadays, several meta-heuristic algorithms have been developed by researchers, which are found to be powerful tools for handling difficult optimization problems. These random search, population-based algorithms are highly flexible, which means that they are appropriate to solve various types of optimization problems, including linear problems, non-linear problems and complex constrained optimization problems. Some of these methods are League Championship Algorithm (LCA) [3], Firefly Algorithm (FFA), Krill Herd Method (KH), Hybrid Firefly and Krill Herd Method (HFA) [9], Neighborhood Knowledge-based Evolutionary Algorithm (NKHA), Bare-Bones Multi-Objective Particle Swarm Optimization (BB-MOPSO), Multi-Objective Imperialist Competitive Algorithm (MOICA), Modified Non-dominated Sorting Genetic Algorithm (MNSGA-II), Multi-Objective Modified Imperialist Competitive Algorithm (MOMICA) [10], Moth Swarm Algorithm (MSA) [11], Multi-objective Evolutionary Algorithm based on decomposition-superiority of feasible (MOEA/D-SF) [12] and many others.

Recently, various hybrid algorithms have been investigated for effectively solving various optimization problems. Alsumait et al. [13] presented a hybrid GA-PS-SQP-based optimization algorithm to solve the economic dispatch (ELD) problem. Attaviriyapap et al. [14] suggested a hybrid optimization technique based on Evolutionary Programming and SQP algorithms for dynamic ELD problems. An integrated predator-prey (PP) optimization and a Powell search method were both proposed for the multi-objective hydrothermal scheduling problem [15]. Mahdad et al. [16] applied a hybrid DE-APSO-PS strategy to solve multi-objective power system planning. A hybrid modified imperialist competitive algorithm and SQP were employed to handle the constrained OPF problem [17]. Recently, considerable attention has been given to the Deep Neural Network approaches to the Energy Management problem [18,19].

It is observed that all Evolutionary Computing (EC)-based algorithms have some advantages and some disadvantages. Two main parts of any EC-based algorithm are

exploration and exploitation. Some algorithms have good exploration capability but poor exploitation and vice versa. The recently developed Jaya algorithm is capable of exploration, while Powell's Pattern Search (PPS) method has good search space exploitation capability. Hence, to boost the operational proficiency of the Jaya algorithm, Powell's Pattern Search method has been incorporated into it. Proper inclusion of the advantages of the Jaya and PPS algorithm would lead to better results for real-world complex, constrained and high-dimensional optimization problems. In the proposed hybrid Jaya-PPS algorithm, the Jaya algorithm was applied to explore a search space that is likely to provide the near-global solution and subsequently, the PPS algorithm was applied to attain a better solution.

The paper's contribution can be summed up as follows:

- The main contribution of this paper is to implement hybridization of two algorithms (Jaya and Powell's Pattern Search) in different manners and at different levels to find the best option for hybridization.
- Powell's Pattern Search method has been incorporated into the Jaya algorithm in three different ways, resulting in three variants, namely, J-PPS1, J-PPS2 and J-PPS3.
- The proposed hybrid Jaya and Powell's Pattern Search method utilizes the exploration property of the Jaya algorithm and the exploitation quality of Powell's Pattern Search method.
- This paper handles the OPF problem considering DG with four objectives functions simultaneously, namely, minimization of fuel cost, emission, real power losses and voltage profile improvement by converting the multi-objective OPF into a single objective OPF.
- In addition to Dragonfly Algorithm (DA), Grey Wolf Optimization (GWO) and Classical Jaya algorithms, three versions of hybrid Jaya and PPS, J-PPS1, J-PPS2 and J-PPS3 for the OPF problem are developed, wherein the excellent search capability of the PPS method has been exploited for further improvement of the solution provided by Jaya algorithm.

This paper is organized as follows: Section 2 presents the formulation of the OPF problem; The proposed hybrid Jaya-PPS algorithm is discussed in Section 3; Section 4 includes the results, while statistical analysis is incorporated into Section 5; Conclusions are presented in Section 6.

2. Problem Formulation

Mathematically, the objective function together and operating constraints of the OPF problem considered in this work are as follows [20]:

$$\text{Optimize } M(W, X) \quad (1)$$

Subject to the constraints given by Equation (2).

$$\begin{cases} g(W, X) = 0 \\ h(W, X) \leq 0 \end{cases} \quad (2)$$

where $M(W, X)$ is the objective function to be minimized, $g(W, X)$ and $h(W, X)$ are the equality and inequality constraints, respectively.

The control variables (X) include: the generator active power output (P_g) except at slack bus, generator bus voltage (U_g), tap-setting of transformer (T_{TR}) and shunt VAR compensation (Q_c). The dependent variables (W) consists of slack bus active power output P_{g1} , Load bus voltage (U_L), generator reactive power output (Q_g) and power flow in transmission lines (S_{il}). The control variables and state variables vectors can be expressed by Equation (3):

$$\begin{bmatrix} W \\ X \end{bmatrix} = \begin{bmatrix} P_{g1}, U_1 \dots U_{NLB}, Q_{g1}, \dots Q_{gNGN}, S_1, \dots S_{NH}, \\ P_{g2} \dots P_{gNGN}, U_{g1} \dots U_{gNGN}, Q_{C1} \dots Q_{CNC}, T_1 \dots T_{NTR} \end{bmatrix} \quad (3)$$

A control or decision variable can have any value within its minimum and maximum limits. In actual practice, transformer tap settings are not continuous variables. However, in this paper, to compare the results with the reported results, all the decision variables are considered to be continuous.

2.1. OPF Objective Functions

This paper handles the OPF problem considering DG with four objectives functions simultaneously, namely, minimization of fuel cost, emission, real power losses and improvement of voltage profile by converting the multi-objective OPF into a single objective OPF.

2.1.1. Fuel Cost Minimization

The prime motive of this objective function is to minimize the total cost of generation/fuel. It can therefore be expressed by Equation (4):

$$Z_{FCM} = \sum_{i=1}^{NGN} f(P_{gi})(\$/h) = \sum_{i=1}^{NGN} A_i + B_i P_{gi} + C_i P_{gi}^2 (\$/h) \quad (4)$$

where A_i , B_i , and C_i are the quadratic fuel cost coefficients of the i th generating unit and P_{gi} is the active power output of i th generating unit.

2.1.2. Emission Cost Minimization

The total emission pollutants such as SO_x (sulfur oxides) and NO_x (nitrogen oxides), which is an approximate combination of a quadratic and an exponential function can be expressed by Equation (5)

$$Z_{ECM} = \sum_{i=1}^{NGN} \alpha_i + \beta_i P_{Gi} + \gamma_i P_{Gi}^2 + \zeta_i \exp(\lambda_i P_{Gi}) \quad (5)$$

where α_i , β_i , γ_i , ζ_i , λ_i are the emission coefficients of i th generating unit.

2.1.3. Real Power Losses Minimization

The aim of the present case is to minimize real power losses. The total real power losses can be computed using Equation (6):

$$Z_{RPLM} = \sum_{i=1}^{NB} P_{gi} - \sum_{i=1}^{NB} P_{di} \quad (6)$$

where NB is the no. of buses, P_{gi} is the active power generation at i th generating unit and P_{di} is the real power load at i th load bus.

2.1.4. Voltage Profile Improvement

Voltage profile improvement means the voltage magnitude at load buses must not deviate much from 1.0 pu. Thus, the main motive, in this case, is to minimize voltage variation from 1.0 pu at all the load buses. In the present case, the objective function can be represented by Equation (7):

$$Z_{TVDM} = \sum_{i \in NLB} |U_i - 1| \quad (7)$$

where U_i is the voltage magnitude at i th load bus.

2.2. Constraints

The equality constraints are a combination of active and reactive non-linear power flow equations. In Equation (2), $g(W, X)$ is a set of equality constraints and is described by Equation (8):

$$g(W, X) = \begin{cases} \sum_{i=1}^{NB} (P_{gi} + P_{DGi}) - \sum_{i=1}^{NB} P_{di} - P_{Loss} = 0 \\ \sum_{i=1}^{NB} (Q_{gi} + Q_{DGi}) - \sum_{i=1}^{NB} Q_{di} - Q_{Loss} = 0 \end{cases} \quad (8)$$

where NB is the no. of buses, P_{gi} , Q_{gi} are the active and reactive power outputs of generating unit, P_{DGi} , Q_{DGi} are the active and reactive power outputs of DG unit, P_{di} , Q_{di} are active and reactive power load demand and P_{Loss} , Q_{Loss} are the total active and reactive power losses occurring in the lines, respectively.

The inequality constraints $h(W, X)$ represents operating limits of various equipment in a power system, which are described by Equation (9):

$$h(W, X) = \begin{cases} P_{gk}^{min} \leq P_{gk} \leq P_{gk}^{max} & k = 2 \dots NGN \\ U_{gk}^{min} \leq U_{gk} \leq U_{gk}^{max} & k = 1 \dots NGN \\ Q_{gk}^{min} \leq Q_{gk} \leq Q_{gk}^{max} & k = 1 \dots NGN \\ T_k^{min} \leq T_k \leq T_k^{max} & k = 1 \dots NTR \\ Q_{Ck}^{min} \leq Q_{Ck} \leq Q_{Ck}^{max} & k = 1 \dots NC \\ U_{Lk}^{min} \leq U_{Lk} \leq U_{Lk}^{max} & k = 1 \dots NLB \\ S_{lk} \leq S_{lk}^{max} & k = 1 \dots ntl \end{cases} \quad (9)$$

where active power output P_g , bus voltage U_g , and reactive power output Q_g , should be regulated by their lower and upper limits for all the generators, including slack bus generator and controllable VAR sources (Q_{Ck}), Transformer taps-setting (T_k) voltage of load buses (U_{Lk}) and power flow in transmission lines (S_{lk}) should vary between their minimum and maximum limits.

2.3. Combined Objective Function (COF)

The multi-objective function, which consists of four contradictory objective functions, i.e., minimization of fuel cost, emission cost, real power loss and total voltage deviation, is transformed into a single-objective function by using weighing factors to combine the four objective functions as given below.

$$COF(U, X) = Z_{FCM} + w_{ECM} \times Z_{ECM} + w_{RPLM} \times Z_{RPLM} + w_{TVDM} \times Z_{TVDM} \quad (10)$$

where w_{ECM} , w_{RPLM} and w_{TVDM} are weighing factors [9].

2.4. Incorporation of Constraints

The constraints are included in the combined objective function in the form of inequalities to find a feasible solution, and thus, the extended objective function can be defined by Equation (11):

$$M_{aug} = COF(U, X) + C_1 \sum_{k=1}^{NGN} h(P_{G_{Slack}} - P_{G_{Slack}}^{lim}) + C_2 \sum_{k=1}^{NGN} h(Q_{gk} - Q_{gk}^{lim}) + C_3 \sum_{k=1}^{NLB} h(U_{Lk} - U_{Lk}^{lim}) + C_4 \sum_{k=1}^{NGN} h(Q_{DGk} - Q_{DGk}^{lim}) \quad (11)$$

3. Jaya Algorithm

The Jaya algorithm is a comparatively new meta-heuristic optimization algorithm developed by Rao [21]. The working principle of the Jaya algorithm is that the numerical solution that has been obtained should go towards the better solution and should avoid

the inferior solutions for a particular optimization problem. The main advantage of the Jaya algorithm is that no algorithm-specific parameters are required, and thus, it is simple to implement this algorithm for solving various kinds of optimization problems.

Maximized (or minimized) value of objective function $M(z)$

Within the lower and upper bounds of the control variables, the initial solution p is randomly selected. After that, all the variables will be eventually updated according to Equation (12). On the basis of the fitness value of the objective function, the best and worst solutions are determined [21].

Let 'm' be the number of design variables (i.e., $j = 1, 2, 3, \dots, m$) and the 'n' is the number of candidate solutions (i.e., population size, $k = 1, \dots, n$). If $z_{i,j,k}$ represents the value of j th variable for the k th candidate in i th iteration; that value is updated according to Equation (12).

$$z_{i+1,j,k} = z_{i,j,k} + \alpha_{i,j,1} (z_{i,j,B} - \text{abs}(z_{i,j,k})) - \alpha_{i,j,2} (z_{i,j,W} - \text{abs}(z_{i,j,k})) \quad (12)$$

In (12), $z_{i,j,B}$ and $z_{i,j,W}$ are the best candidate and worst candidate value of variable j , respectively. The updated value of $z_{i,j,k}$ is $z_{i+1,j,k}$ and throughout the i th iteration, $\alpha_{i,j,1}$ and $\alpha_{i,j,2}$ are two random numbers for the j th variable within $[0, 1]$.

3.1. Powell's Pattern Search (PPS)

In 1962, Powell proposed the Powell search method, which was the expansion of the basic Pattern Search method. It is based on the conjugate direction method. Powell's Pattern Search (PPS) method is a derivative-free optimization technique that is ideal for solving a number of optimization problems beyond the scope of conventional optimization procedures. In general, the advantage of PPS is that the structure of the algorithm is remarkably simple, easy to implement and computationally efficient as well. PPS with meta-heuristic algorithm offers a flexible, balanced operator to enhance local search capability in contrast to another meta-heuristic algorithm. The following is the summary of the PPS algorithm underlying mechanisms [15,22]:

The search direction for l th coordinate for g th dimension of the n dimension search space can be defined as:

$$S_g^l = \begin{cases} 1; & g = l \\ 0; & g \neq l \end{cases} \quad (g = 1, 2, \dots, n; l = 1, 2, \dots, n) \quad (13)$$

The step length λ_g^* for g th decision variable can be determined as:

$$\lambda_g^* = \lambda_g^{\min} + \text{rand} \times (\lambda_g^{\max} - \lambda_g^{\min}) \quad (g = 1, 2, \dots, n) \quad (14)$$

where, λ_g^{\min} , λ_g^{\max} is the minimum and maximum step length for g th decision variable, respectively.

The vector of the decision variable X_g is updated once in the direction of the coordinate (l) as:

$$X_g = X_g + \lambda_g^* \times S_g^l \quad (g = 1, 2, \dots, n) \quad (15)$$

The vector of control variables is modified on the basis of the minimum objective function value. For all 'n' coordinates, this process continues. The pattern search direction is obtained for the next optimization cycle:

$$S_g^l = X_g - Z_g \quad (g = 1, 2, \dots, n; l = n + 1) \quad (16)$$

where Z_g is the initial value of the decision variable X_g .

Additionally, one of the coordinate's direction was discarded in the direction of pattern 'm' as:

$$S_g^m = S_g^l \quad (g = 1, 2, \dots, n; l = n + 1) \quad (17)$$

The process goes on until the entire direction of the coordinate is discarded and the entire operations restart in one of the coordinate directions again. Finally, until the Powell method has reached maximum iterations, the process of updating continues.

3.2. Proposed Hybrid Jaya–PPS Algorithm

Jaya algorithm has a strong capacity to exploit search space globally, but sometimes it suffers from premature convergence and can be stuck simply in local optima [6]. In order to overcome this problem and to make this algorithm more efficient, a hybrid Jaya algorithm, which combines the Jaya algorithm and PPS algorithm, is proposed in this paper. The PPS algorithm is a class of direct search methods. In general, it has immunity to strong local extremist trapping when used for local optimization. The proposed hybrid approach is primarily concerned with balancing the exploration and exploitation steps of the optimization procedure. PPS technique has good search space exploitation capability, while Jaya is able to explore the search space very well. The goal of incorporating PPS with Jaya is to combine the benefits of both algorithms.

Similar to other local search algorithms, the PPS algorithm is also sensitive to the initial or starting point. In selecting the initial point arbitrarily, it requires a large number of function evolutions, computation burden and slow convergence rate. In this research paper, to overcome these demerits of the PPS algorithm, the integration of local search procedure (Powell's Pattern Search) into the classical Jaya algorithm has been carried out in three different ways and the variants of hybrid Jaya–PPS thus developed are termed as J-PPS1, J-PPS2 and J-PPS3. To evaluate the performances of these variants, the common controlling parameters and the total number of function evaluations (NFE) used in J-PPS1, J-PPS2 and J-PPS3 algorithms are set the same as the classical Jaya algorithm. As stated, all the hybrid algorithms have the same number of function evaluations; thus, the additional criterion introduced in the proposed hybrid algorithms is to maintain (balance) the total number of function evaluations. The NFE has been used as a reference to the check efficiency of various algorithms in this paper.

In the first strategy (J-PPS1), the PPS algorithm was applied considering its initial point as the solution offered by the Jaya algorithm after applying it for 25% iterations. In this case, the optimization process is a two-step process. In step 1 (first 25% of Itermax), the Jaya algorithm was applied. However, in step 2 (remaining 75% of iterations), the PPS algorithms were applied using the optimal setting of control variables offered by the Jaya algorithm as initial point setting.

In the second strategy (J-PPS2), the Jaya algorithm and PPS algorithm were applied for an equal number of iterations to maintain the balance between the exploration and exploitation capability of the proposed J-PPS2 algorithm. In other words, the optimization process was completed in two steps. In step 1 (50% of Itermax), the Jaya algorithm was applied, while in step 2 (for the remaining 50% iterations), the PPS algorithms were applied sequentially as in the case of J-PPS1.

In the third strategy (J-PPS3), the PPS algorithm was applied after exploiting the 75% problem-solving capability of the Jaya algorithm, i.e., on the solution achieved by applying the Jaya algorithm for 75% iterations. In other words, the optimization process was divided into two steps. In step 1 (75% of Itermax), only the Jaya algorithm was applied, while in step 2 (for the remaining 25% iterations), only the PPS algorithms were applied. A flowchart of the proposed Jaya–PPS algorithm is shown in Figure 1.

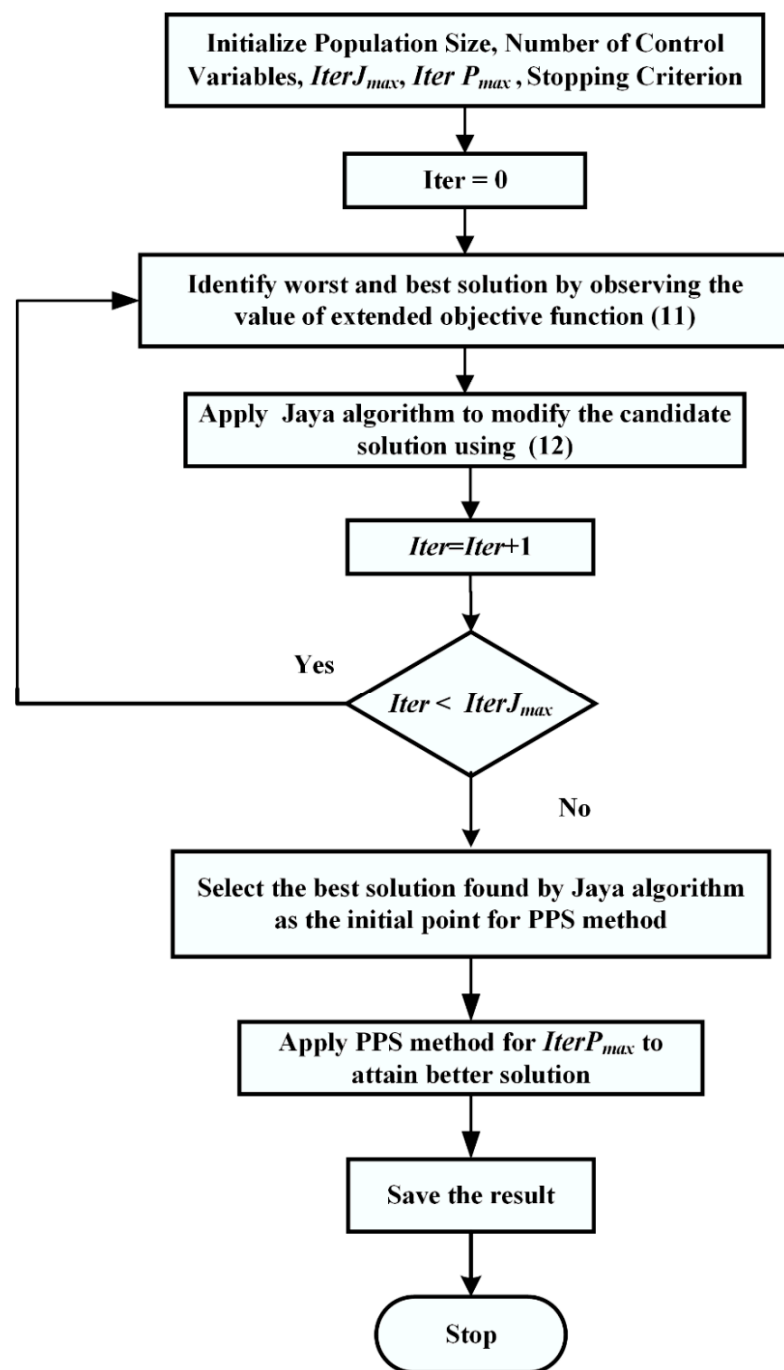


Figure 1. Flowchart of proposed hybrid Jaya–PPS algorithm.

The computational steps of the hybrid Jaya–PPS algorithm are summarized as follows:

- i. Initialize the population with control variables and set maximum iteration count $IterJ_{max}$ and the number of iterations $IterJ_{max}$ for the PPS method.
- ii. Set iteration $Iter = 0$.
- iii. Identify the worst and best solutions in the population on the basis of the extended objective function value Equation (11).
- iv. Modify the solutions using the best and the worst solutions Equation (12).
- v. If the modified solution is found to be better than the previous one, move to step vi, otherwise jump to step vii.
- vi. Replace the previous solution with the modified one and jump to step viii.
- vii. Retain previous solution.

- viii. Increase iteration number by 1, i.e., $\text{Iter} = \text{Iter} + 1$.
- ix. If $\text{Iter} < \text{Iter}_{\text{Jmax}}$, then move to step iii, else move to step x.
- x. Select the best solution found by the Jaya algorithm as the initial point for the PPS method and apply the PPS method for $\text{Iter}_{\text{Pmax}}$ iterations to attain a better solution.
- xi. Stop. Optimal solution achieved.

4. OPF Results and Discussion

In order to demonstrate the effectiveness of the proposed hybrid Jaya–PPS algorithm, DA, GWO, Jaya, J-PPS1, J-PPS2 and J-PPS3 algorithms were applied to the IEEE 30-bus system [6,23], IEEE 57-bus system [24] and IEEE 118-bus system for solving OPF problems with and without considering DG. The lower and upper limits of the 24 control variables, line data, bus data along with their initial settings for the IEEE 30-bus system, are taken from [23], while the emission and fuel cost coefficients are taken from [25]. In a combined single-objective function, the weight of an objective is proportional to the preference factor or weightage assigned to that objective function. This procedure is called a preference-based multi-objective optimization. For comparison, the combined objective function, COF, is obtained by considering the weighting factors W_{ECM} , W_{RPLM} and W_{TVDM} as 19, 22 and 21, respectively, as reported in [9]. The same procedure can be used for different systems also.

The IEEE 30-bus system is modified by including renewable energy source-based DG units. The optimal location for the DG unit is selected using the sensitivity of real power loss and the generation cost to each real and reactive power [6]; in this case, it was bus no. 30. At this bus, the capacity selected for the type 1 DG unit is 5 MW.

The IEEE 57-bus test system has 7 generators and 80 branches. The lower and upper voltage magnitude limits for all the generator and load buses of the system are considered to be 0.94 pu and 1.06 pu, respectively. The limits for the regulating transformers' tap settings are taken as 0.9 pu and 1.1 pu. The generator coefficients, lower and upper limits of all the 33 control variables and system data (bus data, line data) along with their initial settings are taken from [24]. At 100 MVA base, the real power demand and reactive power demand of this test system are 12.508 pu and 3.364 pu, respectively. In the case of the IEEE 57-bus system, the combined objective function, COF, is obtained by considering the weighting factors W_{ECM} , W_{RPLM} and W_{TVDM} as 300, 30 and 600, respectively. IEEE 57-bus system is modified by inserting DG units [26]. The optimal locations of the type 1 DG units are bus nos. 35 and 36 with the capacities of 47.9067 MW and 47.2636 MW, respectively.

To evaluate the scalability of proposed algorithms and prove their efficacy to solve large-scale problems, all three variants of Jaya–PPS algorithms, the GWO and DA algorithm were applied to solve the OPF problem IEEE 118-bus system. The system data, generator coefficients, lower and upper limits of all the 130 control variables, along with their initial settings, are taken from [27]. The active and reactive power demands of this test system are 42.42 and 14.38 pu, respectively, at the 100 MVA base.

To demonstrate the effectiveness of the proposed algorithm, five cases considered are as follows:

- Case 1: OPF no DG in IEEE 30-bus test system.
- Case 2: OPF with DG in IEEE 30-bus test system.
- Case 3: OPF no DG in IEEE 57-bus test system.
- Case 4: OPF with DG in IEEE 57-bus test system.
- Case 5: OPF no DG in IEEE 118-bus system.

Various trials were carried out with different population sizes and no. of iterations. The best results thus achieved and reported in this paper are for population $\text{pop} = 30$ and no. of iterations $\text{Iter}_{\text{max}} = 200$ for IEEE 30-bus test system and $\text{pop} = 40$ and $\text{Iter}_{\text{max}} = 300$ for IEEE 57-bus test system. The OPF results with and without the inclusion of DG obtained using various EC and hybrid Jaya algorithms are included in this section. These algorithms were developed using MATLAB 13a version in a 3.6 GHz Intel Processor, 8 GB RAM Core i7 and 64-bit operating personal computer.

To compare the performance of various algorithms, all the algorithms were run for the same number of function evaluations (NFE), which is equal to 6000 in the case of the IEEE 30-bus test system and 12,000 in the case of the IEEE 57-bus test system. Details of the implementation of various algorithms and inclusion of PPS in the three variants of hybrid Jaya–PPS algorithms are given in Table 1.

Table 1. Details of the DA, GWO, Jaya, J-PPS1, J-PPS2 and J-PPS3 algorithms.

IEEE-30 Bus System			
Algorithm	Population	Iterations	Total NFE = 6000
Dragonfly Algorithm	30	200	30×200
GWO Algorithm	30	200	30×200
Jaya Algorithm	30	200	30×200
J-PPS1	30	200	$30\text{JFE} \times 50 + 30\text{PSFE} \times 150$
J-PPS2	30	200	$30\text{JFE} \times 100 + 30\text{PSFE} \times 100$
J-PPS3	30	200	$30\text{JFE} \times 150 + 30\text{PSFE} \times 50$
IEEE-57 bus system and IEEE 118 Bus System			
Algorithm	Population	Iterations	Total NFE = 12,000
Dragonfly Algorithm	40	300	40×300
GWO Algorithm	40	300	40×300
Jaya Algorithm	40	300	40×300
J-PPS1	40	300	$40\text{JFE} \times 75 + 40\text{PSFE} \times 225$
J-PPS2	40	300	$40\text{JFE} \times 150 + 40\text{PSFE} \times 150$
J-PPS3	40	300	$40\text{JFE} \times 225 + 40\text{PSFE} \times 75$

NFE = Number of function evaluations; JFE = Number of Jaya Function Evaluations; PSFE = Number of PPS. Function evaluation.

4.1. Case 1: OPF No DG in IEEE 30-Bus Test System

In this case, the proposed hybrid Jaya–PPS algorithms, Dragonfly algorithm [28], GWO algorithm [29] and Jaya algorithm [21] were applied to solve the OPF problem considering the combined objective function without DG. Table 2 shows the results of these methods along with optimal control variable settings. The result clearly shows the superiority of the proposed J-PPS3 over other methods. Its combined objective function (965.0228) is less than those attained using other methods with no violation of the pre-specified constraints. The results of hybrid Jaya–PPS algorithms are compared with DA, GWO, Jaya algorithm and also with the reported results available in recent literature in Table 3.

Table 2. OPF results with optimum values of control variables for IEEE 30-bus system.

S. No.	Control Variable	DA	GWO	Jaya	J-PPS1	J-PPS2	J-PPS3
Generator real power output							
1	Pg2	0.52656	0.52553	0.5167	0.5259	0.5266	0.527
2	Pg5	0.31146	0.31068	0.32214	0.3156	0.3165	0.3155
3	Pg8	0.35	0.35	0.3497	0.3496	0.3496	0.35
4	Pg11	0.25774	0.26257	0.27264	0.2699	0.2692	0.2652
5	Pg13	0.21671	0.21185	0.20712	0.2115	0.2091	0.2099
Generator voltage setting							
6	Vg1	1.07429	1.07452	1.0728	1.0724	1.0735	1.0731

Table 2. Cont.

S. No.	Control Variable	DA	GWO	Jaya	J-PPS1	J-PPS2	J-PPS3
7	Vg2	1.05972	1.06035	1.05906	1.0584	1.0597	1.0593
8	Vg5	1.03127	1.03473	1.03371	1.0308	1.0332	1.0318
9	Vg8	1.04147	1.0423	1.04155	1.0406	1.0409	1.0402
10	Vg11	1.05456	1.05344	1.05018	1.0555	1.0431	1.0402
11	Vg13	1.01607	1.01938	1.02735	1.0179	1.0206	1.0186
Transformer tap setting							
12	T6-9	1.06778	1.08906	1.1	1.0991	1.0895	1.1
13	T6-10	1.01404	0.9811	0.94836	0.9499	0.9586	0.9435
14	T4-12	1.02163	1.01232	1.02587	1.0347	1.0304	1.0345
15	T28-27	1.00183	1.00725	1.00342	1.0095	1.0024	1.0023
Shunt VAR source setting							
16	Qc10	0.04965	0.0486	0	0.0104	0.0273	0.0225
17	Qc12	0.00025	0.0009	0.00054	0.0498	0.0031	0.0477
18	Qc15	0.03634	0.01863	0.04966	0.0344	0.0321	0.0474
19	Qc17	0.04876	0.03188	0.05	0.0343	0.0441	0.05
20	Qc20	0.0499	0.04829	0.04985	0.0478	0.0479	0.0481
21	Qc21	0.05	0.05	0.04997	0.05	0.0499	0.0497
22	Qc23	0.04898	0.04621	0.01739	0.0486	0.05	0.04
23	Qc24	0.04978	0.05	0.04986	0.0497	0.0484	0.0482
24	Qc29	0.02535	0.03226	0.03039	0.0344	0.0271	0.0274
COF		965.3516	965.3025	965.2868	965.2159	965.1201	965.0228
Fuel Cost		829.3587	829.2395	831.5493	830.9938	830.8672	830.3088
Emission		0.2370	0.2373	0.2358	0.2355	0.2357	0.2363
Real Power Loss (RPL)		5.6859	5.6843	5.5780	5.6120	5.6175	5.6377
Total Voltage Deviation (TVD)		0.3046	0.3094	0.3114	0.2990	0.2948	0.2949
L-Index		0.1387	0.1389	0.1396	0.1392	0.1393	0.1388
Pg1		122.8389	123.0213	122.1479	121.7620	121.9175	122.2777

Table 3. Results of the proposed method and other methods for case 1.

Algorithm	Comb. Obj Fun (COF)	Fuel Cost (\$/h)	Emission (ton/h)	RPL (MW)	TVD (pu)
Base Case	1336.64501	902.00457	0.22232	5.84233	1.16014
DA	965.35164	829.35878	0.23705	5.68593	0.30469
GWO	965.30257	829.23953	0.23731	5.68435	0.30945
Jaya	965.28681	831.54930	0.23582	5.57800	0.31147
J-PPS1	965.2159	830.9938	0.2355	5.6120	0.2990
J-PPS2	965.1201	830.8672	0.2357	5.6175	0.2948
J-PPS3	965.0228	830.3088	0.2363	5.6377	0.2949
MSA [11]	965.2905	830.639	0.25258	5.6219	0.29385
MPSO [11]	986.0063	833.6807	0.25251	6.5245	0.18991
MDE [11]	973.6116	829.0942	0.2575	6.0569	0.30347

Table 3. Cont.

Algorithm	Comb. Obj Fun (COF)	Fuel Cost (\$/h)	Emission (ton/h)	RPL (MW)	TVD (pu)
MFO [11]	965.8077	830.9135	0.25231	5.5971	0.33164
FPA [11]	971.9076	835.3699	0.24781	5.5153	0.49969
MSA [6]	*	838.9233	0.2116	5.6149	0.1535
ABC [6]	*	835.5230	0.2076	5.3948	0.1380
CSA [6]	*	834.5125	0.2099	5.4250	0.1373
GWO [6]	*	851.0491	0.2057	4.8925	0.2015
BSOA [6]	*	830.7115	0.2251	5.7446	0.1836
MJAYA [6]	*	833.3410	0.2064	5.1779	0.1196
MOEA/D-SF [12]	-	883.322	0.21867	4.4527	0.1322
MOMICA [10]	-	830.1884	0.2523	5.5851	0.2978
MOICA [10]	-	831.2251	0.267	6.0223	0.4046
MNSGA-II [10]	-	834.5616	0.2527	5.6606	0.4308
BB-MOPSO [10]	-	833.0345	0.2479	5.6504	0.3945
NKEA [10]	-	834.6433	0.2491	5.8935	0.4448
FKH [9]	-	828.3271	0.2549	5.3828	0.4925
KH [9]	-	827.7054	0.2526	5.4977	0.4930
FA [9]	-	829.5778	0.2527	5.5104	0.5661

* Different weighting factors.

From Table 3, it can be noted that the proposed J-PPS3 algorithm provides the minimum value of the combined objective function. This demonstrates the effectiveness of the proposed J-PPS3 algorithm as compared to DA, GWO, Jaya, J-PPS1 and J-PPS2 algorithms and other competitors [6,9–11]. Convergence characteristics of DA, GWO, Jaya, J-PPS1, J-PPS2 and J-PPS3 algorithms are shown in Figure 2, while Figure 3 displays the voltage profile provided by the proposed J-PPS3 algorithm. This figure shows that voltages magnitudes at all the buses are within the given upper and lower limits.

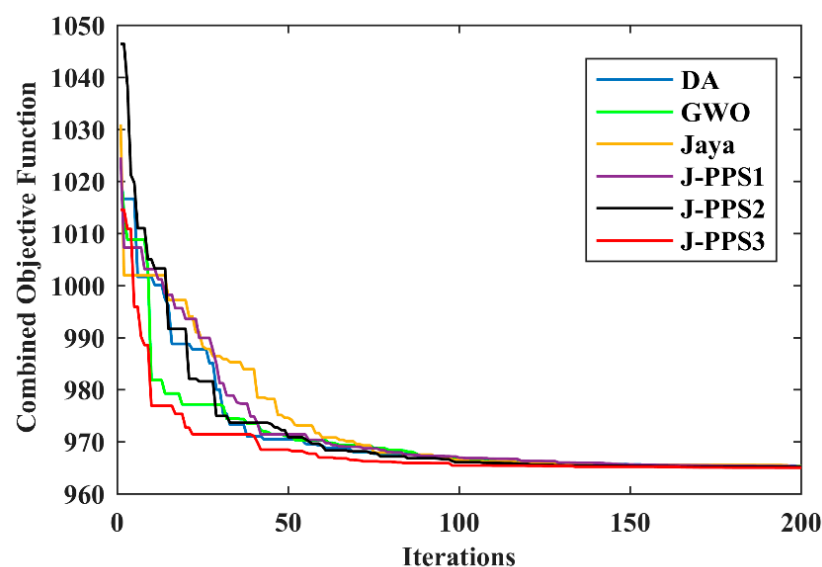


Figure 2. Convergence characteristics for various algorithms for Case 1.

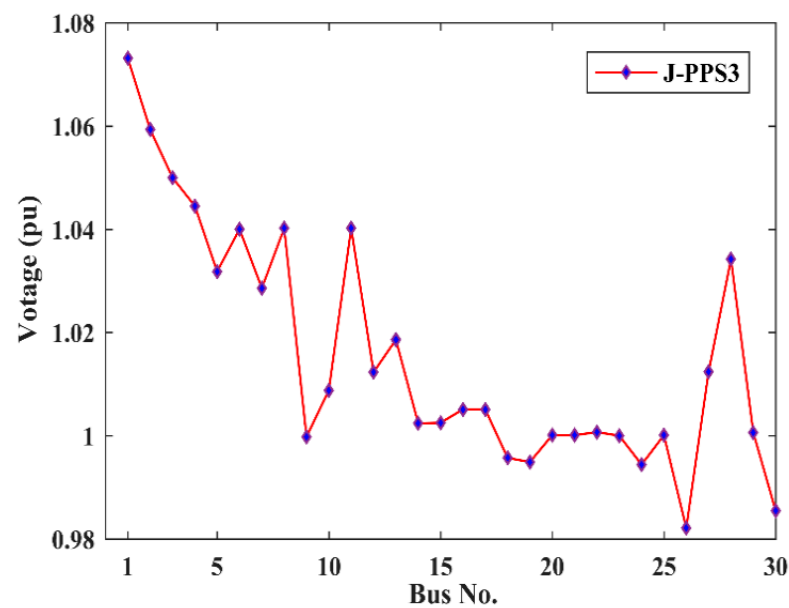


Figure 3. Voltage profile provided by J-PPS3 for Case 1.

4.2. Case 2: OPF with DG in IEEE 30-Bus Test System

In this case, the DA, GWO, Jaya, J-PPS1, J-PPS2 and J-PPS3 algorithms were applied to solve the optimal power flow problem incorporating DG considering the minimization of fuel cost, real power loss, emission and total voltage deviation. Afterward, their results were compared to find the best algorithm. The results of this case for all the algorithms along with optimal control variable settings are shown in Table 4. The numerical outcomes in Table 4 demonstrate that the proposed J-PPS3 algorithm is more effective as compared to other approaches for solving the OPF problem with DG. The combined objective function value obtained using the J-PPS3 algorithm is 937.3486, which is less than those of the DA, GWO, Jaya, J-PPS1 and J-PPS2 methods without any violation of the limits.

Table 4. OPF results with optimum values of control variables for the IEEE 30-bus system.

S. No.	Control Variable	DA	GWO	Jaya	J-PPS1	J-PPS2	J-PPS3
Generator Real power output							
1	Pg2	0.51902	0.51557	0.51579	0.516	0.5161	0.5162
2	Pg5	0.31184	0.31164	0.31143	0.3119	0.3116	0.311
3	Pg8	0.3500	0.34999	0.3500	0.35	0.35	0.35
4	Pg11	0.25881	0.2574	0.26254	0.2596	0.2619	0.261
5	Pg13	0.20565	0.2019	0.20564	0.2063	0.2039	0.2056
Generator voltage setting							
6	Vg1	1.07105	1.07422	1.06371	1.0709	1.0732	1.0724
7	Vg2	1.05748	1.06016	1.04926	1.0579	1.0594	1.0595
8	Vg5	1.03158	1.03316	1.02234	1.0297	1.0329	1.0325
9	Vg8	1.04001	1.04209	1.03217	1.0381	1.0434	1.0423
10	Vg11	1.09934	1.04159	1.04782	1.0459	1.0379	1.0416
11	Vg13	1.02435	1.01592	1.02994	1.023	1.0143	1.0164
Transformer tap setting							
12	T6-9	1.01015	1.09902	1.09958	1.0981	1.0997	1.0998
13	T6-10	1.1	0.92446	0.92521	0.958	0.9565	0.9585

Table 4. Cont.

S. No.	Control Variable	DA	GWO	Jaya	J-PPS1	J-PPS2	J-PPS3
14	T4-12	1.0343	1.02314	1.03347	1.0398	1.0182	1.0241
15	T28-27	1.0054	1.02139	1.00165	1.0028	1.013	1.0089
Shunt VAR source setting							
16	Qc10	0.00001	0.00136	0.00394	0.0436	0.0498	0.0494
17	Qc12	0.01415	0.04996	0.00005	0.0422	0.0344	0.0251
18	Qc15	0.04982	0.03682	0.04071	0.0457	0.0385	0.0454
19	Qc17	0.03936	0.05	0.04992	0.0484	0.05	0.0457
20	Qc20	0.02121	0.00011	0.04959	0.0481	0.05	0.0484
21	Qc21	0.04912	0.05	0.04867	0.0492	0.05	0.05
22	Qc23	0.03649	0.05	0.03892	0.0386	0.04	0.0397
23	Qc24	0.05	0.05	0.04854	0.0482	0.05	0.05
24	Qc29	0.01492	0.05	0.02444	0.012	0.0235	0.0211
COF		938.5816	938.4980	938.3787	937.6646	937.3837	937.3486
Fuel Cost		811.9476	811.2105	812.3347	811.9609	811.8993	811.8635
Emission Cost		0.2328	0.2340	0.2327	0.2329	0.2330	0.2329
Real Power Loss		5.2318	5.2836	5.2871	5.2381	5.2171	5.2214
Total Voltage Deviation		0.3385	0.3142	0.2525	0.2875	0.2990	0.2946
Pg1		119.0998	120.0336	119.1471	119.2581	119.2671	119.2414
L-Index (LI)		0.1046	0.1017	0.1036	0.1027	0.1017	0.1019

It should be noted that the combined objective function of the proposed J-PPS3 decreased from 965.0228 (Case 1) to 937.3486 by 2.86% after placing the DG as anticipated. Type 1 DG has been modeled as a negative load, and hence the total load demand is reduced. This further decreases the fuel cost and hence the combined objective function.

After integrating the DG, the convergence characteristics of DA, GWO, Jaya, J-PPS1, J-PPS2 and J-PPS3 algorithms are depicted in Figure 4. As can be observed from Figure 4, J-PPS3 provides fast and smooth convergence characteristics compared to the other methods. The voltage magnitudes at all the buses provided by the proposed J-PPS3 algorithm are shown in Figure 5, which are within the specified limits.

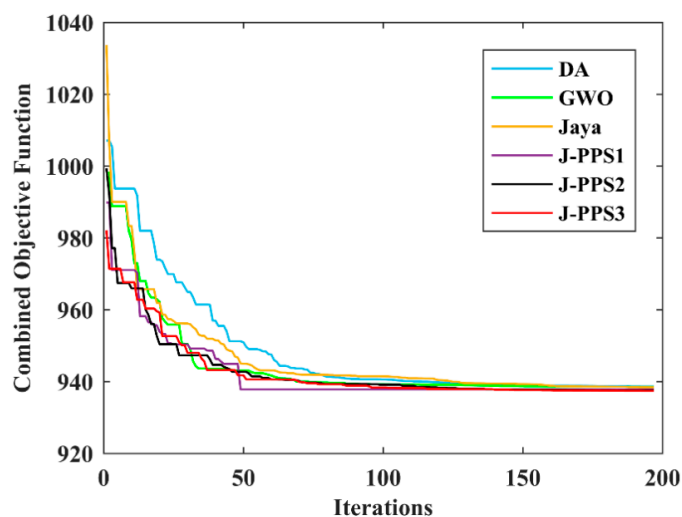


Figure 4. Convergence characteristics for various algorithms for Case 2.

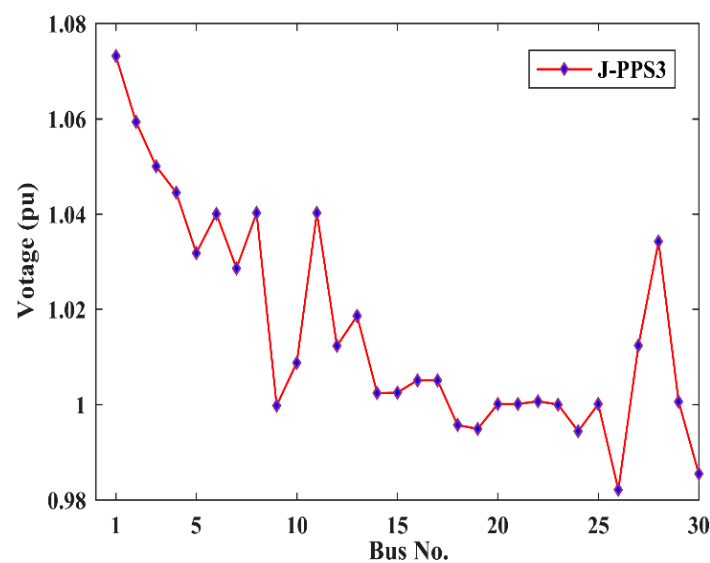


Figure 5. Voltage profile provided by J-PPS3 for Case 2.

4.3. Case 3: OPF No DG in IEEE 57-Bus Test System

In this case, to evaluate the scalability of the J-PPS3 algorithm and to prove its efficacy to solve large scale problems, all six DA, GWO, Jaya, J-PPS1, J-PPS2 and J-PPS3 algorithms were applied to solve the OPF problem in the IEEE 57-bus test system with no DG placed in it. In this case, the combined objective function for OPF comprises fuel cost, emission, real power loss and total voltage deviation. The OPF results and the optimal control variable settings of the J-PPS3 algorithm are compared with DA, GWO, Jaya, J-PPS1 and J-PPS2 in Table 5. Table 6 displays the comparison of numerical outcomes of DA, GWO, Jaya, J-PPS1, J-PPS2 and J-PPS3 algorithms and the reported results [10,12] for the IEEE 57-bus system. Figure 6 displays the convergence characteristics of DA, GWO, Jaya, J-PPS1, J-PPS2 and J-PPS3 algorithms.

Table 5. Optimum values of control variables for IEEE 57-bus system without DG.

S. No	Control Variable	DA	GWO	Jaya	J-PPS1	J-PPS2	J-PPS3
Generator active power output							
1	Pg2	0.9998	1	1	0.9997	0.9994	0.998
2	Pg3	0.52822	0.63533	0.57092	0.6064	0.6052	0.6058
3	Pg6	0.9934	0.92366	0.87967	1	0.9631	0.9196
4	Pg8	3.15449	3.13702	3.21403	3.1116	3.1263	3.1868
5	Pg9	0.99979	1	0.99992	0.9998	0.9994	0.9964
6	Pg12	4.09988	4.0985	4.09921	4.0848	4.1	4.1
Generator voltage setting							
7	Vg1	1.03896	1.04785	1.0333	1.0248	1.0321	1.0292
8	Vg2	0.95129	1.09823	1.09987	1.1	1.0873	1.0761
9	Vg3	1.0799	0.97546	1.08977	0.95	1.1	1.0919
10	Vg6	0.95	1.02	0.97047	1.0289	1.0209	1.0343
11	Vg8	0.99115	0.99545	1.00747	1.011	1.0133	1.0118
12	Vg9	0.95157	1.03031	0.97345	1.0273	1.0485	1.0125
13	Vg12	1.00525	1.01681	1.01609	1.0214	1.0109	1.0211

Table 5. Cont.

S. No	Control Variable	DA	GWO	Jaya	J-PPS1	J-PPS2	J-PPS3
Transformer tap setting							
14	T4-18	1.06804	1.09919	0.98026	0.9344	0.9205	1.0968
15	T4-18	0.90157	0.9	0.91699	1.0891	1.0094	0.9272
16	T21-20	1.00286	0.97166	1.09774	0.9698	0.9791	0.9821
17	T24-25	0.95085	1.03327	1.09303	0.9958	1.0922	1.0199
18	T24-25	1.0109	1.06638	0.90024	1.0045	0.9006	0.975
19	T24-26	1.04559	1.03573	1.0347	1.0213	1.0562	1.0157
20	T7-29	0.92573	0.95287	0.94193	0.9534	0.9446	0.9336
21	T34-32	0.92662	0.93717	0.93982	0.9444	0.916	0.9304
22	T11-41	0.90366	0.9	0.90013	0.9073	0.9	0.9116
23	T15-45	0.9482	0.96314	0.94642	0.9546	0.9364	0.9581
24	T14-46	0.94622	0.96437	0.97134	0.9641	0.976	0.977
25	T10-51	0.98181	1.02617	0.99404	0.9977	0.9789	0.9893
26	T13-49	0.92296	0.9	0.93097	0.9086	0.912	0.9
27	T11-43	0.91315	0.9141	0.92554	0.9278	0.9504	0.95
28	T40-56	1.09814	1.03063	1.06585	1.0017	0.9851	0.9818
29	T39-57	0.90081	0.95428	0.91838	0.9324	0.9307	0.9135
30	T9-55	0.97945	0.94635	0.99248	0.9699	0.9735	1.0015
Shunt VAR source setting							
31	Qc18	0.05841	0.0326	0.00118	0.1531	0	0.0857
32	Qc25	0.08446	0.19243	0.12834	0.1706	0.0792	0.1308
33	Qc53	0.14752	0.10041	0.14808	0.1577	0.0795	0.1171
COF		43,887.4176	43,864.8418	43,833.6421	43,825.8807	43,793.8820	43,788.6319
Fuel Cost		42,584.4552	42,587.9655	42,547.0948	42,575.9726	42,580.0946	42,564.4608
Emission Cost		1.3577	1.3447	1.3708	1.3336	1.3433	1.3566
Real Power Loss		13.6065	13.2727	12.772	12.5408	12.5242	12.5079
Voltage Deviation		0.8124	0.7921	0.8202	0.7893	0.7251	0.7365
Pg1		186.8485	184.6217	187.1970	183.1103	183.9874	182.6473
L-Index		0.2638	0.2429	0.2512	0.2418	0.2669	0.2501

Table 6. OPF results of IEEE 57-bus system without DG.

Algorithm	COF	FCost (\$/h)	Emission (ton/h)	PLoss (MW)	TVD (pu)
Base Case	53,828.14303	51,395.57064	2.76165	28.36589	1.25517
DA	43,887.43700	42,584.46959	1.35770	13.60665	0.81243
GWO	43,864.84184	42,587.97294	1.34478	13.27275	0.79209
Jaya	43,833.62963	42,547.09273	1.37089	12.77199	0.82018
J-PPS1	43,825.8807	42,575.9726	1.33366	12.54089	0.78931
J-PPS2	43,793.88205	42,580.09468	1.34333	12.52422	0.72511
J-PPS3	43,788.63196	42,564.46087	1.35666	12.50792	0.73656
MOMICA [10]	-	41,983.0585	1.496	13.6969	0.797

Table 6. Cont.

Algorithm	COF	FCost (\$/h)	Emission (ton/h)	PLoss (MW)	TVD (pu)
MOICA [10]	-	41,998.5661	1.7605	13.3353	0.8748
MNSGA-II [10]	-	42,070.82476	1.4965	14.4557	0.8896
BB-MOPSO [10]	-	41,994.019127	1.5336	12.609	1.0742
NKEA [10]	-	42,065.9964	1.5174	13.9764	1.042
MOEA/D-SF [12]	-	42,648.69	1.3437	11.8862	0.6713

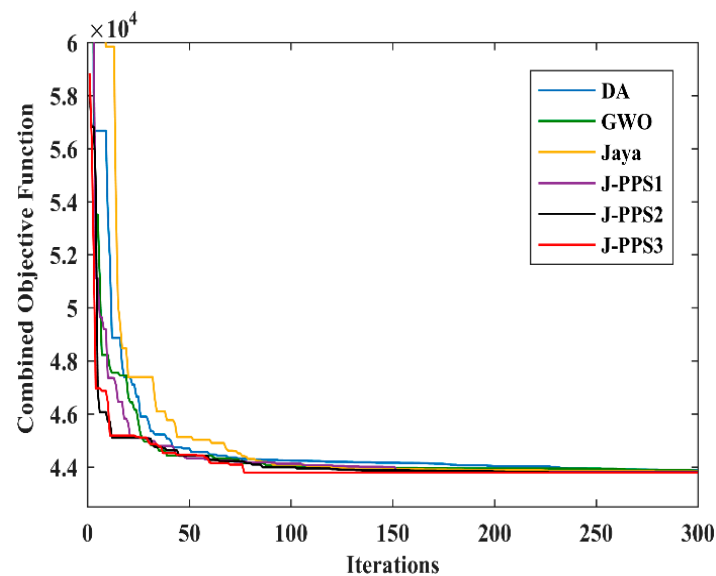


Figure 6. Convergence characteristics for various algorithms for Case 3.

The results in Table 6 prove the dominance of the hybrid J-PPS3 algorithm over other EC-based and hybrid Jaya–PPS algorithms in solving the OPF problem for a large-size power system. The proposed J-PPS3 algorithm provided the combined objective function value as 43,788.631, which is better than the combined objective functions offered by other algorithms with no constraint violation. The bus voltages profile obtained using the J-PPS3 algorithm is within specified limits, as shown in Figure 7.

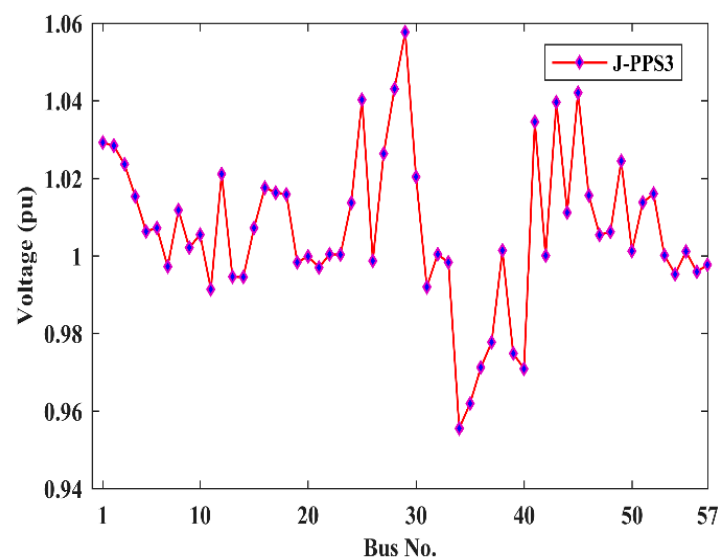


Figure 7. Voltage profile provided by J-PPS3 for Case 3.

4.4. Case 4: OPF with DG in IEEE 57-Bus Test System

In this case, to establish the effectiveness of the J-PPS3 algorithm for solving the OPF problem, the IEEE 57-bus test system with two DGs [26] is considered. The obtained results and the optimal control variable settings of the J-PPS3 algorithm are compared with those of DA, GWO, Jaya, J-PPS1 and J-PPS2 algorithms in Table 7.

Table 7. OPF results for IEEE 57-bus system with two DGs.

S. No	Control Variable	DA	GWO	Jaya	J-PPS1	J-PPS2	J-PPS3
Generator active power output							
1	Pg2	0.90988	0.9991	0.89863	0.9371	0.9368	0.9836
2	Pg3	0.47575	0.49202	0.47457	0.4728	0.4807	0.468
3	Pg6	0.59547	0.36955	0.46213	0.4847	0.3041	0.4147
4	Pg8	3.33676	3.49929	3.47521	3.3894	3.5637	3.4798
5	Pg9	0.99353	0.99999	0.99976	0.9991	0.9981	0.9999
6	Pg12	3.94011	3.86025	3.94345	3.975	3.9554	3.8995
Generator voltage setting							
7	Vg1	1.01833	1.0183	1.01557	1.01	1.0157	1.0166
8	Vg2	1.1	1.09571	1.09894	1.0944	1.1	1.0623
9	Vg3	1.06087	1.05416	1.06987	1.0323	1.0681	1.0749
10	Vg6	0.95	1.08551	1.0955	1.0069	1.0442	1.0497
11	Vg8	1.01291	1.01015	1.00346	0.9835	0.995	1.0026
12	Vg9	1.01131	0.95038	0.98631	1.0073	1.0118	1.0272
13	Vg12	1.01181	1.02021	1.00427	0.9893	1.0088	1.0031
Transformer tap setting							
14	T4-18	0.90037	1.09424	0.9	0.9	1.1	1.0552
15	T4-18	1.09998	0.9	1.0984	1.1	0.9078	0.9054
16	T21-20	1.04432	0.98565	0.98642	0.9794	0.9887	0.9924
17	T24-25	0.90006	1.1	1.034	1.0489	1.0928	1.0455
18	T24-25	1.07936	0.9	0.97881	1.0794	1.0714	1.1
19	T24-26	1.03695	1.00296	1.00993	0.9978	1.0712	1.0429
20	T7-29	0.97201	0.97227	0.95166	0.9815	0.9461	0.9496
21	T34-32	0.94356	1.00622	0.99596	0.986	0.989	0.9887
22	T11-41	0.9784	0.96604	0.95685	0.9594	0.967	0.9653
23	T15-45	0.97868	0.97901	0.98099	0.9703	0.9716	0.9806
24	T14-46	0.97857	0.97699	0.96899	0.9481	0.9764	0.9876
25	T10-51	0.98858	0.99304	0.98007	0.9786	0.9806	0.9783
26	T13-49	0.92431	0.93811	0.93098	0.9007	0.9317	0.9271
27	T11-43	0.99037	1.00372	0.98421	0.9637	0.9821	0.9837
28	T40-56	0.91921	0.90084	0.93647	0.9182	0.9	0.9265
29	T39-57	0.98075	0.99828	0.98771	0.9938	1.0123	0.979
30	T9-55	0.95101	0.98913	0.98154	0.9159	0.9474	0.9383
Shunt VAR source setting							
31	Qc18	0.18621	0.00004	0.12891	0.1062	0.0216	0.0212
32	Qc25	0.06171	0.14674	0.11436	0.1869	0.169	0.1737

Table 7. Cont.

S. No	Control Variable	DA	GWO	Jaya	J-PPS1	J-PPS2	J-PPS3
33	Qc53	0.1296	0.19995	0.16624	0.1797	0.0752	0.065
	COF	39,200.1782	39,173.0979	39,162.8890	39,167.5961	39,165.9645	39,136.3249
	Fuel Cost	38,120.8335	38,114.7354	38,105.9569	38,059.9136	38,048.2507	38,033.8329
	Emission Cost	1.2751	1.3099	1.3218	1.3035	1.3612	1.3115
	Real Power Loss	12.3189	13.1703	12.5706	12.8818	13.3724	12.9742
	Total Voltage Deviation	0.5454	0.4504	0.4721	0.5469	0.5136	0.5329
	Pg1	142.7987	146.7800	142.8253	142.7015	145.1223	144.0540
	L-Index	0.1393	0.1241	0.1290	0.12921	0.1252	0.1250

The results in Table 7 prove the dominance of the proposed hybrid J-PPS3 algorithm over other EC-based and hybrid Jaya–PPS algorithms in successfully handling the OPF problem in large-scale systems penetrated with two DG units. The proposed J-PPS3 algorithm provided the combined objective function value as 39,136.324, which is better than the combined objective functions offered by other algorithms without violating the constraints. The combined objective function of J-PPS3 decreased from 43,778.631 (Case 3) to 39,136.324 (by 10.60%) after implanting two DGs as expected.

In this case, the proposed J-PPS3 algorithm also provided fast and smooth convergence characteristics compared to other algorithms, as shown in Figure 8. The bus voltages profile obtained by the J-PPS3 algorithm is within limits, as shown in Figure 9.

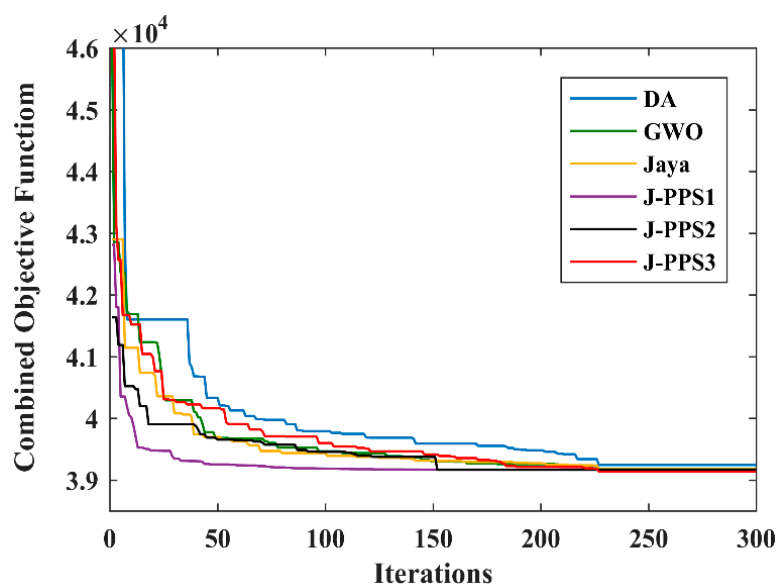


Figure 8. Convergence characteristics for various algorithms for Case 4.

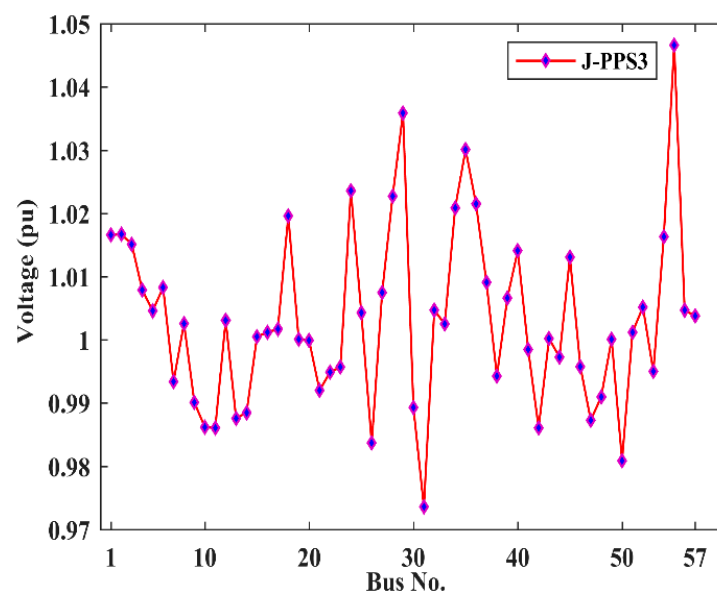


Figure 9. Voltage profile provided by J-PPS3 for Case 4.

4.5. Case 5: OPF No DG in IEEE 118-Bus System

In Case 5, fuel cost is selected as the main objective. The minimum fuel cost obtained by the J-PPS3 algorithm is 129,507.6123 \$/h, while the minimum fuel cost obtained by J-PPS2 and J-PPS1 algorithms is 129,821.4309 \$/h and 129,961.8924 \$/h, respectively. The minimum fuel cost obtained using hybrid Jaya–PPS algorithms and other meta-heuristic algorithms are depicted in Table 8. From Table 8, it is clear that the fuel cost obtained from J-PPS3 algorithm is the least compared to other methods, demonstrating the effectiveness of the proposed J-PPS3 algorithm compared to the J-PPS2, J-PPS1, DA, GWO algorithms and other competitors in handling the OPF problem in a large-sized power system. The fuel cost characteristics for Case 5 are shown in Figure 10.

Table 8. Case 5 (Fuel cost minimization) results in IEEE 118-bus system.

Algorithm	Fuel Cost (\$/h)	TVD (pu)	PG69 (Slack Bus)	Power Loss	
				MW	MVA _r
Base Case	131,220.0208	1.4389	513.8101	132.8101	782.6073
J-PPS1	129,961.8924	1.4402	489.0344	113.4784	745.3196
J-PPS2	129,821.4309	1.5238	430.2158	118.5608	762.0786
J-PPS3	129,507.6123	1.3486	440.1366	109.6528	668.4798
Jaya	130,165.8424	1.4991	482.2581	112.9269	740.0970
DA	130,016.5235	1.4596	450.9608	119.1369	751.3072
GWO	130,053.1453	1.4015	461.0356	108.2561	698.1435
IMFO [20]	131,820.0000	1.5944	407.192	77.6522	−910.020
Interior point [30]	129,720.70	N. A	N. A	N. A	N. A
CC-ACOPF [31]	129,662.0	N. A	N. A	N. A	N. A
NLP [32]	129,700	N. A	N. A	N. A	N. A
QP [32]	129,600	N. A	N. A	N. A	N. A
MIQP [32]	129,600	N. A	N. A	N. A	N. A
ALC-PSO [33]	129,546.0847	N. A	N. A	N. A	N. A

Table 8. Cont.

Algorithm	Fuel Cost (\$/h)	TVD (pu)	PG69 (Slack Bus)	Power Loss	
				MW	MVA _r
PSOGSA [34]	129,733.58	N. A	N. A	73.21	N. A
GPU-PSO [35]	129,627.03	N. A	N. A	76.984	N. A

IMFO = improved moth-flame optimization; ALC-PSO = particle swarm optimization with an aging leader and challengers; PSOGSA = Hybrid Particle Swarm Optimization and Gravitational Search Algorithm; GPU-PSO = Partial swarm optimization-based graphics processing units; CC-ACOPF = Chance Constrained Optimal Power Flow; QP = quadratic programming; MIQP = Mixed Integer quadratic programming.

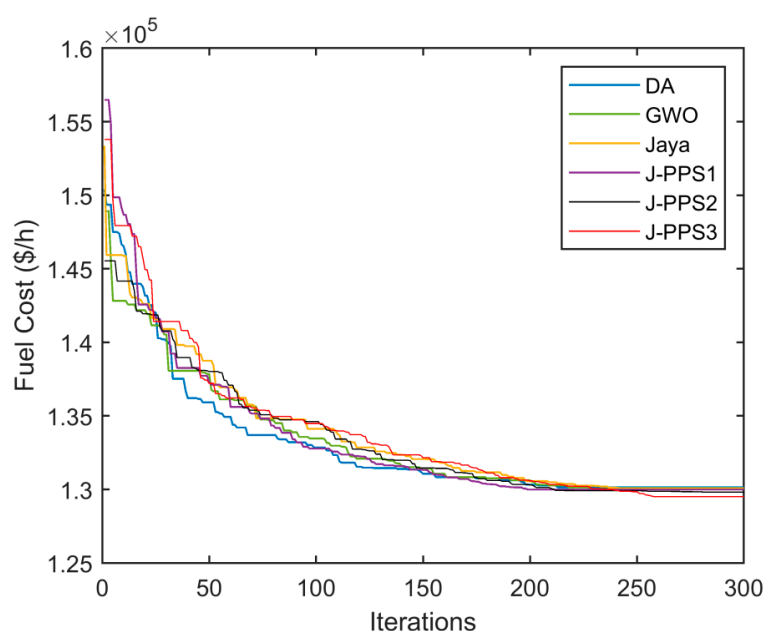


Figure 10. Convergence characteristics for various algorithms for Case 5.

5. Statistical Analysis

Statistical analysis was carried out to evaluate the robustness of DA, GWO, Jaya, J-PPS1, J-PPS2 and J-PPS3 algorithms to solve the OPF problem with and without DG. A total of 50 independent trials were carried out with the same population size and same no. of function evaluations for each case. As previously mentioned, the population sizes and the maximum NFE were 30 and 6000 for the IEEE 30-bus test system, respectively, and as 40 and 12,000 for the IEEE 57-bus system, respectively, which provided the best results. These trials were utilized to find out the best value, worst value, average (mean) value of OPF results and standard deviation (SD) required for statistical analysis of various algorithms implemented in this paper and are shown in Tables 9 and 10, respectively. These tables show that, for all the considered cases of IEEE 30-bus and IEEE 57-bus test systems, the best, worst and mean values are nearest to each other; therefore, the standard deviation values are low. The smallest SD values offered by the proposed J-PPS3 algorithm in all the cases clearly indicate that statistically meaningful results are obtained by the proposed J-PPS3 method. This affirms the robustness of the proposed algorithm.

Table 9. Performance measures of various algorithms for IEEE 30-bus system.

Algorithm	Without DG				Incorporating DG			
	Best	Worst	Mean	Std. Deviation	Best	Worst	Mean	Std. Deviation
DA	965.3516	966.4352	965.8734	0.02526	938.5816	939.1761	938.7554	0.02615
GWO	965.3025	966.7339	965.7564	0.02151	938.4980	939.2469	938.8678	0.02387
Jaya	965.2868	966.8154	965.8975	0.01983	938.3787	939.2543	938.9781	0.01955
Jaya-PPS1	965.2159	965.6587	965.4260	0.01851	937.6646	937.8942	937.7815	0.01829
Jaya-PPS2	965.1201	965.4089	965.2481	0.01809	937.3837	937.6582	937.4982	0.01785
Jaya-PPS3	965.0228	965.3261	965.2094	0.01132	937.3486	937.5803	937.4623	0.01105

Table 10. Performance measures of various algorithms for IEEE 57-bus system.

Algorithm	Without DG				Incorporating DG			
	Best	Worst	Mean	Std. Deviation	Best	Worst	Mean	Std. Deviation
DA	43,887.437	43,973.873	43,893.893	0.02988	39,200.178	39,218.879	39,207.656	0.02887
GWO	43,864.841	43,896.887	43,871.698	0.02917	39,173.097	39,181.365	39,178.432	0.02828
Jaya	43,833.629	43,845.953	43,839.894	0.02820	39,162.889	39,175.542	39,168.764	0.02812
Jaya-PPS1	43,825.880	43,839.720	43,833.542	0.02588	39,167.596	39,176.742	39,172.427	0.02609
Jaya-PPS2	43,793.882	43,804.659	43,800.752	0.02602	39,165.964	39,174.694	39,169.524	0.02531
Jaya-PPS3	43,788.631	43,797.462	43,793.298	0.01299	39,136.324	39,140.437	39,138.542	0.01297

6. Conclusions

This paper proposes a hybrid Jaya-PPS algorithm using Jaya and Powell's Pattern Search method to solve the multi-objective optimal power flow problem incorporating DG to minimize generation fuel cost, emission, real power loss and voltage profile improvement simultaneously. The multi-objective optimization problem has been solved by transforming it into a single-objective optimization problem using weighting factors. Three versions of hybrid Jaya-PPS techniques J-PPS1, J-PPS2 and J-PPS3, were developed by integrating the PPS method in different ways. In order to evaluate the performance of the proposed hybrid Jaya-PPS algorithms, these algorithms were employed to solve the OPF problem in standard IEEE 30-bus and IEEE 57-bus systems with/without DG and IEEE 118-bus systems for fuel cost minimization. The results achieved by the hybrid Jaya-PPS algorithms were compared to the Dragonfly algorithm, Grey Wolf Optimization and Jaya algorithms, and the reported results published in recent literature. The numerical outcomes demonstrate that the proposed J-PPS3 algorithm dominates other approaches when solving the OPF problem. For example, the combined objective function found by Jaya-PPS1 for the 30-bus system is 937.3486, with a reduction of 2.86% of the original system, with a 0.01105 standard deviation. This benefit increases further with the size of the system. Statistical analysis has shown that the hybrid J-PPS3 algorithm is a reliable and robust optimization algorithm. As the hybrid J-PPS3 algorithm has good exploration and exploitation properties, it can reliably solve the OPF problem in practical power systems.

Author Contributions: All authors have contributed equally in technical and non-technical work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature

W	dependent variable
X	control variable
$M(W, X)$	objective function
$g(W, X)$ & $h(W, X)$	equality and inequality constraints
NB	number of buses
NLB	number of load buses
Ntl	number of transmission lines
NGN	numbers of generators
NC	number of VAR compensation units
NTR	number of regulating transformers
P_{di} and Q_{di}	active and reactive load
P_{gi} and Q_{gi}	Active & reactive power generations
P_{Loss} and Q_{Loss}	real and reactive power loss
U_{gk}^{min} and U_{gk}^{max}	minimum & maximum voltage limits of kth generator bus
Q_{gk}^{min} and Q_{gk}^{max}	minimum and maximum limits of reactive power output of kth generator
P_{gk}^{min} and P_{gk}^{max}	minimum and maximum active power limits of kth generating unit
T_k^{max} and T_k^{min}	maximum and minimum tap setting of kth transformer
U_{Lk}^{max} and U_{Lk}^{min}	maximum and minimum voltage limit of kth load bus
S_{lk}^{max}	maximum MVA flow in kth transmission line
C_1, C_2, C_3 and C_4	penalty factors corresponding to limit violations
A_i, B_i , and C_i	fuel cost coefficients of the ith generating unit
P_{g1}	slack bus generator's active power output
$\alpha_i, \beta_i, \gamma_i, \xi_i, \lambda_i$	emission coefficients of ith generating unit
pop	Population size
Itermax	Maximum No. of iterations
IterJmax	Maximum No. of Jaya iterations
IterPmax	Maximum No. of PPS iterations
NFE	Number of function evaluations
JFE	Number of Jaya Function Evaluations
PSFE	Number of PPS Function Evaluation

References

1. Carpentier, J. Optimal power flows. *Int. J. Electr. Power Energy Syst.* **1979**, *1*, 3–15. [\[CrossRef\]](#)
2. Hazra, J.; Sinha, A.K. A multi-objective optimal power flow using particle swarm optimization. *Eur. Trans. Electr. Power* **2010**, *21*, 1028–1045. [\[CrossRef\]](#)
3. Bouchekara, H.; Abido, M.; Chaib, A.; Mehasni, R. Optimal power flow using the league championship algorithm: A case study of the Algerian power system. *Energy Convers. Manag.* **2014**, *87*, 58–70. [\[CrossRef\]](#)
4. Quadri, I.A.; Bhowmick, S.; Joshi, D. A comprehensive technique for optimal allocation of distributed energy resources in radial distribution systems. *Appl. Energy* **2018**, *211*, 1245–1260. [\[CrossRef\]](#)
5. Khaled, U.; Eltamaly, A.M.; Beroual, A. Optimal Power Flow Using Particle Swarm Optimization of Renewable Hybrid Distributed Generation. *Energies* **2017**, *10*, 1013. [\[CrossRef\]](#)
6. Elattar, E.E.; ElSayed, S.K. Modified JAYA algorithm for optimal power flow incorporating renewable energy sources considering the cost, emission, power loss and voltage profile improvement. *Energy* **2019**, *178*, 598–609. [\[CrossRef\]](#)
7. Srivastava, L.; Singh, H. Hybrid multi-swarm particle swarm optimisation based multi-objective reactive power dispatch. *IET Gener. Transm. Distrib.* **2015**, *9*, 727–739. [\[CrossRef\]](#)
8. Low, S.H. Convex Relaxation of Optimal Power Flow—Part I: Formulations and Equivalence. *IEEE Trans. Control. Netw. Syst.* **2014**, *1*, 15–27. [\[CrossRef\]](#)
9. Khelifi, A.; Bentouati, B.; Chettih, S. Optimal Power Flow Problem Solution Based on Hybrid Firefly Krill Herd Method. *Int. J. Eng. Res. Afr.* **2019**, *44*, 213–228. [\[CrossRef\]](#)

10. Ghasemi, M.; Ghavidel, S.; Ghanbarian, M.M.; Gharibzadeh, M.; Vahed, A.A. Multi-objective optimal power flow considering the cost, emission, voltage deviation and power losses using multi-objective modified imperialist competitive algorithm. *Energy* **2014**, *78*, 276–289. [\[CrossRef\]](#)
11. Mohamed, A.-A.A.; Mohamed, Y.S.; El-Gaafary, A.A.; Hemeida, A.M. Optimal power flow using moth swarm algorithm. *Electr. Power Syst. Res.* **2017**, *142*, 190–206. [\[CrossRef\]](#)
12. Biswas, P.P.; Suganthan, P.N.; Mallipeddi, R.; Amaratunga, G.A.J. Multi-objective optimal power flow solutions using a constraint handling technique of evolutionary algorithms. *Soft Comput.* **2019**, *24*, 2999–3023. [\[CrossRef\]](#)
13. Alsumait, J.; Sykulski, J.; Al-Othman, A. A hybrid GA–PS–SQP method to solve power system valve-point economic dispatch problems. *Appl. Energy* **2010**, *87*, 1773–1781. [\[CrossRef\]](#)
14. Attaviriyapap, P.; Kita, H.; Tanaka, E.; Hasegawa, J. A hybrid EP and SQP for dynamic economic dispatch with nonsmooth fuel cost function. *IEEE Trans. Power Syst.* **2002**, *17*, 411–416. [\[CrossRef\]](#)
15. Narang, N.; Dhillon, J.S.; Kothari, D.P. Multi-objective fixed head hydrothermal scheduling using integrated predator-prey optimization and Powell search method. *Energy* **2012**, *47*, 237–252. [\[CrossRef\]](#)
16. Mahdad, B.; Srairi, K. Multi objective large power system planning under sever loading condition using learning DE-APSO-PS strategy. *Energy Convers. Manag.* **2014**, *87*, 338–350. [\[CrossRef\]](#)
17. Ben Hmida, J.; Chambers, T.; Lee, J. Solving constrained optimal power flow with renewables using hybrid modified imperialist competitive algorithm and sequential quadratic programming. *Electr. Power Syst. Res.* **2019**, *177*, 105989. [\[CrossRef\]](#)
18. Wu, J.; Wei, Z.; Li, W.; Wang, Y.; Li, Y.; Sauer, D.U. Battery Thermal- and Health-Constrained Energy Management for Hybrid Electric Bus Based on Soft Actor-Critic DRL Algorithm. *IEEE Trans. Ind. Inform.* **2021**, *17*, 3751–3761. [\[CrossRef\]](#)
19. Wu, J.; Wei, Z.; Liu, K.; Quan, Z.; Li, Y. Battery-Involved Energy Management for Hybrid Electric Bus Based on Expert-Assistance Deep Deterministic Policy Gradient Algorithm. *IEEE Trans. Veh. Technol.* **2020**, *69*, 12786–12796. [\[CrossRef\]](#)
20. Taher, M.A.; Kamel, S.; Jurado, F.; Ebeed, M. An improved moth-flame optimization algorithm for solving optimal power flow problem. *Int. Trans. Electr. Energy Syst.* **2019**, *29*, e2743. [\[CrossRef\]](#)
21. Rao, R.V. Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems. *Int. J. Ind. Eng. Comput.* **2016**, *7*, 19–34. [\[CrossRef\]](#)
22. Rao, S.S. *Engineering Optimization: Theory and Practice*; John Wiley & Sons: Hoboken, NJ, USA, 2019.
23. Lee, K.Y.; Park, Y.M.; Ortiz, J.L. A united approach to optimal real and reactive power dispatch. *IEEE Trans. Power Appar. Syst.* **1985**, *PAS-104*, 1147–1153. [\[CrossRef\]](#)
24. Zimmerman, R.D.; Murillo-Sánchez, C.E.; Thomas, R.J. Matpower. Available online: <https://matpower.org/docs/ref/matpower5.0/case57.html> (accessed on 11 May 2021).
25. Niknam, T.; Narimani, M.R.; Jabbari, M.; Malekpour, A.R. A modified shuffle frog leaping algorithm for multi-objective optimal power flow. *Energy* **2011**, *36*, 6420–6432. [\[CrossRef\]](#)
26. Charles, J.K.; Otero, N.A. Effects of distributed generation penetration on system power losses and voltage profiles. *Int. J. Sci. Res. Publ.* **2013**, *3*, 1–8.
27. Power System Test Cases. Available online: <http://www.ee.washington.edu/research/pstca> (accessed on 11 May 2021).
28. Meraihi, Y.; Ramdane-Cherif, A.; Acheli, D.; Mahseur, M. Dragonfly algorithm: A comprehensive review and applications. *Neural Comput. Appl.* **2020**, *32*, 16625–16646. [\[CrossRef\]](#)
29. Mirjalili, S.; Mirjalili, S.M.; Lewis, A. Grey Wolf Optimizer. *Adv. Eng. Softw.* **2014**, *69*, 46–61. [\[CrossRef\]](#)
30. Jiang, Q.; Geng, G.; Guo, C.; Cao, Y. An Efficient Implementation of Automatic Differentiation in Interior Point Optimal Power Flow. *IEEE Trans. Power Syst.* **2009**, *25*, 147–155. [\[CrossRef\]](#)
31. Brust, J.J.; Anitescu, M. Convergence Analysis of Fixed Point Chance Constrained Optimal Power Flow Problems. *arXiv* **2021**, arXiv:2101.11740.
32. Fortenbacher, P.; Demiray, T. Linear/quadratic programming-based optimal power flow using linear power flow and absolute loss approximations. *Int. J. Electr. Power Energy Syst.* **2019**, *107*, 680–689. [\[CrossRef\]](#)
33. Singh, R.P.; Mukherjee, V.; Ghoshal, S. Particle swarm optimization with an aging leader and challengers algorithm for the solution of optimal power flow problem. *Appl. Soft Comput.* **2016**, *40*, 161–177. [\[CrossRef\]](#)
34. Radosavljević, J.; Klimenta, D.; Jevtić, M.; Arsić, N. Optimal Power Flow Using a Hybrid Optimization Algorithm of Particle Swarm Optimization and Gravitational Search Algorithm. *Electr. Power Compon. Syst.* **2015**, *43*, 1958–1970. [\[CrossRef\]](#)
35. Roberge, V.; Tarbouchi, M.; Okou, F. Optimal power flow based on parallel metaheuristics for graphics processing units. *Electr. Power Syst. Res.* **2016**, *140*, 344–353. [\[CrossRef\]](#)

A Hybrid Model for Combining Neural Image Caption and k-Nearest Neighbor Approach for Image Captioning

Kartik Arora, Ajul Raj, Arun Goel, Seba Susan^{[0000-0002-6709-6591]*}

Department of Information Technology,
Delhi Technological University,
Bawana Road, Delhi, India-110042
seba_406@yahoo.in

Abstract. A hybrid model is proposed that integrates two popular image captioning methods to generate a text-based summary describing the contents of the image. The two image captioning models are the Neural Image Caption (NIC) and the k -nearest neighbor approach. These are trained individually on the training set. We extract a set of five features, from the validation set, for evaluating the results of the two models that in turn is used to train a logistic regression classifier. The BLEU-4 scores of the two models are compared for generating the binary-value ground truth for the logistic regression classifier. For the test set, the input images are first passed separately through the two models to generate the individual captions. The five-dimensional feature set extracted from the two models is passed to the logistic regression classifier to take a decision regarding the final caption generated which is the best of two captions generated by the models. Our implementation of the k -nearest neighbor model achieves a BLEU-4 score of 15.95 and the NIC model achieves a BLEU-4 score of 16.01, on the benchmark Flickr8k dataset. The proposed hybrid model is able to achieve a BLEU-4 score of 18.20 proving the validity of our approach.

Keywords: Image captioning, k -nearest neighbor, Neural networks, Long-Short Term Memory (LSTM), Logistic regression, Hybrid model, BLEU scores.

1 Introduction

Image captioning is the task of generating text that describes a given image. Describing the contents of an image in a textual way has many applications, for example, describing contents on a screen for visually impaired, real time captioning of videos, and in robotics. Image captioning is different from image classification since it involves not only identifying the objects in the image, but also summarizing the relation between the objects in the image using natural language. A lot of research work has been done in this topic in recent times [1] [2] [3] [4]. One method is to use the k -Nearest Neighbor (kNN) approach [2] to select a caption in the dataset that accurately describes the image. This involves finding a consensus caption from a set of captions that describe images that are similar to the test image. If the set of images are diverse, one would expect the selected caption to be generic (example- a dog). If the images are similar, the caption selected would be more specific (example- a black dog).

Another approach is to use the Neural Networks to generate novel captions that describe the test image. The model in [1] uses a recurrent neural network for generating the sentences and is also called Neural Image Caption (NIC). This approach uses a combination of pre-trained convolutional neural network VGG16 that processes the input image, and Long-Short Term Memory (LSTM) [5] which is well suited for processing sequential data i.e. the captions in this case. We propose to integrate the two approaches- NIC and kNN into a hybrid model that uses a trained logistic regression classifier to choose the better caption. If the test image is quite similar to the images in the training set, one would expect the captions generated by Nearest Neighbor be better than the Neural Network approach. Otherwise, the novel captions generated by NIC tend to be better. We seek to find a set of criteria to choose the model that would provide the better captions for an input image. We use Flickr8K dataset to evaluate our model. The organization of this paper is as follows: the related work is discussed in section 2, the proposed hybrid model is presented in section 3, the results are analyzed in section 4 and the conclusion is drawn in section 5.

2 Related Work

Image caption generation is mostly implemented either by distance-based matching or by training neural networks like LSTM. Distance or similarity based classifiers have managed to carve their own niche despite the success of neural networks for image classification [10]. This fact is reconfirmed through our own experiments which prove that for several examples, the distance-based classifier outperforms the neural network in caption generation. A hybrid model incorporating the goodness of both distance-based and neural network approaches is proposed in our work and will be described in detail in subsequent sections. We discuss works pertaining to both approaches in this section. Devlin *et al.* proposed a k-Nearest Neighbor (kNN) approach for image captioning [2]; this is the kNN model in our hybrid technique. This approach generates a caption for the test image using the captions of images in the training set that are similar to the test image. This approach finds the nearest k images for the test image using the cosine similarity metric, for three different feature spaces: GIST, fc7 and fc7-fine. Each of the k images have 5 captions each, so the candidate caption set C consists of $n=5*k$ captions. Then the Consensus Caption c^* according to [2] is the caption with highest similarity score (BLEU-4 score [7]) with all the captions within subset M of C .

$$c^* = \underset{c \in C}{\operatorname{argmax}} \max_{M \subset C} \sum_{c' \in M} \operatorname{Sim}(c, c') \quad (1)$$

Vinyals *et al.* proposed a neural network model called Neural Image Caption (NIC) to generate novel image captions; this is the second model used in our hybrid technique. It consists of an encoder CNN connected to an LSTM network. The CNN is pre-trained on image classification task and the last hidden layer of CNN is used as input to the LSTM network. The model maximized the probability $p(S|I)$ where I is the image and S is a sequence of words $\{S_1, S_2, \dots\}$ that describes the image. The model used the CNN to extract a feature vector from the input image which is then used as an input to the

LSTM circuit. Then, using the encoded image and the partial caption (which at the beginning would be null or a special start token), the output of the LSTM would be the probability of each word in the dictionary to be the next word in the sequence, out of which we either take the one with the maximum probability (greedy search), or choose the top i words (beam search). The word would be added to the previous partial caption to generate the new partial caption. The caption would end when a special end token was selected or a specified length was reached. The model was trained using stochastic gradient descent [8] minimizing the loss function:

$$L(I, S) = -\sum_t \log p_t(S_t) \quad (2)$$

The loss function is the summation of negative log probabilities of correct word S_t at each step t . Before training, basic pre-processing is done on captions in the dataset. All words with occurrences greater than 5 are kept in the dictionary. Unlike the k -Nearest Neighbor (kNN) model which chooses a caption from the training set that best represents the test image, NIC Generator can construct novel captions which are not present in the training set. Recent literature focusses on modifying LSTM-based network architectures for improvising the natural language in image captions. A hierarchical LSTM in a recent work [9] comprises of a phrase decoder and a sentence decoder for generating image description. Xu *et al.* [4] developed an attention-based model which was able to focus on relevant parts of the image to generate better captions. Ding *et al.* [3] proposed that instead of assigning equal weights to all words, one could assign different weights to words according to their importance in the sentence. For example, for the images of a bird, the word 'bird' would have a larger weight. This also prevents mis-recognition since the main subjects in the image are identified correctly.

3 Proposed Hybrid Model

In this section, we propose a hybrid model that combines two state-of-the-art models: Neural Image Caption (NIC) [1] and k -Nearest Neighbor approach [2] to generate captions for an input image. Both models were described in detail in section 2. In our NIC implementation, the image is first fed to a pre-trained convolutional neural network, Inception-V3 [6], that produces a rich representation of the input image by encoding it into a fixed-length vector of size 2048. This vector is the output of the last hidden layer of the Inception-V3 model and it is given as input to a LSTM which is a recurrent neural network.

Our hybrid technique incorporates a meta-classifier (logistic regression) that will choose the better model for a given input image and use the caption generated by this model. The general idea for a hybrid model is shown in Fig. 1. We propose a generic algorithm which requires classifying an image into either category A or category B, where category A is the category of images that are better modeled by NIC and category B is the category of images that are better modeled by k -Nearest Neighbor model. We use logistic regression for this classification and discuss some possible set of features which can help us to produce a robust classifier.

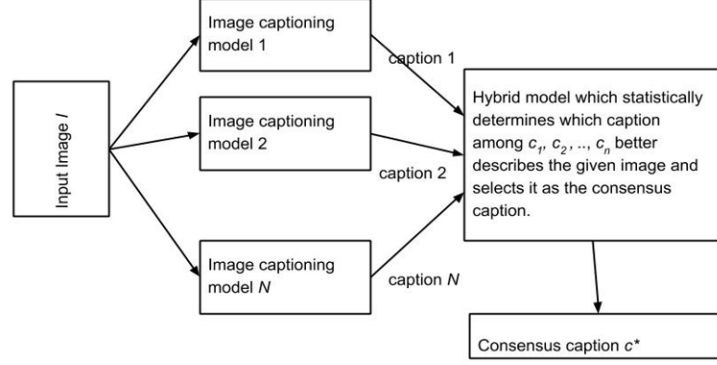


Fig. 1. Generic layout of a hybrid model which combines N image captioning models

3.1 Methodology

Let M_1 and M_2 be the NIC model [1] and the k -nearest neighbor model [2], respectively that are trained individually on the training set of images and their ground-truths (captions). We propose a hybrid model M which selects a consensus caption c^* for a given input image I using the following steps.

1. Generate caption c_1 using M_1 for given input image I .
2. Generate caption c_2 using M_2 for given input image I .
3. A set of five features (section 3.2) are extracted from the two models using the validation set and fed as input to the logistic regression classifier.
4. The BLEU-4 scores of the two models are compared for generating the binary-value ground truth for the logistic regression classifier (0 if $\text{BLEU}(M_1) \geq \text{BLEU}(M_2)$, else 1 if $\text{BLEU}(M_1) < \text{BLEU}(M_2)$).
5. For the test set, the input images are first passed separately through the two models to generate the individual captions. The five-dimensional feature set extracted from the two models is passed to the logistic regression classifier to take a decision regarding the final caption generated which is the best of two captions generated by the models.
6. If the logistic regression classifier predicts that M_1 produces a better caption i.e. predicted value $y=0$ for I , then $c^* = c_1$, else if $y=1$, then $c^* = c_2$.

The block diagram for the process is shown in Fig. 2.

3.2 Feature extraction and normalization

We propose a set of five features extracted from the classification results of models M_1 and M_2 that is used for training the meta-classifier in our hybrid model. The qualitative definitions of the new features are enlisted as follows.

- a. The confidence score that the NIC model has for the caption it generated for the given image.

- b. The confidence score that the k -nearest neighbor model has for the caption it generated for the given image.
- c. A measure of similarity between the images in the training data to the input image in consideration.
- d. The length of the captions generated by both the models.

The above conditions led us to formulate the five features quantitatively as follows.

1. Length-normalized log probability p^* of c_I (from M_1) which is a measure of the confidence M_1 has on c_I .
2. The average (BLEU-4) similarity score of c^* (from M_2), from (1).
3. Cosine similarity S_c between input image I and the image Y , where Y belongs to K the set of the k nearest images to I , summed over all Y . Averaging the similarity scores across multiple samples of a class improves accuracy as observed in [12]. The features are derived from the fc7 layer of VGG16 pretrained network used in model M_2 [2].

$$S_c = \sum_{Y \in K} \cos_sim(I, Y) \quad (3)$$

4. The two features: length of caption $c_I (=l_I)$ and length of caption $c_2 (=l_2)$.

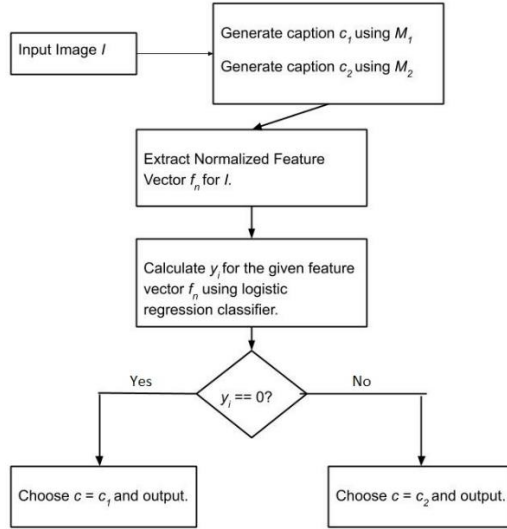


Fig. 2. Process flow of generating consensus caption c^* for input image using proposed model M

These features are now normalized so that their absolute values lie in the range $[0, 1]$. We divide both l_I and l_2 by the length of the longest caption in the Flickr8k dataset ($=35$ words). We divide S_c by five times the number of summands in its summation (i.e. $5k$) in (3) and we divide $\sum_{c' \in M} Sim(c^*, c')$ by the number of summands in its summation (i.e. $|M|$) in (1). Finally, we have the normalized feature vector f_n given by (4) that is the input to the logistic regression classifier.

$$f_n = \left\{ p^*, \frac{\sum_{c' \in M} \text{Sim}(c^*, c')}{|M|}, \frac{S_c}{5k}, \frac{l_1}{35}, \frac{l_2}{35} \right\} \quad (4)$$

4 Results

We use Flickr8K dataset [11] to evaluate our hybrid model. It contains 8092 images with 5 captions each, out of which 6000 are used for training, 1000 for testing and the rest for development. We first compare BLEU-1 and BLEU-4 scores for various LSTM beam sizes in Table 1. In Table 2, we present results for: 1) Neural network based NIC model [1], with beam size $i=3$ which has the highest BLEU-1 score in Table 1, 2) k -nearest neighbor model [2] with $k=30$ and $|M|=50$, 3) Proposed hybrid model which integrates the above two models using the logistic regression classifier. All the scores reported have been evaluated on the Flickr8k dataset on a system with Intel® Core™ i5-8300H Processor, with 8 GB RAM and GTX 1050 graphics running on Windows 10 Pro 64 bit. The code was compiled on Python 3.6.9 using TensorFlow 2.1.0. The proposed hybrid model was able to achieve higher BLEU-1 and BLUE-4 scores on the Flickr8k test data than the individual models as observed from Table 2.

Table 1. BLEU-1 and BLEU-4 scores for NIC for different beam sizes.



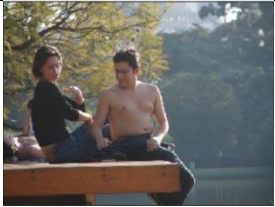


Beam Size (i)	BLEU-1	BLEU-4
1	57.59	14.44
3	58.13	16.01
5	58.09	16.29
7	57.89	16.03

Table 2. BLEU-1 and BLEU-4 scores for kNN, NIC and Hybrid Model.

Model	BLEU-1	BLEU-4
kNN with $k=30, M=50$	56.02	15.95
NIC with beam size 3	58.12	16.01
Hybrid model	59.67	18.20

Table 3 shows some examples of captions generated using our hybrid scheme. One of the captions (either NIC or kNN) shown in each of the five cases is incorrect. As observed, in two cases out of five, the kNN model outperforms the neural network approach (NIC). Our hybrid model chooses the best caption that describes the scene adequately in all five cases. The code of our hybrid model is made available online at <https://github.com/rizal-rovins/hybrid-image-captioning-model>

Table 3. Images and their captions generated by the hybrid model.

1		NIC Caption	Hybrid model classifies image to category A (NIC). <i>Final caption:</i> A boy is jumping off a dock into a lake.
		A boy is jumping off a dock into a lake.	
		kNN Caption	
		A woman in a bikini jumping off a dock into a lake.	
2		NIC Caption	Hybrid model classifies image to category B (kNN). <i>Final caption:</i> A boy jumping in the air on the beach.
		A little girl in pink bathing suit is jumping into the water.	
		kNN Caption	
		A boy jumping in the air on the beach.	
3		NIC Caption	Hybrid model classifies image to category A (NIC). <i>Final caption:</i> A man and a woman are sitting on a park bench.
		A man and a woman are sitting on a park bench.	
		kNN Caption	
		A girl doing a handstand on a trampoline.	
4		NIC Caption	Hybrid model classifies image to category A (NIC). <i>Final caption:</i> A mountain biker rides through the woods.
		A mountain biker rides through the woods.	
		kNN Caption	
		A man riding a bike down a hill.	
5		NIC Caption	Hybrid model classifies image to category B (kNN). <i>Final caption:</i> A dog running through the water.
		A white dog fetches a stick in his mouth.	
		kNN Caption	
		A dog running through the water.	

5 Conclusion

We have presented a hybrid model that combines two existing image captioning models- NIC and k-Nearest Neighbor (kNN) trained separately on images from the training set. We extract a novel set of five features from the validation set for evaluating the captions generated by the two models, that is used to train a logistic regression classifier. The BLEU-4 scores of the two models are compared for generating the $\{0, 1\}$ ground truth values for the logistic regression classifier. Our hybrid model chooses the best caption that describes the scene adequately for a given test image. The proposed method was able to achieve higher BLEU-1 and BLUE-4 scores on the benchmark Flickr8k dataset. The technique can be further extended to combine more than two image captioning models and advanced forms of LSTM incorporating attentional mechanism could be used in place of NIC in our model.

References

1. Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156-3164. 2015.
2. Devlin, Jacob, Saurabh Gupta, Ross Girshick, Margaret Mitchell, and C. Lawrence Zitnick. "Exploring nearest neighbor approaches for image captioning." arXiv preprint arXiv:1505.04467 (2015).
3. Ding, Guiguang, Minghai Chen, Sicheng Zhao, Hui Chen, Jungong Han, and Qiang Liu. "Neural image caption generation with weighted training and reference." *Cognitive Computation* 11, no. 6 (2019): 763-777.
4. Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention." In International conference on machine learning, pp. 2048-2057. 2015.
5. Schmidhuber, Jürgen, and Sepp Hochreiter. "Long short-term memory." *Neural Comput* 9, no. 8 (1997): 1735-1780.
6. Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818-2826. 2016.
7. Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "BLEU: a method for automatic evaluation of machine translation." In Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311-318. Association for Computational Linguistics, 2002.
8. Bottou, Léon. "Large-scale machine learning with stochastic gradient descent." In Proceedings of COMPSTAT'2010, pp. 177-186. Physica-Verlag HD, 2010.
9. Tan, Ying Hua, and Chee Seng Chan. "Phrase-based image caption generator with hierarchical LSTM network." *Neurocomputing* 333 (2019): 86-100.
10. Susan, Seba, and Gitin Kakkar. "Decoding facial expressions using a new normalized similarity index." In *2015 Annual IEEE India Conference (INDICON)*, pp. 1-6. IEEE, 2015.
11. Hodosh, Micah, Peter Young, and Julia Hockenmaier. "Framing image description as a ranking task: Data, models and evaluation metrics." *Journal of Artificial Intelligence Research* 47 (2013): 853-899.
12. Susan, Seba, and Srishti Sharma. "A fuzzy nearest neighbor classifier for speaker identification." In *2012 Fourth International Conference on Computational Intelligence and Communication Networks*, pp. 842-845. IEEE, 2012.

A Language-Independent Speech Sentiment Analysis Using Prosodic Features

Monil Bansal*

Department of Information Technology
Delhi Technological University
Delhi, India
bansalmonil7@gmail.com

Sampriti Yadav*

Department of Information Technology
Delhi Technological University
Delhi, India
ysampriti@gmail.com

Dinesh K. Vishwakarma

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

* Both authors have contributed equally

Abstract— Sentiment Analysis from audios is a challenging field of ongoing research because, though humans can recognize emotions from facial expressions, gestures, and tone of the ongoing conversation, it can be a challenging task for machines. In the following report and its corresponding research work, an audio emotion detection system has been proposed and is used to perform a comparative study to detect emotions from 3 datasets. In this research, a comprehensive study of diverse datasets (TESS and RAVDESS) along with a custom dataset is made. Our proposed model aggregates results from diverse baseline machine learning models trained on different parameters and hyperparameters and its performance is calculated and compared with existing research. The proposed model uses different features of the audio such as MFCC, Mel Spectrogram, and Chroma which are extracted from the datasets which make our model independent of language barriers. The efficacy of this model is evaluated using various evaluation metrics such as confusion matrix, overall accuracy, and F1-score. As for the outcome of the research and experiment, the overall accuracy is 99.46% and 89.62% for TESS and RAVDESS respectively. Furthermore, an accuracy of 78.28% has been reported for the custom dataset. It is also found that, the most predictable emotion is anger while the most misclassified is fear.

Keywords—Emotion detection, Mel Spectrogram, MFCC

I. INTRODUCTION

In today's world, the basis for communication amongst human beings is the exchange of information through speech. Humans express their feelings through emotions. It is easier for other humans to understand and interpret them based on hand gestures, facial expressions, and tone of the speech, while this is not the case for machines. But this gap is diminishing day by day as technology is enhancing and new research is being carried out each day. This research work has attempted to explore the effectiveness of various machine learning models and build an aggregator model (Speech Recognition System). It also aims to compare the efficacies of these individual architectures. The algorithmic changes in the proposed model include extending our comprehensive study to aggregator models and their hyperparameter optimization. These models have been fed with engineered data (feature extraction using Librosa and selection using techniques like PCA) with the best features. In the case of emotions, while some extreme ones (e.g. anger) can be easy to identify, some soft and neutral ones

(e.g. neutral) can be very difficult to recognize. For machines to identify emotions from audio signals, they need to analyze various features such as pitch, loudness, energy at the very core. This research has enabled us to perform multi-modal emotion recognition using the TESS [1], RAVDESS [2], and our custom dataset. Every human expresses some emotions more than others [3]. The proposed research work has classified emotions into one of six Ekman's emotions [4] (Anger, Joy, Sadness, Fear, Disgust, and Surprise). This aggregator model can be brought to use in many applications. To design systems that give more personalized user preferences, it is extremely important to analyze how emotion impacts both modes of interaction between humans (verbal and nonverbal).

Some of the common applications are Healthcare, Counseling Security, and AI assistants at call centers. For example, if an AI assistant could determine if a user is sad or Angry, it can switch to more informed communication.

II. LITERATURE REVIEW

A lot of work in the field of sentiment/emotion analysis from human audio has been done previously using different datasets and using several different baseline machine learning models. Neiberg et al. [1] worked on emotion recognition in spontaneous speech. According to them, it is more difficult to detect emotions from live or spontaneous sounds than using the pre-recorded dataset. For emotion recognition in spontaneous real-time speech, they have proposed an approach in which they have used three classifiers and combined their results. Since the end of the 20th century, a lot of research has been done on Sentiment Analysis on different modalities such as text, audio, and video. Proposed methodologies vary from linguistic analysis [20, 21], to ML approaches [22, 23], to data mining techniques [24, 25, 26].

Indira et al. [2] have used the RAVDESS dataset and applied various models like SVM, Random Forest, and MLP Classifier. They achieved an overall accuracy of 79% using Random Forest. Rajwinder Singh [3] in their Acoustic Emotion Classification System achieved an overall accuracy of 64.15% using SV-Classifier on the RAVDESS dataset. We managed to increase the accuracy by 25.57% using the same dataset. Rohit Joshi et al [29] used NLP Techniques to see sentiments with the assistance of a tool Sentiment

Analyzer that extracts sentiments and is employed to get all the references for the given subject efficiently.

of these datasets are small, their combination becomes a good set for experimentation. EmoDB consists of speech files from 10 actors speaking 10 sentences each whereas

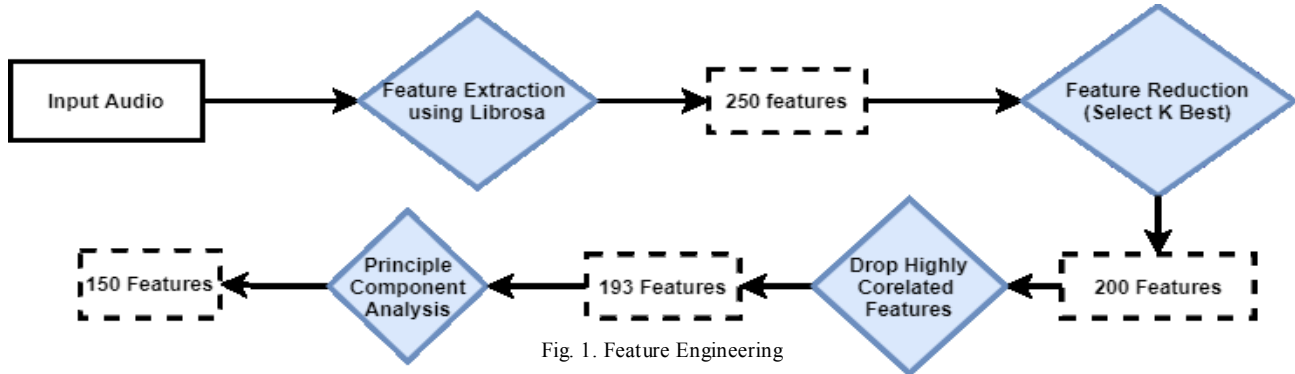


Fig. 1. Feature Engineering

III. Dataset

Collecting data for Speech Emotion Recognition has been a challenge since the majority of datasets are anonymized and don't reflect diverse accents. This research has experimented on sentences from datasets such as TESS, RAVDESS, and a custom dataset. Various important characteristics of these datasets such as the emotions present and their number of samples, class number, etc. are presented in Table 1. In each of these 20% of total samples has been used for testing, 10% for validation, and 70% for training.

Toronto Emotional Speech Set (TESS) contains a collection of sentences of two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions as shown in Table 1. There are a total of 2800 audio records. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) has a total of 7356 files, which requires a memory of 24.8 GB. RAVDESS contains speech by 24 actors, 12 of which are male and the remaining 12 are female. These actors vocalized two sentences in a North American accent. Each emotion is vocalized at 2 intensities: normal and strong. The audio consists of a resolution of 16 bits with a frequency of 48 kHz and is present in both .wav and .mp4 format. There are a total of 1440 + 1012 files (1440 Speech and 1012 Song files).

TABLE 1. EMOTION COUNT OF DATASETS

Emotions	TESS	RAVDESS	Custom Dataset
ANGRY	400	344	170
SAD	400	344	170
HAPPY	400	344	170
NEUTRAL	400	172	165
DISGUST	400	344	170
SURPRISED	400	344	-
FEAR	400	344	170
CALM	-	344	-
Total	2800	2580	1015

Our custom dataset consists of recordings from EmoDB [10] and SAVEE [11] along with noise additions. Since both

SAVEE consists of 4 actors, 7 emotions, and 480 utterances. A total of 1015 audio files with a frequency of 48 kHz are present. Our model classifies audio signals into emotions that can be affected by noise. Therefore altering the original audio signals with noises of varying signal-to-noise ratios and helps us in evaluating the effectiveness of our model under these adverse conditions. This dataset also gives us diverse accents and makes the system more robust.

IV. FEATURE EXTRACTION

In any speech recognition system, one of the most important tasks is to extract features:

- Identify the important parts which signify linguistic content along with emotion expressed
- Discard unimportant information like background noise etc.

A characteristic of audio files is their changing nature. But if we consider the audio on a short time scale it can be estimated to a constant signal i.e. a signal which does not change much by which we mean statistically constant. For this reason, we keep the sampling rate low and divide the audio into small frames of about 20-40 ms each. It is worth noting that it is important to choose a correct length of signals as if we continue to reduce the size then we get fewer and fewer samples which are not enough for a good prediction whereas longer samples mean that the audio signal is no longer constant and is again of changing nature.

Along with providing the fundamental tools necessary to retrieve data/features out of music, the Librosa package also allows and helps one to visualize the audio signals and also helps to extract the features out of the given audio file at different sampling rates using various techniques for signal processing. Librosa's load() helps us read one audio file at a time. It returns a time-series in the format of a 1D array (in the case of mono) and 2D array (case of stereo). The array is composed of amplitudes of audio signals. It also returns sr (which is the sampling rate) whose default value is set as 22400Hz. This package has been used in this research for extracting various features from the chosen dataset. There are majorly two kinds of features:

- Prosodic features*: Characteristics of speech that go beyond phonemes and constitute auditory qualities of audio signals are known as prosodic features [6] (or suprasegmentally phonology). We never really

think about the use and interpretation of prosodic features while communicating. These features appear with the sound which comes when speech is connected together. For example, Rhythm, Intonation, and stress. One of the prosodic features which we take into account for the purpose of our research is MFCC (Mel Frequency Cepstral Coefficients) [7] [8] .

- B. *Paralinguistic features*: Characteristics of speech that do not consist of spoken words are paralinguistic features. They emphasize what people mean rather than what people say. These are very important features as they are capable of changing the emotion and the meaning completely. For example pitch of audio signal, tone, expressions

Prosodic features have been used in this research which helps us accurately analyze [9] the audio signal and determine the emotion corresponding to the sound. The features under consideration for this research work are MFCC, MEL, Chroma, Tonnetz, and Contrast.

In humans like other mammals the sound produced is mainly dependent on the shape of the vocal tract. Determining this shape (of the vocal tract) will mean that the phoneme that is produced can accurately be represented. MFCC can accurately represent the envelope of a short-time power spectrum which in turn actually signifies the shape of the vocal tract. The relation of perceived frequency which is also known as a pitch to the actual frequency is calculated by Mel Scale. An accurate representation of audio signals where the spectrum was shown as 12 parts constitute 12 semitones of an octave. Chroma is related to the 12 distinct pitch groups. These features, known as pitch profiles, are tools for exploring music whose pitches are grouped into 12 classes, and whose tuning approximates to the equal-tempered scale. Contrast refers to the difference in the sound of the phonemes produced. Minimizing the features is important because it can result in a meaningful model. This approach helped us select the most important 150 features from a total of 250 audio features extracted (refer to Fig.1.). This selection of features helped reduce overfitting when applying decision tree-like models along with reducing the overall training time and, hence improving effectiveness. In Fig.1. Feature reduction has been performed in 3 steps: First selecting K-best features, then dropping highly correlated features using correlation matrix, and finally applying PCA(Principal Component Analysis) to extract the best possible 150 features. The last step helps keep as much variance by dimensionality reduction while feature selection.

V. PREDICTION MODELS

The problem at hand is a classification task. Given an audio file, we detect the emotion of the speaker. In this research the following baseline models are being used:

- A. *K-Nearest-Neighbors (KNN)*: KNN is used as a simple baseline model whose output will be used to compare different model's accuracy. The only hyperparameter in this model is the k value which is not easy to find. The value of k should be

optimal as a low value can result in a model in which noise has a high effect on the output whereas a larger value can overfit the model and the computation becomes expensive.

- B. *Support Vector Machine (SVM)*: SVM is a machine learning model which can train itself on complex nonlinear data similar to the model's output. It does so by the use of kernel functions to map the primary features in a higher dimension plane. After doing so the data can easily be classified by linear classifiers [12]. SVM works well with high-dimensional space, hence it suits our case.
- C. *Decision Trees*: A tree-like structure of features and their possible outcomes. Since they allow us to explore all possible options, decision trees are more likely to produce good results. In this algorithm, a particular feature is selected at each level, and based on the outcome the data is divided into multiple categories to the next level.
- D. *Multilayer Perceptron (MLP)*: Artificial Neural Network is implemented using Multi-layer Perceptron is the simplest neural network classifier. It includes an input layer, several hidden layers, and an output layer (preferably softmax). It is also known as a feed-forward network [13] [14], i.e. learns weights moving forward, calculates the loss at the output layer, and back-propagates rectifying/correcting the weights and biases learned. Parameter optimization is one of the biggest concerns when dealing with Multi-Layer Perceptrons [4]. MLP classifies noisy inputs concerning their similarity with pure inputs hence allows us to correctly predict noisy inputs making the system more efficient.

These baseline models are very diverse in their working and their comparative study helps us in analyzing all perspectives of our data.

Aggregator Models (Ensemble Learning): Aggregator models in machine learning operate on a similar idea. They combine the decisions from multiple models to improve the overall performance [15].

- A. *Max Voting Algorithm*: Predictions are made for each data point using multiple models in this algorithm. Each model's prediction is assigned a weight and the overall output depends on the weighted 'votes' derived from each of the individual models used. The majority prediction is the final output. Lower variance is provided in the final output predicted by the voting classifier over the individual baseline models. The max-voting classifier helps in reducing the dispersion or distribution of the prediction and model's efficacy. The two major ways in which max-voting ensemble helps us are: Giving better results than any individual baseline models in terms of performance and accuracy. Providing lower variance than any individual baseline models. Fig.2. shows the max voting architecture along with the weights associated with each baseline

model. The final prediction depends on each models' predictions proportionate to their weights.

B. Random Forest: This model is an ensemble model (follows a bagging technique) which includes different decision trees trained using different training parameters on different mini datasets produced out of the given input. In the end average is taken to enhance the overall confidence of the produced output. This also helps in controlling the over-fitting of the model.

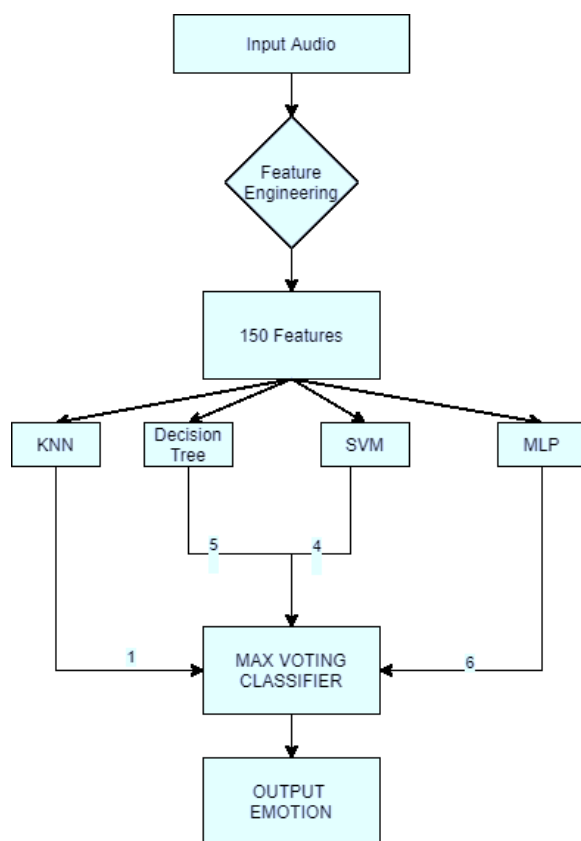


Fig. 2. Max Voting Architecture

C. XGBoost: XGBoost is a type of advanced Gradient Boosting [16]. It is also called 'regularized boosting' because it decreases over-fitting and makes the model more robust also increasing its performance.

VI. EVALUATION METRICS

The results of our prediction models are below, along with a summary of metrics for each model. A major issue was overfitting because of the relatively small size of the dataset. To reduce overfitting K-fold cross-validation was used. To optimize hyperparameters, a standard Grid Search CV method was used to find out the best-fit hyperparameters. Results are bound to vary (1-2%) unless seed values are fixed. The most commonly used methods to evaluate the performance of an emotion detection model are confusion matrix, precision, recall, F1-score, and overall accuracy.

Accuracy is calculated as the number of correctly classified values divided by the total number of values. Overall

accuracy is a good measure because all the emotions have the same weight and/or value. Through this research, we aim to build a speech emotion recognition system that minimizes the cost of misclassified data points (false positives and true negatives). This can be evaluated through two other metrics Precision and Recall.

Precision = True Positive / (Total Predicted Positive)

Recall = True Positive / (Total Actual Positive)

F1-Score = $2 * ((\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}))$

We consider F1-score as a better performance measure to seek a balance between Precision and Recall and to rule out uneven class distributions misclassifications.

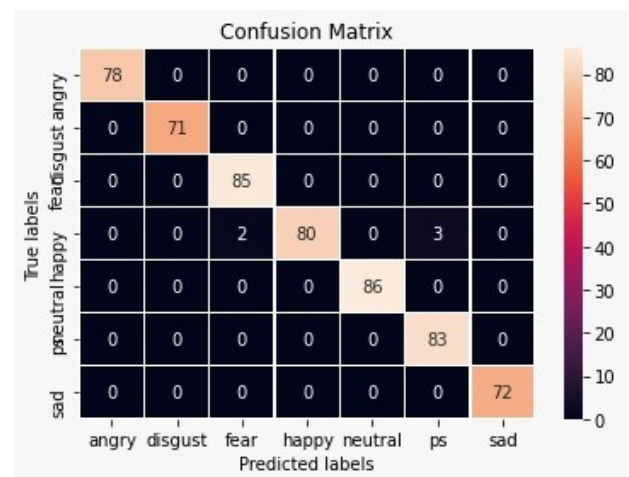


Fig. 3. TESS

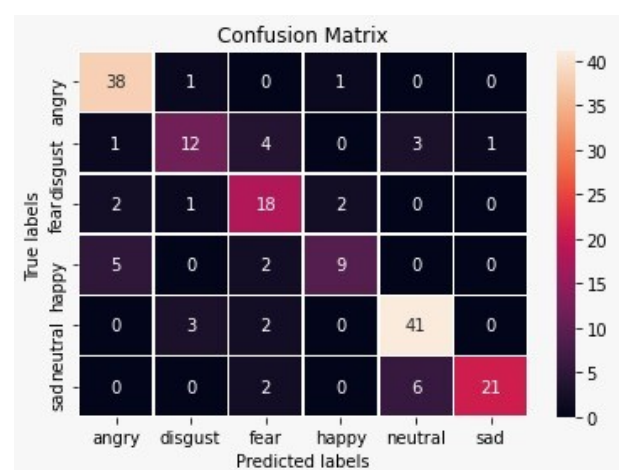


Fig. 4. Custom Dataset

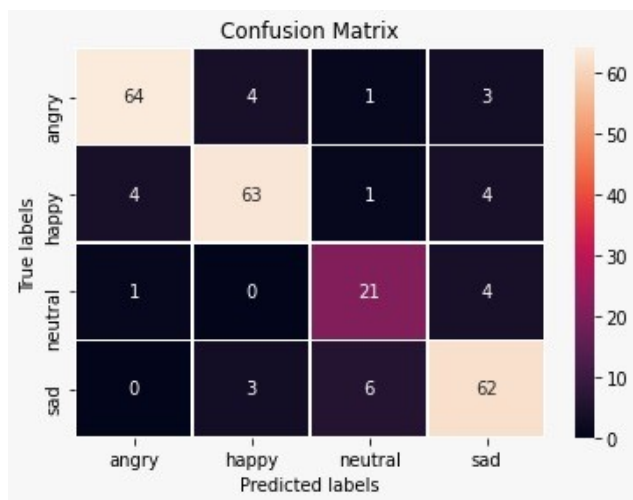


Fig. 5. RAVDESS

TABLE 2. ACCURACY FOR BASELINE MODELS

	TESS	RAVDESS	Custom DS
KNN	77.34%	60.13%	59.36%
SVM	89.66%	76.41%	62.58%
MLP	96.76%	82.39%	72.57%
Decision Trees	94.56%	80.35%	65.71%

TABLE 3. ACCURACY FOR AGGREGATOR MODELS

	TESS	RAVDESS	Custom DS
Random Forest	99.01%	83.05%	82.28%
Max Voting	99.12%	85.89%	75.28%
XG Boost	99.46%	89.62%	78.28%

TABLE 4. CLASSIFICATION REPORT FOR XGBOOST ON TESS

Emotion	Precision	Recall	F1 Score
Angry	1.00	0.99	0.99
Disgust	1.00	1.00	1.00
Fear	0.95	0.98	0.97
Happy	0.99	0.94	0.96
Neutral	1.00	1.00	1.00
Surprise	0.97	1.00	0.98
Sad	1.00	1.00	1.00

TABLE 5. CLASSIFICATION REPORT FOR XGBOOST ON RAVDESS

Emotion	Precision	Recall	F1 Score
Angry	0.88	0.88	0.88
Happy	0.82	0.84	0.83
Neutral	0.87	0.81	0.84
Sad	0.80	0.80	0.80

TABLE 6. CLASSIFICATION REPORT FOR RANDOM FOREST ON CUSTOM DATASET

Emotion	Precision	Recall	F1 Score
Angry	0.79	0.89	0.84
Disgust	0.64	0.64	0.64
Fear	0.63	0.61	0.62
Neutral	0.89	0.75	0.81
Sad	0.89	0.95	0.80
Happy	0.88	0.68	0.77

VII. RESULTS

The results from the experiments also show the efficiency of different ensemble models compared to the baselines, and the state of the art on TESS, RAVDESS, and custom datasets. The outcomes in Table. 2-6 show that out of all the baseline models (KNN, SVM, MLP, and Decision Trees) the best performing one is MLP. Whereas different ensemble models perform better in different cases given the varied data.

The results from Table. 4-6 exhibits the precision, recall, and F1 score values calculated for each emotion class in each dataset. For TESS (refer to Table 4.) these values are stabilized which allows us to attain distributed F1-score around 0.99 for all classes. The minimal difference of the F1 score shows us the robustness and efficiency of the model. The model is comparatively less accurate on classes 'happy' and 'surprised' and this result seems apt because the aforementioned emotions are known to be difficult to distinguish. For the RAVDESS dataset, (refer to Table 5.) precision and recall are stabilized which allows us to attain distributed F1 score around 0.84 for all classes.

In order to evaluate the effectiveness of the proposed methodology, we decided to experiment on our custom dataset. The custom dataset was tested on the same baseline and ensemble models and gave us an overall accuracy of 82.28% using Random Forest Classifier (refer to Table 3). It also gave us an F1 score of 0.78 which depicted the models' effectiveness in classifying 6 emotions accurately. According to the results 'anger' is the most accurately classified emotion in our custom dataset. This dataset contains different accents by a large number of actors which makes the prediction task even more challenging for the model. In addition to this, the noise addition allows us to transform the dataset into a more real-life scenario.

VIII. CONCLUSION

With this report, we evaluated and analyzed the efficiency of baseline models such as K-Nearest Neighbors' (KNN), Random Forests and Support Vector Machine (SVM) as well as aggregator models such as Max Voting, ADA Boost, XG Boost, and Gradient Boost for the task of emotion detection from audios using 3 different datasets: TESS, RAVDESS and our custom-made dataset. Apart from the difference in datasets, more complex aggregator models were used for the comparative study. This led to a speech emotion recognition system on TESS, RAVDESS, and our custom dataset with F1 scores of 0.99, 0.84, and 0.76 respectively.

The best classifier for the given task is XG Boost for TESS and RAVDESS whereas for our Custom Dataset Random Forest works best. The efficacies calculated from the experimental setup were very satisfactory keeping in

mind the challenging trade-off between accuracy and dataset size. It is also noted that anger and neutral are the two of the easiest to predict emotions (refer to Table 6) whereas fear emotion can pose a challenge to the model.

In the comprehensive study during the research, it was found that random forest makes a better prediction and gives better results than any baseline machine learning model. Though this model acts as a black box and leaves very little control to the user as to what the model does. All that we can do is tune parameters at random seeds. To optimize parameters, a standard Grid Search CV method was used to find out the best-fit parameters. XG Boost tends to overfit more than random forest but when provided a robust set of data points and conservative hyperparameters, it provides higher accuracy.

IX. FUTURE WORK

The future vision for this work is to analyze and include diverse features and come up with an aggregator model which can prove to be a robust approach towards handling such problems. Moreover, we can also use the information [2] hidden in the spectrogram and use various advanced neural network models (like CNN and RNN) to make our model even more accurate and real-time.

X. REFERENCES

- [1] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)." Scholars Portal Dataverse, doi: doi:10.5683/SP2/E8H2MF.
- [2] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS One*, vol. 13, no. 5, pp. e0196391, May 2018, [Online]. Available: <https://doi.org/10.1371/journal.pone.0196391>.
- [3] J.-D. Haynes and G. Rees, "Decoding mental states from brain activity in humans," *Nat. Rev. Neurosci.*, vol. 7, no. 7, pp. 523–534, 2006, doi: 10.1038/nrn1931.
- [4] I. Dnvsls, B. Lakshmi, H. Prasanna, C. Pavani, and G. Vandana, "European Journal of Molecular & Clinical Medicine Assessment of Patient Health Condition based on Speech Emotion Recognition (SER) using Deep Learning Algorithms," vol. 7, p. 2020.
- [5] F. Daneshfar and S. Kabudian, "Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm," *Multimed. Tools Appl.*, vol. 79, Jan. 2020, doi: 10.1007/s11042-019-08222-8.
- [6] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*. Macmillan Pub. Co., 1993.
- [7] Y. Ü. Sonmez and A. Varol, "New Trends in Speech Emotion Recognition," in *2019 7th International Symposium on Digital Forensics and Security (ISDFS)*, 2019, pp. 1–7, doi: 10.1109/ISDFS.2019.8757528.
- [8] L. Muda, M. Begam, and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," 2010.
- [9] I. Idris, M. S. Salam, and M. S. Sunar, "Speech emotion classification using SVM and MLP on prosodic and voice quality features," *J. Teknol.*, vol. 78, Dec. 2015, doi: 0.11113/jt.v78.6925.
- [10] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech." [Online]. Available: <http://www.expressive-speech.net/emodb/>.
- [11] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *Univ. Surrey Guildford, UK*, 2014.
- [12] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011, doi: <https://doi.org/10.1016/j.patcog.2010.09.020>.
- [13] G. Zhang, Q. Song, and S. Fei, "Speech Emotion Recognition System Based on BP Neural Network in Matlab Environment," in *Proceedings of the 5th International Symposium on Neural Networks: Advances in Neural Networks, Part II*, 2008, pp. 801–808, doi: 10.1007/978-3-540-87734-9_91.
- [14] Y. Huang, G. Zhang, and X. Xu, "Speech Emotion Recognition Research Based on Wavelet Neural Network for Robot Pet BT - Emerging Intelligent Computing Technology and Applications. With Aspects of Artificial Intelligence," 2009, pp. 993–1000.
- [15] T. G. Dietterich, "Ensemble Methods in Machine Learning BT - Multiple Classifier Systems," 2000, pp. 1–15.
- [16] A. Iqbal and K. Barua, "A Real-time Emotion Recognition from Speech using Gradient Boosting," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2019, pp. 1–5, doi: 10.1109/ECCE.2019.8679271.
- [17] D. Neiberg, K. Elenius, I. Karlsson, and K. Laskowski, *Emotion Recognition in Spontaneous Speech*. 2006.
- [18] I. Dnvsls, B. Lakshmi, H. Prasanna, C. Pavani, and G. Vandana, "European Journal of Molecular & Clinical Medicine Assessment of Patient Health Condition based on Speech Emotion Recognition (SER) using Deep Learning Algorithms," vol. 7, p. 2020.
- [19] R. Singh, H. Puri, N. Aggarwal, and V. Gupta, "An Efficient Language-Independent Acoustic Emotion Classification System," *Arabian Journal for Science and Engineering*, vol. 45, no. 4, pp. 3111–3121, 2020, doi: 10.1007/s13369-019-04293-9.
- [20] V. Hatzivassiloglou and J. M. Wiebe, "Effects of adjective orientation and gradability on sentence subjectivity," in *Proceedings of the 18th Conference on Computational Linguistics-Volume 1*, pp. 299–305, 2000.
- [21] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, pp. 174–181, 1997.
- [22] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10*, pp. 79–86, 2002.
- [23] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. A. Reis and J. Reynar, "Building a sentiment summarizer for local service reviews," in *WWW Workshop on NLP in the Information Explosion Era*, 2008.
- [24] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *Proceedings of the National Conference on Artificial Intelligence*, pp. 755–760, 2004.
- [25] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177, 2004.
- [26] S. Morinaga, K. Yamanishi, K. Tateishi and T. Fukushima, "Mining product reputations on the web," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 341–349, 2002.
- [27] Sungeetha, Akey, and Rajesh Sharma. "Transcapsule model for sentiment classification." *Journal of Artificial Intelligence* 2, no. 03 (2020): 163–169.
- [28] Mitra, Ayushi. "Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset)." *Journal of Ubiquitous Computing and Communication Technologies (UCCT)* 2, no. 03 (2020): 145–152.

- [29] Joshi, Rohit, and Rajkumar Tekchandani. "Comparative analysis of Twitter data using supervised classifiers." In 2016 International Conference on Inventive Computation Technologies (ICICT), vol. 3, pp. 1-6. IEEE, 2016.



A Review on Computation Methods Used in Photoplethysmography Signal Analysis for Heart Rate Estimation

Pankaj¹ · Ashish Kumar² · Rama Komaragiri¹ · Manjeet Kumar³ 

Received: 31 December 2020 / Accepted: 24 April 2021
© CIMNE, Barcelona, Spain 2021

Abstract

Photoplethysmography (PPG) sensor-enabled wearable health monitoring devices can monitor realtime health status. PPG technology is a low-cost, noninvasive optical method used to measure a volumetric change in blood during a cardiac cycle. Continues analysis of change in light signal due to change in the blood helps medical professionals to extract valuable information regarding the cardiovascular system. Traditionally, an electrocardiogram (ECG) has been used as a dominant monitoring technique to detect irregularities in the cardiovascular system. However, in ECG for monitoring cardiac status, several electrodes have to be placed at different body locations, limiting its uses under medical assistantship and in a stationary position. Therefore, to fulfill the market demand for wearable and portable health monitoring devices, researchers are now showing interest in the PPG sensor enable wearable devices. However, the robustness of PPG sensor-enabled wearable devices is highly deviating due to motion artifacts. Therefore before extracting vital sign information like heart rate with PPG sensor, efficient removal of motion artifact is very important. This review orients the research survey on the principles and methods proposed for denoising and heart rate peak detection with PPG. The efficacy of each method related to heart rate peak detection with PPG technologies was compared in terms of mean absolute error, error percentage, and correlation coefficient. A comparative analysis is formulated to estimate heart rate based on the literature survey from the last ten years on PPG technology. This review article aims to explore different methods and challenges mentioned in state-of-the-art research related to motion artifacts removal and heart rate estimation from PPG-enabled wearable devices.

1 Introduction

In today's world, monitoring cardiovascular health status for early diagnosis is one of the leading research areas. The heart rate study is a prominent approach to analyze cardiovascular health status during daily routine [1]. Due to its

simplicity, accuracy, and low cost, Photoplethysmography (PPG) is gaining importance and becoming an alternative approach to monitoring and studying vital body signs. PPG technology uses optical sensors and is popular due to its lightweight, fashionable, simplicity, and more importantly, it can be used as wearable devices like the smart fitness band [2]. Generally, abnormalities in the functionality of the heart are identified using heart rate and percentage of oxygen. Initially, PPG technology is used in pulse oximetry to monitor oxygen levels in the blood. Due to PPG's noninvasive nature, it has now become a standard of care in the operating theatre, intensive care unit [3]. Pulse oximetry has the flexibility to observe the body vitals both qualitatively and quantitatively. PPG is a noninvasive tool that can continuously monitor heart rate, respiratory rate, cardiac outputs, and blood pressure.

Even though PPG technology has many advantages, the major drawback is erroneous data in certain circumstances, mainly due to noise from motion artifacts. Hence the accuracy of PPG technology depends upon the suppression of noises [4].

✉ Manjeet Kumar
manjeetchhillar@gmail.com

Pankaj
er.pankaj08@gmail.com

Ashish Kumar
akumar.1june@gmail.com

Rama Komaragiri
rama.komaragiri@gmail.com

¹ Department of Electronics and Communication Engineering, Bennett University, Greater Noida, India

² School of Electronics Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India

³ Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India

1.1 Principle of PPG

PPG technology measures the change in blood volume in the tissue during a heart cycle using a light source. This volumetric measurement provides important information regarding the cardiovascular system. A PPG sensor mainly consists of two electronic components, a light emitter and a light intensity sensing component. Typically, LED is used as a light emitter and a photodetector to detect (sense) the change in light intensity [5]. A PPG pulse corresponding to one heartbeat includes the systolic and diastolic phases. During the systolic phase, the volume of blood in arteries is more; this is because during this phase heart contracts and pushes oxygen-rich blood to all the tissues and organs. The systolic phase causes more light is absorbed by the blood cells. Therefore the amount of light detected by the photodetector during the systolic phase is low. During the diastolic phase, the blood has flown back into the heart. Therefore, during the diastole phase, the light detected by the photodetector increases due to a decrease in the blood volume. Depending upon application and sensor placement, PPG can be used either in transmissive mode or in reflection mode, as shown in Fig. 1 [6].

When a photodetector and LED are placed on parallel sides of a finger to detect the transmitted light, this mode is known as a transmissive mode. In transmissive mode, the probe is in a projection that the photodetector and LED face each other with a layer of tissues between them [7]. Detection in transmissive mode depends upon transmission of light from body parts, so thin structures like the earlobe and finger are preferred in this mode. When both photodetector and LED are placed on the same side of a finger to detect the reflected light, it is a reflective mode. In reflection mode, both the sensors are placed next to each other with an approximate spacing of 3 cm. Therefore reflection mode can use anybody site like the forehead and wrist. Choice of the site to place PPG sensors depends on the patient's blood perfusion, comfortability of the subject, and application [8].

The role of the photodetector is to detect and quantify the light absorbed during pulsatile and non-pulsatile flow [9]. During pulsatile flow, light is absorbed by the change in

blood flow inside the arteries, which is synchronous with a heartbeat. During the non-pulsatile flow, light is absorbed by background tissues. Therefore, a photodetector detects the volumetric change in blood flow in arteries by detecting the light intensity difference [10]. Measurement of this change in light intensity thus helps to analyze the functionality of the heart.

A PPG signal mainly consists of AC and DC components. AC component in the PPG output waveform indicates the change in light intensity during the systolic and diastolic phase due to the blood in arteries [11]. The steady DC part of the PPG waveform indicates the light absorbed by tissues, skin, and bone, as shown in Fig. 2. Analysis of the DC component provides valuable information regarding venous blood flow, respiration, and thermoregulation. Variation in light intensity detected due to arterial blood flow is around 1% only, which provides information on the heart's functionality [12].

1.2 PPG Analysis Using Multiple Wavelengths

Light absorption during systolic and diastolic phases of a heart cycle follows Beer's law and Lambert's law, jointly known as Beer–Lambert's law. According to Beer's law, light absorbed by the blood is proportional to the concentration of oxygenated hemoglobin and deoxygenated hemoglobin. As per Lambert's law, light absorption is proportional to light penetration in the skin [13].

Therefore according to Beer–Lambert law, the amount of light absorption (A_λ) through a substance, given by Eq. (1) is directly proportional to the light absorber concentration (C), optical path length traversed by the light signal (L), and light absorptivity at a particular wavelength (ϵ_λ)

$$A_\lambda = \epsilon_\lambda CL \quad (1)$$

Body skin mainly consists of three layers, as shown in Fig. 3. Due to absorption, only light waves with a larger wavelength can penetrate through all three layers.

Therefore the measurement mode and the body vitals that need to monitor, determine the selection of LED. Oxygenated hemoglobin absorbs light at near infra-red (NIR)

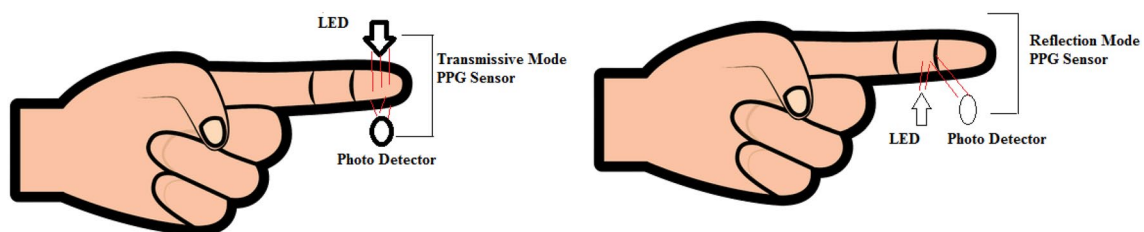


Fig. 1 Placement of sensor in transmissive mode PPG (left) and reflection mode PPG (right)

Fig. 2 Variation in light intensity during pulsatile and non-pulsatile flow

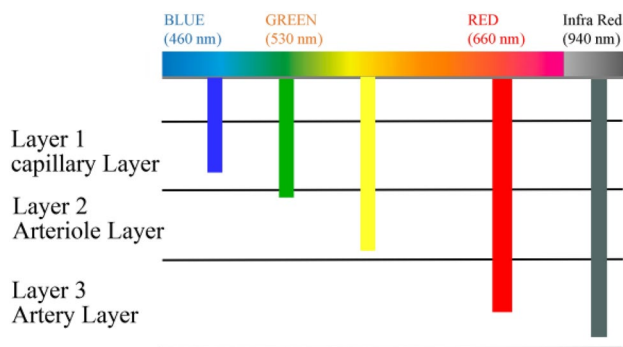
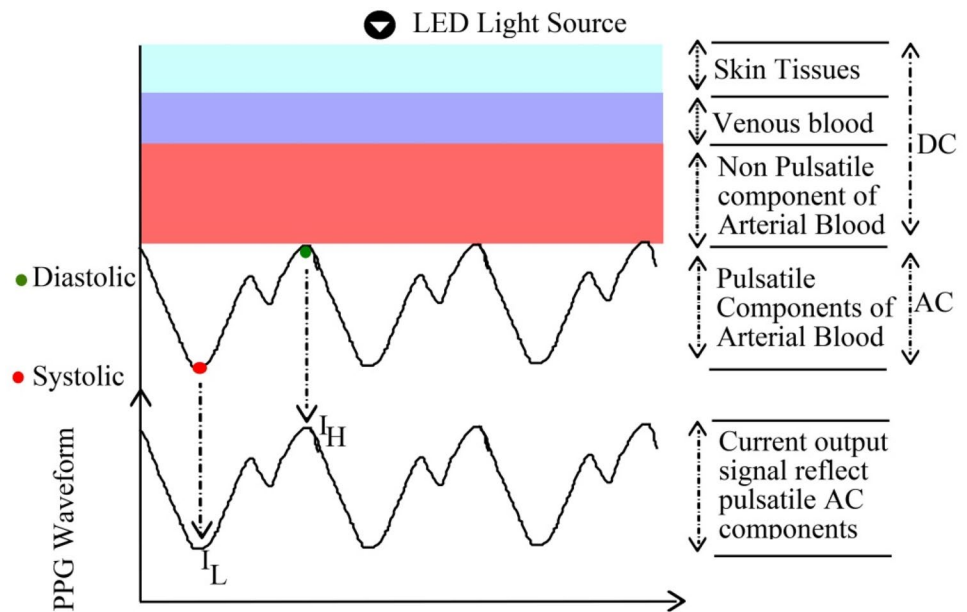


Fig. 3 A schematic representation of the penetration of light through the skin at different wavelengths

wavelength, while deoxygenated hemoglobin absorbs light at red wavelength. Hence, PPG employing NIR and red light wavelength LEDs and photodetectors is generally used for a clinical checkup to calculate the hemoglobin concentration. The effect of motion artifact on the PPG signal also depends on the wavelength of the light used. Longer wavelength light like infra-red light gets affected more due to motion artifact as it penetrates deep inside the tissue.

On the other hand, the light of a shorter wavelength (green light) is generally free from motion artifacts. Light at a shorter wavelength penetrates less inside the body tissue. Thus, to mitigate the effect of motion artifacts and the absorption of light by body tissues, PPG based on multi-wavelength optical sensors has been proposed to detect blood flow variations at different skin depths [14].

The light emitted by the diode is absorbed by tissues, and the amount of absorption in terms of detected light intensity is determined by photodetector [15]. When used as a pulse

oximeter, PPG uses two LEDs of a different wavelength. One LED emits light in the red spectrum around 660 nm, at which light absorption due to deoxyhemoglobin is greater than that of oxyhemoglobin. Another LED emits light in the infrared spectrum at a wavelength of 940 nm, at which oxyhemoglobin absorbs more light than deoxyhemoglobin. Accurate information on the blood circulation during a heart cycle is obtained by fixing the wavelength of LEDs between 660 and 940 nm.[16]. Finally, a Microprocessor unit analyzes the light absorption at each wavelength to determine the concentration of oxyhemoglobin and deoxyhemoglobin.

The rest of the paper is organized as follows: Estimation of heart rate from PPG is outlined in Sect. 2; Sect. 3 describes different methodologies proposed to date to remove motion artifacts. Section 4 highlights different datasets available for heart rate estimation using PPG. A literature survey based on different algorithms and methods proposed for heart rate identification is presented in Sect. 5. Challenges, and Discussion are drawn in Sects. 6, and 7 summarizes the work.

2 Heart Rate Estimation Using PPG

Realtime estimation of heart rate using a wearable device is one of the demanding applications in the health care system for the early diagnosis of cardiovascular diseases. Heart rate is the average number of times a heart beats per minute. Fluctuation in the time interval between subsequent heartbeats in milliseconds is called heart rate variability (HRV). Heart rate and HRV are standard markers for detecting health status. In a human body, the behavior of sympathetic and the parasympathetic branches of the autonomic nervous

system (ANS) indicate the status of HRV [17]. The sympathetic branch is related to the acting condition of the body, and the parasympathetic branches are related to the resting and digesting phase of the body. Depending upon day-to-day activities, the brain processing signal through ANS to the other parts of the body, through which the body can either react or stay relaxed. The human body tackles all kinds of signals received through the ANS system in a balanced way [18]. However, if a body persistently involves an unhealthy diet, irregular sleep, stress, and laziness, the balance between the ANS system's branches may be disturbed.

A subject with a high HRV means that the ANS system is in balance and responding to both sympathetic and parasympathetic inputs. Low HRV indicates that the subject is working under stress or fatigue and sympathetic branches dominate parasympathetic branches. A body with high HRV has a healthy status, but a low HRV indicates more stress, due to which the risk of cardiovascular disease may increase. Therefore from the last few years, HRV analysis has become a valuable tool for the early diagnosis of cardiovascular disease. Therefore both heart rate and HRV are used to measure cardiovascular health status. Heart rate and HRV are determined by measuring the volumetric change in blood during a heart cycle by passing the light through the skin. The PPG output waveform shown in Fig. 4 depicts the fluctuation in light absorption during a systolic and diastolic phase of a heart. When the heart contracts, the volume of blood flow increases, which increases the hemoglobin; therefore, the light absorption due to increased hemoglobin also increases—the amount of light detected by the detector decreases. In the dilation phase, when the blood volume

reduces, the hemoglobin decreases. Therefore, the amount of absorbed light decreases, hence the light detected by the photodetector increases. As a result, a pulsatile waveform in response to a cardiac cycle is observed as a PPG waveform [19].

The volumetric change of blood in tissue is synchronous to the heartbeat, which is used to estimate the heart rate. A PPG waveform mainly consists of four points O–S–N–D. As shown in Fig. 4, the S-point (Systolic Peak) represents the peak value in a PPG signal. The calculation of the Peak-to-Peak interval of consecutive PPG signals (S–S) provides information on the heart rate. The Peak-to-Peak interval correlates closely with the R–R interval in an ECG waveform. Analysis of pulse interval (O–O) provides information about HRV.

For the estimation of heart rate and HRV using PPG, it is necessary to analyze different properties of pulsatile PPG waveform like time interval between two consecutive systolic peaks (t_{S-S}), systolic peak amplitude (P_s), and the amplitude of diastolic Peak (P_d) [20]. After calculating the accurate value t_{S-S} , the instantaneous heart rate due to a single heartbeat is calculated using Eq. (2).

$$HR_i = \frac{60}{t_{S-S}} \quad (2)$$

For a time window H , the heart rate is calculated by using Eq. (3).

$$HR_{true} = \frac{60H}{t_{S-S}} \quad (3)$$

PPG waveform recorded from a healthy subject consists of three feature points, systolic Peak (S), diastolic Peak (D), and dicrotic notch (N). However, some of the feature points may be missing in some PPG waveforms. As the morphology of a PPG wave depends on age, gender, and health status, some of the feature points may miss the recorded PPG signal. The accuracy of cardiovascular functionality estimation depends on the accurate analysis of these features. The first derivative and second derivative of a PPG signal help identifying the PPG feature points [21]. By analyzing the features extracted from these three waveforms, namely the PPG signal, the 1st derivative of the PPG waveform, and the 2nd derivative of the PPG waveform, adequate information related to cardiac function can be processed [7]. A schematic representation of these three waveforms is shown in Fig. 5. It is mandatory to detect feature Point S in PPG signals to detect the heart rate accurately. It is important to note that reliable estimation of the heart rate and HRV is only possible if the Point-S in the PPG signal is detected.

In a healthy subject, the subsequent cardiac cycle's morphological structure possesses almost similar properties as its predecessor. A missing feature point indicates

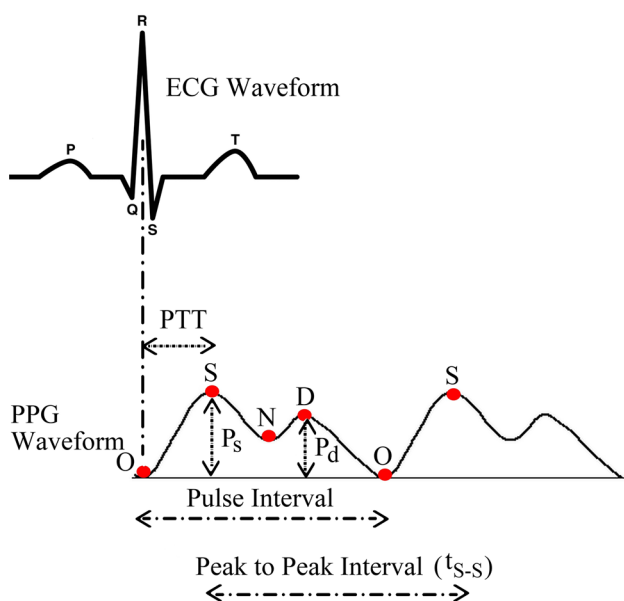


Fig. 4 Different feature points related to the PPG waveform

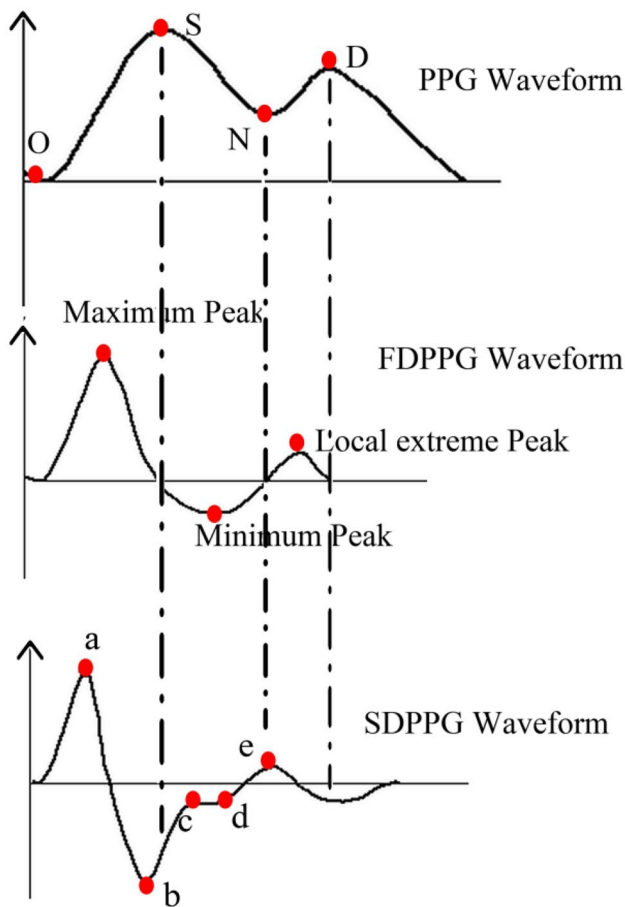


Fig. 5 Schematic representation of a PPG waveform, its First Derivative PPG (FDPPG) and Second Derivative PPG (SDPPG)

a sign of abnormality in cardiac function. To accurately locate feature points in a PPG signal, the derivatives of a PPG signal are used.

Between the first and second derivatives of a PPG signal, the second derivative is widely-used to locate the missing feature points. A PPG waveform, along with its first derivative and second derivative, is shown in Fig. 5. Normalized amplitude values, namely b/a , c/a , d/a , e/a , can be used to detect arterial stiffness [22].

In elderly subjects, the normalized amplitude b/a increases and other normalized amplitudes decrease. Analysis of change in amplitude value is used to measure the subject's cardiovascular age index, studied by aging index (AGI) as in eq. (4) [23].

$$AGI = \frac{b - c - d - e}{a} \quad (4)$$

Moreover, different intervals between different peaks from the second derivative of the PPG signal are used to identifying a subject with abnormalities [24].

Using the correlation between consecutive heartbeats within a time window, pulse transit time (PTT) and pulse wave velocity (PWV) provide vital information about heart rate and HRV. PTT is defined as the time required by an arterial pulse wave to travel from an aortic valve to a body site perfuse by optical light [20]. In reference to the ECG waveform, PTT is the time interval between the R-wave peak and any feature point on the PPG signal. PWV is used to measure the heart rate and heart rate variability. PWV is the velocity of a pressure wave when the blood flows through arteries. PWV has an inverse relation with PTT as given by Eq. (5). Therefore, PTT and PWV form a noninvasive method to analyze cardiac functionality.

$$PWV = \frac{D}{PTT} \quad (5)$$

Here D is the vessel length through traversed by a pressure pulse.

The PPG signal analysis is also affected by various noises like motion artifacts, variation due to baseline drift, and ambient light noise due to sensor position variation. Out of these noises, motion artifact has a significant effect on heart rate analysis as the frequency of the motion artifact lies inside the required heart rate information band. Hence, accurate heart rate peak identification when the PPG sensor is in motion is challenging. For accurate heart rate estimation, the effect of motion artifact in the PPG signal must be removed. The following section describes the motion artifacts reduction techniques and their properties proposed to date.

3 Motion Artifacts Removal Techniques

Accurate and reliable peak detection with wearable PPG sensors for heart rate estimation has become a demanding application in the health care industry. Physical motion during daily activities drastically reduces the accuracy of heart rate identification using a PPG sensor. In this section, several approaches proposed to date to mitigate the effect of motion artifacts from raw PPG signals are summarized.

Due to physical movement, sensor light passes from the body tissue deviates from its path, which provides erroneous data. The frequency spectrum of motion artifact is greater than 0.1 Hz and usually lies inside the heart signal's desired spectrum [25]. Hence, motion artifact is a leading noise source that influences various factors in the PPG signal analysis, potentially limiting the PPG sensor's usage to study and monitor the cardiac system information for health monitoring. Thus, the suppression of the noise spectrum from PPG signals is one of the leading research topics in the healthcare industry.

In [26], decomposition-based independent component analysis (ICA) is proposed to suppress the motion artifacts

components from a PPG spectrum. Moreover, the ICA-based approach provides reliable output only if noise and information signal possesses a mutually exclusive spectrum. In the realtime analysis of cardiac health monitoring, independent spectral conditions are not met. Thus, the efficiency of the ICA approach becomes suboptimal. Another widely used approach to suppress motion artifacts is adaptive filtering. As motion artifact behavior is random, a fixed coefficient filtering process is not suitable. Therefore adaptive filtering based motion artifact removal was proposed in [25]. However, the adaptive filtering performance depends upon the nature of the reference noise signal [27]. Therefore, adaptive filters can only provide reliable noise suppression when the correlation between the reference accelerometer signal and the motion spectrum is high, which is not possible in realtime.

Moreover, the high computational complexity of the adaptive filter limits their usage in wearable PPG. To make the system computationally efficient and to reduce the requirement of an additional accelerometer, deep learning convolutional neural network (CNN) is proposed to detect the noise in a PPG signal. The proposed CNN-based PPG signal classification in [25] uses a 1-D CNN network and provides the flexibility to the user to select any PPG segment of 5-s duration to detect motion artifacts [28]. CNN network can automatically extract the features by classification, thus reduces the need for threshold setting and segmentation. The correlation feature between both left and right hands was used to detect motion artifacts without an additional accelerometer sensor [29]. Since the nature of the PPG signal is nonlinear and varies between subjects, the proposed work in [28] uses the artificial neural network approach to analyze PPG signal characteristics to detect motion artifacts, and by using ANFIS based algorithm, the lost part of the PPG signal due to noise is retrieved. In [30], a method based on neural-network-based classification was proposed to detect the PPG signal accurately.

Based on the penetration depth of different light wavelengths, one more approach to removing motion artifacts without using accelerometer sensors was proposed [31]. A shorter wavelength green light source to estimate heart rate and a longer-wavelength infrared light source to provide a reference noise signal is used. Moreover, light sources with different wavelengths also detect noise that arises due to micro motions. In [29], to reduce the computational complexity, a multi-sensor method with multiple wavelengths is proposed to study the infected frame instead of analyzing the whole PPG signal. As a PPG signal is of pulsating nature, the most pulsating signal is used to extract a clean PPG signal. Multi-wavelength (Red, Green, Infrared) have different penetration depths. ICA approach is used to extract the pulsatile component. A method based on the fusion of signals from multiple sensors was proposed in [32] to remove

motion artifacts from the PPG signal. The method in [33] extracts the reference signal through the PPG signal, thus reducing the hardware cost.

Most of the proposed methods related to motion artifact removal deal with simple exercise or limited physical movement. Therefore, to remove strong-motion artifacts, discrete wavelet signal decomposition and thresholding-based approaches are proposed to remove the noise spectrum from the PPG signal [34]. A decomposition-based empirical mode decomposition (EMD) approach was implemented to extract the correct PPG segment from the corrupted PPG signal. A modified nonlinear approach named ensemble empirical mode decomposition (EEMD) was proposed in [35] to reduce motion artifacts from the PPG signal to resolve the mode mixing problems that arise during time–frequency distribution. In the EEMD method, reference noise is added to decompose the given PPG signal into IMF, without any prerequisite selection criterion on window width.

The potential of the principal component analysis (PCA) approach was combined with the EEMD method for accurate extraction of vital sign information from the PPG signal. Generally, motion artifact removal techniques are either based on time analysis or frequency analysis, which possess their inherent limitations. Therefore time–frequency based approach was proposed in [35]. However, time–frequency based approaches failed to provide reliable results when the nature of motion noise is periodic and strong. In that case, the extraction of a clean PPG signal becomes very difficult. Therefore the demand for accurate and reliable motion artifact removal methods for analyzing accurate vital signs is still an important research topic.

4 PPG Database

There are several data sets publicly available to test proposed algorithms. Table 1 highlights all the publicly available databases recorded with PPG-enabled wrist-worn devices. One of the most standard datasets is IEEE signal processing competition (SPC) 2015. IEEE SPC 2015 dataset was first used in [36]. IEEE SPC 2015 dataset consists of recordings from 23 subjects, in which the first 12 subjects have undergone simple physical exercises like walking (IEEE SPC-12 Training). The subjects numbered 13–23 performed arm exercises to introduce some motion noise (IEEE SPC-11 Testing). Two PPG signals and three-axis accelerometers are used on the wrist while recording the PPG. To test the efficacy of the work, the IEEE SPC dataset also recorded ECG signals while the subject is at rest. One more publicly available recent dataset is named PPG dataset for heart rate estimation in *daily life activities* (PPG DaLiA) [37], which is introduced to overcome the limitation on low physical activity used while recording the IEEE SPC dataset. In PPG DaLiA, fifteen subjects have

Table 1 Summary of publically available PPG databases

Database	Wristband embedding sensor	Description
IEEE Signal Processing Competition (IEEE SPC-12) –Training [36]	2-Channel PPG (green LEDs wavelength: 609 nm), 3-axis accelerometer	Twelve male subjects aged 18–35 years ECG (HRreference) recorded simultaneously from the chest Sampling frequency: 125 Hz
IEEE Signal Processing Competition (IEEE SPC-11) –Testing [36]	2-Channel PPG (green LEDs, wavelength: 609 nm), 3-axis accelerometer	Eleven subjects aged 19–58 years ECG (HRreference) recorded simultaneously from the chest The sampling frequency is 125 Hz
IEEE Signal Processing Competition (IEEE SPC-23) –Testing + Training [38]	2-Channel PPG (green LEDs, wavelength: 609 nm), 3-axis accelerometer	IEEE SPC-23 dataset includes both IEEE SPC Training and Testing dataset
IEEE Signal Processing Competition (IEEE SPC-22) –Testing + Training [38]	2-Channel PPG (green LEDs, wavelength: 609 nm), 3-axis accelerometer	IEEE SPC-22 dataset does not consider subject number 13 from the IEEE SPC-23 dataset
Wrist PPG during exercise [39]	1-Channel PPG (green LEDs, wavelength: 510 nm) A low noise 3-axis accelerometer A wide-range 3-axis accelerometer 3-axis gyroscope for orientation	Out of nine subjects, only one subject participated in all exercise ECG (HRreference) is recorded simultaneously from the chest
Wrist PPG during walking/running [40]	3-Channel PPG (green LEDs, wavelength: 525 nm), 3-axis accelerometer, 3-axis gyroscope	24 subject with an average age of 26.9 ± 4.8 year ECG signal captured using Holter device The sampling frequency is 50 Hz
PPG dataset for heart rate estimation in <i>daily life activities</i> (PPG DaLiA) [37]	4 LEDs (two green and two red) three-axis accelerometer	15 subjects aged 21–55 years The sampling frequency of 64 Hz

undergone physical activities that are similar to daily activities. PPG DaLiA dataset is specially designed to identify heart rate under a motion noise environment. Besides this real-life exercise feature, the PPG DaLiA dataset has limited information on the age group.

The limitations posed by the accelerometer during recording on the accuracy of the PPG data set are improved by introducing a gyroscope along with accelerometers in the PPG signal recorder. During the signal recordings, the subjects underwent physical exercise activities like walking, running on a treadmill [39].

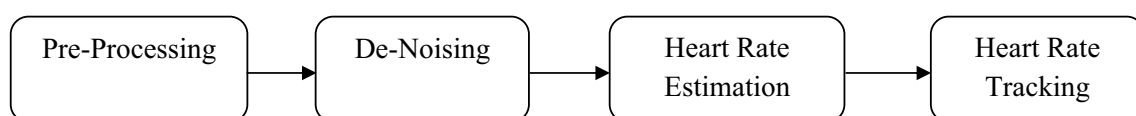
5 Literature Survey Based on Heart Rate Estimation

Accurate estimation of heart rate is essential to detect any abnormalities in body function. The reliability of heart rate estimation is always affected due to the presence of motion artifacts. Therefore denoising motion artifacts and correct heart rate estimation in realtime are current research areas while designing smart wearable healthcare devices. This

motivates researchers to develop and implement a faster and reliable way to identify the correct heart rate during physical activities. The majority of the proposed work to date related to heart rate detection follows a four-step approach, as shown in Fig. 6.

Input to the preprocessing stage consists of sensor information like accelerometer, PPG, and gyroscope [39]. The role of the preprocessing stage is to filter out undesired frequency spectrum (out of the desired window) by using bandpass filters. For reliable and correct estimation of heart rate, the role of the denoising stage is crucial. Using a reference noise signal (output of the accelerometer sensor) while recording a PPG signal helps the denoising algorithm remove the noise spectrum from the information signal. After removing motion artifacts, by identifying the correct peak, the heart rate is estimated in *stage-3*. A post-processing stage known as the heart rate tracking stage is used to provide exact information. The algorithms proposed to date showed a tradeoff between complexity and accuracy.

This literature review summarizes the research articles related to heart rate estimation using the

**Fig. 6** Flowchart indicating the four main stages in heart rate estimation

Photoplethysmography (PPG) method. The heart rate estimation performance is studied in the literature by evaluating average absolute error (AAE), absolute error percentage (AEP), and Pearson correlation coefficient. AAE and AEP are computed using a reference ground truth heart rate value, estimated using ECG. The performance of the heart rate algorithm is estimated using the following indexing. $HR_{true}(i)$ represents the ground truth ECG heart rate in the i^{th} time window, and $HR_{est}(i)$ is the estimated heart rate value using the proposed method. The output of each proposed work was analyzed and compared in terms of mean absolute error, error percentage, and Pearson correlation coefficient.

The average absolute error is calculated by using Eq. (6).

$$AAE = \frac{1}{W} \sum_{i=1}^W \left| \sum_{i=1}^W HR_{est}(i) - HR_{true}(i) \right| \quad (6)$$

For a total number of windows W , the average absolute error percentage (AEP) is calculated using Eq. (7).

$$AEP = \frac{1}{W} \sum_{i=1}^W \frac{|HR_{est}(i) - HR_{true}(i)|}{HR_{true}(i)} \times 100 \quad (7)$$

The other set of parameters used in some works in the literature include accuracy (ACC), sensitivity (SCC), and specificity, given by Eqs. (8)–(10), respectively.

$$\text{Accuracy(ACC)} = \frac{(TP + TN)}{(TP + TN) + (FP + FN)} \quad (8)$$

$$\text{Sensitivity(SCC)} = \frac{TP}{TP + FN} \quad (9)$$

$$\text{Specificity} = \frac{TN}{FP + TN} \quad (10)$$

In Eqs. (8)–(10) true positive (TP) is the number of segments that are classified correctly. NP is the true negative, which shows the number of segments affected due to motion artifacts. False-positive (FP) indicates the segment which is affected but also classified incorrectly. FN false negative shows segment, which is artifact affected. Table 2 summarized all the techniques proposed, along with their evaluation results.

An algorithm to minimize the motion artifact effect on heart rate estimation is proposed in [69]. Due to lower complexity and normalization features, the Normalized Least Mean Square (NLMS) adaptive filter is used to remove motion artifacts. After removing the motion artifact, the heart rate is calculated from the autocorrelation-based fundamental period extraction unit. A threshold-based approach is used as a post-processing step to extract heart rate information. The proposed algorithm extracts heart rate with a

correlation of more than 0.98. The accuracy of denoising using an adaptive filter always depends on the accuracy of the reference noise signal recorded using the accelerometer. An algorithm named signal decomposition for denoising, sparse signal reconstruction for high-resolution spectrum estimation, and spectral peak tracking (TROIKA) [36] is proposed in a wearable PPG device that does not require a reference signal to estimate heart rate. TROIKA technique for heart rate estimation consists of a three-step process. *Step 1* consists of the signal decomposition method to denoise the motion artifacts components. *Step 2* used the sparsity-based spectrum estimation approach to estimate heart rate. *Step 3* is a post-processing step to track and verify the desired peak related to heart rate. An AAE of 2.34 ± 0.82 BPM and AEP of 1.80% was calculated with IEEE SPC 12 candidate dataset. TROIKA approach has shown good results during physical activities also. To further improve the performance [41], proposed an approach named *joint sparse spectrum reconstruction* (JOSS), which follows a modified procedure to improve the accuracy of previous work TROIKA. It utilizes the PPG signals and acceleration signals jointly for heart rate spectrum estimation under the multiple measurement vectors model. Noise due to motion from PPG signal is removed by spectral subtraction instead of signal decomposition. Selection and verification of peak were used as a post-processing step to track heart rate. The authors calculated an AAE of 1.28 ± 2.61 BPM and an AEP of $1.01\% \pm 2.29\%$ with the proposed technique. JOSS provides a reduction in the error compared to TROIKA implemented on the IEEE SPC 12 candidates' dataset. Despite the improvement in the result recorded with [36, 41], both approaches faced a limitation in terms of computational complexity. A novel method called *spectrum subtraction, peak tracking, and post-processing* (SPECTRAP) is proposed to reduce the computational complexity [43]. Asymmetric least squares spectrum subtraction approach is used to denoise the PPG signal. Instead of using heuristic rules based spectral peak tracking, a Bayesian decision theory was used for reliable estimation of heart rate. An AAE of 1.50 ± 1.95 BPM and AEP of $1.12 \pm 1.47\%$ were calculated with IEEE SPC 12 candidate dataset. SPECTRAP showcased the reduction in computation complexity at the expense of an increase in the AEP. Using random forest-based spectral peak tracking algorithm, a method to reduce computational complexity by reducing AAE is proposed in [46]. The power spectral density of the PPG data segment and the accelerometer are compared to remove motion artifacts. Using the method in [46], an AAE of 1.23 ± 0.80 BPM with IEEE SPC 12 candidate dataset and 1.65 ± 1.56 BPM with IEEE SPC 22 candidates' dataset were showcased with a reduced computational complexity with and reduced APE.

Like TROIKA, a method to estimate the heart rate by using spectral peak tracking is proposed [42]. The spectral tracking method involves multiple heart rate trajectories,

Table 2 Summary of heart rate estimation techniques using wrist PPG technology

Author	Year	Proposed algorithm	Signals used	Data set	Pre-processing	Denoising technique	Heart rate estimation technique	Heart rate tracking technique	Results
Zhang et al. [36]	2015	[TROIKA] signal decomposition based denoising and sparse signal reconstruction based heart rate estimation	2 Channel PPG and 3-axis Acceleration Signal	IEEE SPC-12 (Training)	2nd order BPF (0.4–5 Hz) Signal downsampled to 25 Hz	SSA and signal decomposition	Sparsity based spectrum estimation	Spectral peak tracking and verification	AAE = 2.34 ± 0.82 BPM AEP = $1.80\% \pm 0.992$
Zhang [41]	2015	[JOSS] joint model for heart rate estimation	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	2 nd order Butterworth BPF from 0.4 to 4 Hz. Down sampled 125–25 Hz	MMV model-based sparse signal recovery and spectral subtraction	MMV model-based sparse signal recovery and spectral subtraction	Peak selection, verification and discovery	AAE = 1.28 ± 2.61 BPM AEP = $1.01 \pm 2.29\%$ $r = 0.993$
Murthy et al. [42]	2015	MISPT: Multiple initializations based on heart rate peak tracking	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	N/A	ANC	MISPT (Trajectory strength)	N/A	AAE = 1.11 ± 2.33 BPM
Sun and Zhang [43]	2015	[SPECTRAP] based on asymmetric least square and Bayesian decision theory	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	BPF (0.5–3 Hz)	ALS-SS algorithm	Spectral peak selection approach	Moving average smoothing filter	AAE = 1.50 ± 1.95 BPM AEP = $1.12\% \pm 1.47\%$ $r = 0.995$
Zhang et al. [44]	2015	Decomposition base reduction of motion noise	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	2nd order Butterworth filter (0.4–5 Hz)	Spectrum subtraction	Spectral peak selection	Window's based	AAE = 1.83 ± 1.21 BPM AEP = $1.40\% \pm 0.989$
Fallet and Vesin [45]	2015	Time-domain NLMS denoise algorithm	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	2nd Order BPF (0.4–5 Hz) Signal Down sampled to 35 Hz	NLMS adaptive filtering	Adaptive frequency tracking	OSC-ANF	AAE = 1.71 ± 0.49 AEP = $1.41\% \pm 0.994$
Salehizadeh et al. [46]	2016	SPAMA: Spectral filtering based removal of motion artifact and heart rate estimation	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	BPF (0.5–3 Hz) Down sampled to 31.25 Hz	Spectral Filtering	Largest peak comparing algorithm	Largest peak comparing algorithm	AAE = 0.89 ± 0.6 BPM AEP = $0.65 \pm 0.4\%$ $r = 0.98$

Table 2 (continued)

Author	Year	Proposed algorithm	Signals used	Data set	Pre-processing	Denoising technique	Heart rate estimation technique	Heart rate tracking technique	Results
Mashhadi et al. [47]	2016	SVD to decompose acceleration signal for denoising and heart rate estimation using the iterative method with a threshold value	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	BPF (0.4–5 Hz) Signal	SVD	IMAT: Iterative method with adaptive thresholding	Peak selection and thresholding	AAE = 1.25 ± 0.6 BPM AEP = 0.99% $r = 0.999$
Khan, et al. [33]	2016	Time–frequency hybrid approach (EEMD and RLS) for denoising	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	BPF (0.6–1.7 Hz) Signal down-sampled to 25 Hz	EEMD and RLS	periodogram based approach	Peak Tracking and thresholding	AAE = 1.02 ± 1.79 BPM, AEP = 0.79% $r = 0.996$
Ye et al. [35]	2016	Time–frequency hybrid method combines adaptive filter with SSA based signal decomposition approach	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	BPF (0.4–5 Hz) Signal	RLS + SSA	Sparsity based spectrum estimation	Spectral peak tracking and thresholding	AAE = 1.16 BPM AEP = 0.93% $r = 0.9975$
Fujita et al. [48]	2016	PARHELIA: Particle filter-based heart rate estimation under strong influence from motion artifacts.	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	BPF (1 Hz–3.5 Hz)	NA	Spectral analysis	Particle filter	AAE = 1.17 BPM AAP = 0.43% $r = 0.995$
Chowdhury et al. [49]	2016	MURAD: Multiple reference noise signals based denoising algorithm	2 Channel PPG and 3 axis Acceleration Signal and the difference between 2 PPG signal	IEEE SPC-12 (Training)	Sgolay filter	RLS based adaptive filtering	Periodogram based	spectral peak tracking and verification	AAE = 0.9726 ± 1.831 BPM AEP = $0.76\% \pm 1.5\%$ $r = 0.9972$
Dubey et al. [50]	2016	HSUM: Harmonic sum model used to remove motion artifact and heart rate estimation	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	NA	Sum of two different truncated Fourier series models	3-point median filtering	Use frequency locations of peak of the STFT magnitude	AAE = 0.73 ± 0.83 BPM $r = 0.9978$

Table 2 (continued)

Author	Year	Proposed algorithm	Signals used	Data set	Pre-processing	Denoising technique	Heart rate estimation technique	Heart rate tracking technique	Results
Dao et al. [51]	2016	SVM classifier used to evaluate feature for heart rate estimation	2 Channel PPG	Four different dataset	6th order IIR BPF with cut-off 0.1 Hz and 10 Hz	SVM	Time-frequency spectrum	Reference usability index	NA
Farhadi et al. [52]	2016	Used spectral property of heart rate and cumulative feature of all frequency to detect the peak	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-23 (Training + Testing)	NA	Spectral division	Cumulative spectrum calculation	Lazy tracker algorithm	AAE= 1.19 BPM
Temko [53]	2017	WFPV: Wiener filter to attenuate motion artifacts and phase vocoder to refine the heart rate estimation	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	4th order Butterworth BPF (0.4–4 Hz) downsampled to 25 Hz	Wiener filter	Phase vocoder	Online post-processing and Offline post-processing	AAE= 1.02 BPM AEP= 0.81 % $r = 0.997$
Zhao et al. [54]	2017	SFST: Combination of STFT and spectral analysis for accurate heart rate estimation	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	Zero-phase forward and reverse digital IIR filter	STFT	Signal segmentation and spectral peak tracking	Moving average filter	AAE= 1.06 ± 0.69 BPM AEP= $0.94 \% \pm 0.53 \%$ $r = 0.991$
Tariqul et al. [55]	2017	Multi-stage cascaded adaptive RLS filtering combined with SSA approach	2 Channel PPG, average of two-channel and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	BPF (0.4–3.5 Hz)	Multi-stage RLS adaptive filtering and singular spectrum analysis	Spectral Peak detection	Weighted moving average based tracking algorithm	AAE= 1.16 ± 1.74 BPM $r = 0.9958$
Galli et al. [56]	2017	A zero padded DFT to estimate heart rate and kalman filter for tracking	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	NA	Sub-space decomposition	FFT	Kalman filter	AAE= 1.85 ± 1.00 BPM AEP= $1.45 \% + 0.79 \%$
Islam et al. [57]	2017	PREHEAT: Hybrid time frequency approach for denoising and wavelet-based heart rate estimation	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	BPF (0.4–3.5 Hz)	EEMD + cRLS	Wavelet-Fourier based	Heuristic decision based approaches	AAE= 0.83 ± 0.96 BPM $r = 0.998$

Table 2 (continued)

Author	Year	Proposed algorithm	Signals used	Data set	Pre-processing	Denoising technique	Heart rate estimation technique	Heart rate tracking technique	Results
Yalan et al. [58]	2017	Two stage approach: one for hybrid motion artifacts removal and second random forest base heart rate estimation	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-23 (Training + Testing)	BPF (0.4–5 Hz)	ANC + SSA	Random forest	Random forest-based spectral peak tracking	AAE= 1.23 ± 0.80 r=0.992
Nathan and Jafari [59]	2017	Property of both particle filter and fusion approach utilized for correct heart rate estimation	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-23 (Training + Testing)	BPF (0.5–15 Hz)	STFT	Particle Filter	Fusion approach	AAE= 1.4 ± 1.55 BPM
Chung et al. [60]	2018	FSM based real-time heart rate estimation	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	4th order Butterworth BPF (0.4–4 Hz) downsampled to 25 Hz	Wiener filter	Crest factor from the Periodogram	FSM Triggered by CF	AAE= 0.79 ± 0.6 BPM AEP=0.60% r=0.997
Lee et al. [40]	2018	Gyroscope sensor-based filtering and DDFD + FSM combined approach to estimate heart rate	3-channel PPG, 3-axis gyroscope, 3-axis accelerometer signals	gyro_acc_ppg-24 (Training)	BPF (0.4–4 Hz) down sampled to 25 Hz	Power spectrum filtering technique	DDFD + FSM	FSM Triggered by CF	AAE= 1.92 BPM AEP= 1.76% r=0.9756
Islam et al. [61]	2018	Cascade and parallel combination of adaptive filters was used with convex combination to reduce the effect of motion artifacts	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	BPF (0.4 Hz–3.5 Hz)	Cascade and parallel connection of adaptive filters	Spectral-domain spectral peak estimation	Search range	AAE= 1.12 ± 2.30 r=0.994
Biagetti et al. [62]	2019	KNN classifier for realtime implementation	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	BPF (0.5–4 Hz)	Automatic activity intensity classifier	Signal subspace decomposition	KNN classifier	AAE= 2.17 r=0.99

Table 2 (continued)

Author	Year	Proposed algorithm	Signals used	Data set	Pre-processing	Denoising technique	Heart rate estimation technique	Heart rate tracking technique	Results
Wang et al. [63]	2019	An LMS-Newton-based cascaded automatic noise cancellation. Notch filters to restore the PPG signal	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	4th order Butterworth BPF (0.4 Hz–4.0 Hz)	ANC	HRFT scheme	Heuristic correction scheme	AAE=0.92 $r=0.997$
MOTIN et al. [64]	2019	Weiner filter used to estimate correct heart rate peak	2 Channel z-scored PPG and 3 axis Acceleration Signal	IEEE SPC—23 (Training + Testing)	4th order Butterworth BPF (0.2–5 Hz)	Recursive Wiener filtering	FFT	Consecutive window Threshold Approach	AAE=1.02±0.44 BPM $r=0.997$
Chung et al. [65]	2019	The method proposed to minimize the error in beat per minute by modifying the signal power of the current heart rate window	2 Channel PPG and 3 axis Acceleration Signal	Two different dataset	4th-order Butterworth BPF (0.4–4) Hz	Subtraction method	FFT	FSM crest factor	AAE=1.20 BPM and AEP=1.05%
Roy and Gupta [66]	2020	MODTRAP: Decomposition and neural network-based heart rate estimating approach	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-12 (Training)	LPF (with zero-phase, 10 Hz cut-off, 60-dB stop band attenuation)	EEMD and neural network model	VMD and window algorithm	Heuristic algorithm	AAE=0.57 BPM AEP=0.43% $r=0.999$
Kumar and Bhaskar [67]	2020	Motion artifacts removal was done using cascaded RLS, NLMS and LMS filters using the softmax activation function	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-22 (Training + Testing)	BPF (0.4–3.5 Hz)	RLS, NLMS, LMS	Softmax activation function + FFT	Phase vocoder	AAE=1.86±2.36 $r=0.988$
Kumar and Bhaskar [68]	2020	CASINOR: Combination of adaptive filters using a single noise reference signal	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC—23 (Training + Testing)	BPF (0.4–3.5 Hz)	RLS and NLMS	Sigmoid + FFT	Phase vocoder	AAE=1.20±1.77 $r=0.990$

Table 2 (continued)

Author	Year	Proposed algorithm	Signals used	Data set	Pre-processing	Denoising technique	Heart rate estimation technique	Heart rate tracking technique	Results
Chung et al. [38]	2020	An eight-layer model was proposed to design a heart rate estimation wearable device using PPG technology	2 Channel PPG and 3 axis Acceleration Signal	IEEE SPC-23 (Training + Testing)	Butterworth BPF (0.4–4 Hz) downsampled to 25 Hz	Convolution and maxpooling layer	Concatenate Layer	LSTM layers	AAE=0.76 BPM AEP=0.66%

BPF bandpass filter, *SSA* singular spectrum analysis, *r* Pearson correlation coefficient, *LPF* low pass filter, *VMD* variational mode decomposition, *FFT* fast Fourier transform, *SVD* singular value decomposition, *ALS-SS* asymmetric least squares spectrum subtraction, *FSM* finite state machine

while TROIKA uses a single heart rate trajectory. Using an adaptive noise cancellation (ANC) filter for denoising, this work calculated an AAE of 1.11 ± 2.33 BPM with the same PPG dataset used in [37]. TROIKA's major disadvantage is that it uses SSA for signal decomposition, which partially removes the motion artifacts from raw a PPG signal. An algorithm based on EEMD to minimize the motion artifact from the PPG signal, which occurs due to strong physical exercise, is proposed [44]. In [44], EEMD is used for signal decomposition to remove the motion artifacts from a raw PPG signal partially. After signal denoising, the spectrum subtraction approach is used to find the correlation between motions affected PPG signal and reference noise signal to remove the remaining motion artifact from the spectrum. This approach resulted in better noise performance than TROIKA in terms of AAE and AEP, which are 1.83 ± 1.21 BPM and 1.40%, respectively.

Instead of signal decomposition, an adaptive motion artifact reduction approach to suppress motion was proposed in [45] using an NLMS adaptive filter to reduce motion artifact. A time-varying bandpass filter is used for accurate heart rate estimation. Filter updates its coefficient at constant intervals so that it can efficiently track the frequency component. In this method, AAE and AEP have calculated as 1.71 ± 0.49 BPM and 1.41%, respectively, resulting in a 27% reduction in AAE. This approach works well when the motion artifact is weak. To improve the performance of the adaptive filter to suppress the strong motion artifact from the PPG signal, in [47] singular value decomposition (SVD) stage is introduced before the adaptive filter to decompose the three-axis accelerometer signal having different periodic components. SVD eases the convergence of the adaptive filters. The decomposed output and reference noise signal used by the adaptive filter to suppress the motion artifact from the PPG signal. An AAE and AEP of 1.25 ± 0.6 BPM and 0.99% calculated respectively.

One major issue faced by benchmark techniques like TROIKA and JOSS was the runaway error problem. A hybrid approach that abolishes the dependency of heart rate estimation over the previous window is proposed [33] to overcome the runaway error problem. In this method, a two-channel PPG signal is used to estimate heart rate. EEMD approach is used to obtain a noise-free PPG signal, and the RLS adaptive filter is used to remove motion artifacts and identify the heart rate peak. An AAE of 1.15 ± 2.37 BPM using a single channel and 1.02 ± 1.79 BPM with two channels are reported.

Apart from various advantages, the proposed in [33] does not denoise the signal effectively when the motion artifact frame exists close to the heart rate frame. A method named *precise heart rate tracking* (PREHEAT) is proposed in [57] by introducing a dynamic order correlation-based recursive least-squares (cRLS) adaptive filter to minimize the effect

of motion artifact effectively. After denoising, Wavelets are used in addition to Fourier transform to detect the correct heart rate peak. PREHEAT calculates an improved AAE of 0.83 ± 0.96 BPM. PPG time–frequency features based motion artifact removal approach was proposed in 2016 [51], named *time–frequency spectra of PPG signal* (TifMA) for realtime heart signal analysis. Compared to published work related to motion artifact removal and heart rate detection, TifMA also tests the noise frame usability for heart rate peak detection instead of deleting them. Using frequency modulated and amplitude modulated data from the usable signal, the proposed algorithm accurately estimates heart rate value using a subsequent window approach. The affectivity of TifMA was tested in terms of specificity and selectivity.

Various methods proposed in the literature to denoise a PPG signal are based on signal decomposition or adaptive filtering that failed to provide reliable results in realtime applications. An approach based on cascaded RLS adaptive filter and EEMD is proposed in [35] to overcome the limitations posed by realtime PPG applications. The author computed an AAE of 1.16 ± 2.23 BPM and AEP of 0.93% with the IEEE SPC 12 candidate's dataset.

In *particle filter-based algorithm for heart rate estimation using photoplethysmographic signals* (PARHELIA) [48], a method based on particle filter for heart rate estimation is proposed with tracking multiple candidates. A particle filter can help recover an incorrect track to the correct track. PARHELIA uses the acceleration signals to update the weight of particles in the particle filter to reduce the effects of motion artifacts. Updating weight depends on three steps, namely prediction, weight calculation, and resampling. An AAE of 1.17 BPM was calculated with PARHELIA, which showed an improvement of 8.6% compared to the TROIKA. Another work based on particle filter proposed in [59] the heart peak by focusing on those consistent with time. Instead of three axes reference noise signal, a single reference noise signal was used to reduce computational complexity having the highest peak frequency. Instead of relying on any reference characteristics points for measurement, the proposed filter considers noisy signals as input and modifies the weight of selected particles to analyze heart information. Heart rate was estimated by detecting the highest weight particle assigned to each window. To further refine the heart rate estimation, a fusion method was used, in which an AAE of 1.4 ± 1.55 BPM is calculated [59].

An algorithm named *multiple reference adaptive noise cancellation technique* (MURAD) is proposed in [49] to improve the effectiveness of adaptive filters for accurate heart rate estimation. In this method, the three-axis accelerometer reference noise signal and the difference between two PPG signals are used as the reference noise signal. Instead of using a fixed reference noise signal for each window, the proposed work provides flexibility to select a

realtime reference noise signal for accurate and reliable heart rate estimation. An AAE of 0.97 ± 1.83 BPM and AEP of $0.76 \pm 1.5\%$ were calculated with MURAD algorithm. In [50], a different approach to separate motion artifacts spectrum and PPG spectrum from raw PPG data is proposed. The harmonic sum model retrieves the fundamental frequency component of the reference noise acceleration signal within a short window range to estimate the heart rate spectrum from raw a PPG signal. An AAE of 0.73 ± 0.83 BPM was calculated, which showed improved error performance over methods already reported.

As observed from the literature, the frequency-domain approach, like EMD [33, 44], increases the computational complexity. In [70], a modified EMD approach with variance characterization to identify motion-affect periods in the whole PPG signal from a predefined time window is proposed to overcome the computation complexity issue. An AEP is calculated as 1.03%, which demonstrated the use of a modified EMD approach introduced in wearable devices [70]. A method that uniquely detects heart rate peak frequency under the realtime environment with reduced system complexity is proposed in [53] to reduce computational complexity. A unique property of this work was that it does not rely on heart rate information recorded in the previous window for heart rate detection. To avoid large-amplitude reference noise signal detection in detecting heart rate, a spectral division approach is used to extract the reference accelerometer spectra from the PPG signal. A composition of all frequency components is used to measure the highest peak frequency under the desired range. Finally, a constant value based jump procedure was introduced to track the heart rate in the noisy spectrum.

Wiener filter and phase vocoder based new approach named WFPV is proposed in [53] to overcome the limitations of computational complexity faced by methods based on heuristic rules or thresholds detection for heart rate estimation. A Wiener filter is used to attenuate the effect of strong motion artifacts. A phase vocoder was used, which allows the user to estimate heart rate for a short period. Compared with previously presented methods, WFPV improved AAE to 1.02 BPM and AEP to 0.81%. The Wiener filter used reference noise signals from accelerometers from all three axes to filter motion artifacts. In [64], a modified method to remove motion artifacts by using a three-axis acceleration reference noise signal is proposed.

Some zeros were added at the end of the signal to make heart rate resolution less than 1 BPM to identify heart rate peak frequency. The heart rate is further tracked by comparing the estimated result with a predefined threshold. An AAE of 1.02 ± 0.44 BPM was calculated, which is better than most of the proposed work. Conceptually similar work was also presented in [56] to estimate the correct heart rate peak. A one-variable Kalman filter was employed to refine

the heart rate value. To reduce the effect of the motion noise SVD technique filters out a subset matrix of noise-free PPG signal. To assess the present work compared with [53], the authors calculated two more parameters for maximum absolute deviation and standard deviation. Maximum absolute deviation provides the capability to assess the algorithm's accuracy at each point in a window and a standard deviation computed over the whole window.

Considering the advantages of time–frequency approaches simultaneously, a time–frequency based short-time Fourier spectral tracking (SFST) approach was proposed to estimate heart rate in a short period. As FFT provides limited resolution to study heart rate, [54] replaced FFT with STFT for realtime heart rate estimation. After the preprocessing step, the signal is divided in a short time window using the STFT approach to reduce motion artifacts. A cyclic moving average filter is used to filter out unexpected variance values in heart rate due to complex motion artifacts. Using IEEE SPC 12 candidates' dataset, calculated results showed improved AAE results of 1.06 ± 0.69 BPM and AEP of $0.94\% \pm 0.53\%$. In [55], a new method to utilize the potential of a time–frequency based approach for heart rate estimation is proposed. A combination of RLS adaptive filter (a time-domain approach) output and SSA (frequency domain approach) output was used to minimize the motion artifacts in [58]. By considering the previous heart rate time window, a conditional sum approach was used to avoid false estimation of heart rate. For reliable heart rate peak detection, tracking of heart rate within a search range is implemented as a post-processing step, which resulted in an AAE of 1.16 ± 1.74 BPM.

Researchers have devoted many efforts to provide low computational complexity approaches to estimate heart rate for wearable devices accurately in recent years. In [58], an approach based on the random forest binary decision algorithm for accurate heart rate estimation is proposed. A binary decision algorithm helps in deciding between two algorithms used for motion artifacts removal. For feature extraction, wavelet-based techniques were used. Compared with the result of a similar approach, this work calculated an AAE of 1.23 BPM with low computational complexity.

Another concern in developing wearable devices is the accurate estimation of heart rate during intensive physical activity. In [60], an algorithm to identify heart rate in a real-time environment is proposed. The main objective of this work is to remove the motion artifacts spectrum that occurs due to physical movement across the sensor. For denoising, the Wiener filtering approach was used. To solve the difficulties faced in heart rate estimation during intensive exercise, the finite state machine (FSM) based algorithm was used under the post-processing step, ignoring inaccurate estimations. Compared with the previously reported method, an improved result in terms of AAE 0.79 ± 0.6

BPM was calculated with IEEE SPC 23 candidate dataset. Even though the accelerometer signals cancel out the motion artifact, they introduce gravitational acceleration error. To solve the problem of gravitational acceleration, a gyroscope is used to record the reference noise signal [40].

For heart rate estimation using wearable devices, properties like tracking ability, robustness, and computational cost are considered important design parameters and can be realized by a combination of adaptive filters [71]. By assigning different weights to the combined layers of an adaptive filter, the adaptive filter's denoising performance can be improved [61]. The output of two parallel cascaded networks was combined using a convex combination to improve the output efficiency, which depends on the choice of filters and adaptive filter parameters. A three-stage cascaded network model was proposed to filter out motion artifacts in three directions. The output from the cascaded RLS and cascaded LMS stage were combined using the convex combination. An AAE of 1.12 BPM was calculated on the same dataset used in [36]. Using the LMS filter properties, a method to minimize motion artifacts was also introduced to estimate heart rate accurately. A notch filter was used to reproduce the PPG signal from the detected heart rate peak [63]. An AAE of 0.92 BPM was calculated, which showed an improved result compared to the state-of-the-art techniques.

Despite this improvement in error performance, the performance of the LMS filter depends upon an adjustment of tap weight, which is directly related to the input vector. If the input vector is not bounded, then the LMS filter may face gradient noise amplification due to the incorrect selection of step size. To avoid the gradient noise amplification and step size issues, a three-stage cascaded adaptive filter RLS, NLMS, LMS based approach is proposed in [72]. In [72], two different pairs of adaptive filters are combined using a convex combination to effectively denoise the PPG signal. Sigmoid function based parameters are assigned to each pair of adaptive filters were updated at each iteration to improve the filtering performance. The FFT-based approach is used to estimate the heart rate. Convex combination assigns constant value at each combinational layer consists of different output combinations of the adaptive filter. It provides maximum value to those layers that perform well in that iteration. Using IEEE SPC 12 candidates' data set, an AAE of 0.92 BPM is calculated. For reliable denoising and heart rate, in [73], three stages of cascaded adaptive filters output are combined using the softmax normalized function. The FFT approach estimates the heart rate value by using a phase vocoder. An AAE of 1.86 BPM was calculated on large datasets, which showed less error than other techniques that used the same data set to test the algorithms. By combining the output of adaptive filters, estimation of heart rate becomes more accurate, but computational time increases. In [68], a new denoising algorithm named combination of adaptive

filters using single noise reference signal (CASINOR) is proposed to reduce computational time and error values. Only RLS and NLMS adaptive filters are used to denoise the signal. A sigmoid function was also used to combine the output of both filters. The main feature of CASINOR was that it requires only a single reference acceleration noise signal instead of a three-direction reference noise signal. The accelerometer signal with maximum power is chosen as a reference noise signal. After spectral estimation, a phase vocoder is used to refine the heart rate peak values. Using CASINOR, an AAE of 1.92 BPM is calculated with IEEE SPC 23 candidates' dataset.

Following the decomposition approach for denoising in [73], a method based on VMD is introduced to study the PPG signal in small data length to improve heart rate estimation accuracy. Further to the identified heart rate peak, the PCA approach was used to select the more heart rate relevant mode. With shorter length data, the proposed [73] decompose method identified heart rate peak with less error. Further, to identified accurate heart rate spectrum peak during physical exercise, in [74], a personalized deep learning approach was introduced. For accurate estimation, the algorithm was trained according to the realtime situation. An AAE of 1.47 ± 3.37 BPM was calculated with IEEE SPC 23 candidate's dataset. In realtime, the nature of noise cannot be predicted. A fixed reference noise model may not work effectively to analyze the signal in a realtime environment. In [75], a neural network-based classification approach to separate clean segments without reference noise acceleration signals is proposed for realtime applications. The main feature of this work was that instead of assessing the complete PPG frame, it access individual pulse behavior. The efficacy of the work depends upon the accuracy of the reference template. In [66], a hybrid approach comprised of VMD and neural network classification to estimate heart rate in a realtime environment is proposed. This work identifies the beat morphological structure of beat besides heart rate estimation using a neural network model-based template matching feature. An AAE of 0.53 BPM was calculated on IEEE SPC 23 candidate's dataset, which showed improved performance over the state-of-the-art techniques. In [76], a hybrid approach to jointly estimate heart and respiratory information from the IMF spectrum is proposed. In this method, the EEMD approach is used to generate the desired frequency window's IMF function. PCA technique was used to extract the most relevant feature for heart rate estimation. The method showed similar results on IEEE SPC 23 candidate's dataset obtained, but the accuracy and reliability of this work are far greater than the EEMD approach. Effectiveness of work is calculated in terms of mean and variance with a value of 99.95% and 0.0010% respectively.

Most of the techniques presented were tested with the common dataset IEEE SPC 2015. However, this dataset was

recorded with little physical exercise, and each dataset has a duration of less than one hour. In [75], to design a more robust system, a new dataset PPG DaLiA is introduced, which contains recording with some real-life daily activities with a duration of more than 36 h. Two-channel PPG signal and three-axis accelerometer signal are firstly separated in a short window duration of eight seconds. Then Fast Fourier approach was implemented on each window for heart rate estimation. The tracking step is introduced in the CNN layer to improve accuracy and reliability, which relies heavily on the correlation property of the subsequent window of the heart cycle.

In [62], to reduce the computation complexity problem faced by benchmark techniques [44], an SVD based algorithm to estimate heart rate from motion corrupted raw PPG signal is introduced. A genetic algorithm was used to optimize the value of parameters used under the heart rate tracking step to deal with the different motion artifacts cases. From the acceleration signal, the KNN classifier is used to detect the intensity of physical activities. The proposed [62] approach produced comparable results but required less complex processing stages. An AAE of 2.17 BPM was calculated on the same dataset [42]. One more technique based on neural networks for heart rate estimation is introduced in [38], which uses an eight-layer filter model to track the heart rate. The Gaussian distribution function is used to improve the accuracy of the estimation signal. Complex mathematical calculations limit the application of the eight filter model to use in the realtime analysis of heart rate.

To improve mean absolute error performance, A method based on the power spectrum of the desired signal to improve mean absolute error performance is proposed in [65]. This approach deal with the signal's power for measuring accurate heart rate peak during body movement. Estimating the true heart rate of the present window depends on the accuracy of the previous window; hence the crest factor property of FSM is used to check the response of heart rate in the

subsequent window. The mean value of the previous heart rate window in terms of the Gaussian kernel function is multiplied by the current time window to improve the SNR value. Improved results in terms of AAE of 1.20 BPM and AEP of 1.05% were calculated with IEEE SPC 23 candidates' dataset. After considering problems faced in time and frequency-based approaches, in [67], a modified approach simultaneously uses both the PPG modes to reduce the effect of noise. The effect of noise imposes on the PPG signal depends on the penetration depth of light used to capture the signal. A total of six sensors of different wavelengths were used to illuminate the skin. Out of six sensors, four sensors were used for reflection mode and two for transmissive mode. Separate LEDs were used because the transmissive mode needs a light source that penetrates deeper into the skin. Blue, green, and infrared light

show superior results compared to other light sources for estimating heart signals.

6 Challenges and Discussion

In the last decade, monitoring cardiovascular health has become an essential feature for the early diagnosis and prevention of cardiovascular diseases. Due to a lack of efficient monitoring tools, the mortality rate due to cardiovascular diseases increases year by year. To prevent any accidents related to cardiovascular disease, personal health monitoring devices are gaining importance. Therefore the demand for battery-operated wearable sensing devices is ever increasing. Wearable devices with PPG sensor technology will give people the flexibility to measure their health status at any time and any place.

Based on the literature review, PPG technology can monitor heart rate in wearable devices like bands and watches. The accuracy of wearable PPG-based monitoring tools suffers from effects related to motion artifacts. Researchers devoted a lot of effort to design an accurate and reliable monitoring tool in the healthcare system to tackle motion artifacts. We have also highlighted the algorithm proposed to reduce the effect of motion artifacts from the PPG signal. In the literature, time-domain approaches like adaptive filtering and frequency domain approach like signal decomposition are used to denoise. Later on, some of the methods combined the positive feature of both techniques to provide accurate results. Signal-based techniques can give noise-free signals, but they faced computational complexity problems.

On the other hand, adaptive noise cancellation showed reliable results only when reference noise signals correlate highly with the motion spectrum, which is not possible in realtime. In addition to this work, proposed related to heart rate estimation using PPG provide inaccurate results if the noise spectrum lies close to the heart rate peak. Moreover, due to the non-stationary nature of the biological signal, Fourier-based heart rate estimation also not provides reliable results.

Despite the outstanding progress in the past few years related to motion artifact removal from PPG signal discussed in section (III), an effective and computational efficient motion artifact removal algorithm is still in great demand. Therefore there are still many issues to be resolved to implement a realtime continuous method using PPG to monitor cardiovascular behavior during physical activities.

7 Conclusion

This paper presents a review of the potential of Photoplethysmography technology in the field of biomedical signal processing. This paper presented a comprehensive review

of state-of-the-art research on suppressing motion artifacts and heart rate estimation using a PPG-enabled wearable device. In the last decade, the ratio of death worldwide due to cardiovascular diseases increases day by day. This hike is due to faster changing lifestyle, stress level, and people's food habits across the world. To reduce the risk of cardiovascular diseases, a frequent medical checkup is needed for continuous assessment. So regular monitoring of cardiovascular health status is important for early diagnosis and timely treatment of cardiovascular disease. Therefore the need for a portable and wearable device for early diagnosis is growing day by day. Due to their small size and low cost, PPG sensor-based wearable devices showed their potential to use as a health monitoring device in the future. This review paper summarized different techniques proposed in the last ten years for noise suppression and heart rate estimation with PPG technology. Some of the methods were computationally inefficient, and others were inefficient under realtime monitoring. Despite many advantages of the Photoplethysmography sensor, it can produce erroneous data in certain circumstances. One of the main reasons for error is the occurrence of motion artifacts. Therefore the role of the PPG sensor for extracting vital information is limited due to motion artifact. A reliable health monitoring device in a realtime environment requires signal processing algorithms that effectively remove motion artifacts and are computationally efficient.

Declarations

Conflict of interest All Authors of this work declare no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Kamal AAR, Harness JB, Irving G, Mearns AJ (1989) Skin photoplethysmography—a review. *Comput Methods Programs Biomed* 28(4):257–269
2. Vashist SK, Schneider EM, Luong JH (2014) Commercial smartphone-based devices and smart applications for personalized healthcare monitoring and management. *Diagnostics* 4(3):104–128
3. Tamura T (2019) Current progress of photoplethysmography and SPO 2 for health monitoring. *Biomed Eng Lett* 9(1):21–36
4. Warren KM, Harvey JR, Chon KH, Mendelson Y (2016) Improving pulse rate measurements during random motion using a wearable multichannel reflectance photoplethysmograph. *Sensors* 16(3):342
5. Allen J (2007) Photoplethysmography and its application in clinical physiological measurement. *Physiol Meas* 28(3):R1
6. Moraes JL, Rocha MX, Vasconcelos GG, VasconcelosFilho JE, De Albuquerque VHC, Alexandria AR (2018) Advances in

- photoplethysmography signal analysis for biomedical applications. *Sensors* 18(6):1894
7. Moço AV, Stuijk S, de Haan G (2018) New insights into the origin of remote PPG signals in visible light and infrared. *Sci Rep* 8(1):1–15
8. Hartmann V, Liu H, Chen F, Qiu Q, Hughes S, Zheng D (2019) Quantitative comparison of photoplethysmographic waveform characteristics: effect of measurement site. *Front Physiol* 10:198
9. Chan ED, Chan MM, Chan MM (2013) Pulse oximetry: understanding its basic principles facilitates appreciation of its limitations. *Respir Med* 107(6):789–799
10. Joyner MJ, Casey DP (2015) Regulation of increased blood flow (hyperemia) to muscles during exercise: a hierarchy of competing physiological needs. *Physiol Rev* 95:549–601
11. Tamura T, Maeda Y, Sekine M, Yoshida M (2014) Wearable photoplethysmographic sensors—past and present. *Electronics* 3(2):282–302
12. Sun Y, Thakor N (2015) Photoplethysmography revisited: from contact to noncontact, from point to imaging. *IEEE Trans Biomed Eng* 63(3):463–477
13. Liu J, Yan BPY, Dai WX, Ding XR, Zhang YT, Zhao N (2016) Multi-wavelength photoplethysmography method for skin arterial pulse extraction. *Biomed Opt Express* 7(10):4313–4326
14. Spigulis J, Gailite L, Lihachev A, Ertz R (2007) Simultaneous recording of skin blood pulsations at different vascular depths by multiwavelength photoplethysmography. *Appl Opt* 46(10):1754–1759
15. Hertzman AB (1938) The blood supply of various skin areas as estimated by the photoelectric plethysmograph. *Am J Physiol Leg Content* 124(2):328–340
16. Elgendi M (2012) On the analysis of fingertip photoplethysmogram signals. *Curr Cardiol Rev* 8(1):14–25
17. Singh N, Moneghetti KJ, Christle JW, Hadley D, Froelicher V, Plews D (2018) Heart rate variability: an old metric with new meaning in the era of using mhealth technologies for health and exercise training guidance. Part two: prognosis and training. *Arrhythm Electrophysiol Rev* 7(4):247
18. Hernando A, Peláez-Coca MD, Lozano MT, Aiger M, Izquierdo D, Sánchez A et al (2018) Autonomic nervous system measurement in hyperbaric environments using ECG and PPG signals. *IEEE J Biomed Health Inform* 23(1):132–142
19. Castaneda D, Esparza A, Ghamari M, Soltanpur C, Nazeran H (2018) A review on wearable photoplethysmography sensors and their potential future applications in health care. *Int J Biosens Bioelectron* 4(4):195
20. Elgendi M, Fletcher R, Liang Y, Howard N, Lovell NH, Abbott D et al (2019) The use of photoplethysmography for assessing hypertension. *NPJ Digit Med* 2(1):1–11
21. Chakraborty A, Sadhukhan D, Mitra M (2019) An automated algorithm to extract time plane features from the PPG signal and its derivatives for personal health monitoring application. *IETE J Res* 1–13. <https://doi.org/10.1080/03772063.2019.1604178>
22. Elgendi M, Liang Y, Ward R (2018) Toward generating more diagnostic features from photoplethysmogram waveforms. *Diseases* 6(1):20
23. Pilt K, Ferenets R, Meigas K, Lindberg LG, Temitski K, Viigimaa M (2013) New photoplethysmographic signal analysis algorithm for arterial stiffness estimation. *Sci World J* 2013:1–9
24. Chakraborty A, Sadhukhan D, Pal S, Mitra M (2020) Automated myocardial infarction identification based on interbeat variability analysis of the photoplethysmographic data. *Biomed Signal Process Control* 57:101747
25. Ram MR, Madhav KV, Krishna EH, Komalla NR, Reddy KA (2011) A novel approach for motion artifact reduction in PPG signals based on AS-LMS adaptive filter. *IEEE Trans Instrum Meas* 61(5):1445–1457
26. Kim BS, Yoo SK (2006) Motion artifact reduction in photoplethysmography using independent component analysis. *IEEE Trans Biomed Eng* 53(3):566–568
27. Ram MR, Madhav KV, Krishna EH, Komalla NR, Sivani K, Reddy KA (2013) ICA-based improved DTCWT technique for MA reduction in PPG signals with restored respiratory information. *IEEE Trans Instrum Meas* 62(10):2639–2651
28. Goh CH, Tan LK, Lovell NH, Ng SC, Tan MP, Lim E (2020) Robust PPG motion artifact detection using a 1-D convolution neural network. *Comput Methods Programs Biomed* 196:105596
29. Tarvirdizadeh B, Golgouneh A, Tajdari F, Khodabakhshi E (2020) A novel online method for identifying motion artifact and photoplethysmography signal reconstruction using artificial neural networks and adaptive neuro-fuzzy inference system. *Neural Comput Appl* 32(8):3549–3566
30. Zhang Y, Song S, Vullings R, Biswas D, Simões-Capela N, Van Helleputte N et al (2019) Motion artifact reduction for wrist-worn photoplethysmograph sensors based on different wavelengths. *Sensors* 19(3):673
31. Lee J, Kim M, Park HK, Kim IY (2020) Motion artifact reduction in wearable photoplethysmography based on multi-channel sensors with multiple wavelengths. *Sensors* 20(5):1493
32. Motin MA, Karmakar CK, Palaniswami M (2017) Ensemble empirical mode decomposition with principal component analysis: a novel approach for extracting respiratory rate and heart rate from photoplethysmographic signal. *IEEE J Biomed Health Inform* 22(3):766–774
33. Khan E, Al Hossain F, Uddin SZ, Alam SK, Hasan MK (2015) A robust heart rate monitoring scheme using photoplethysmographic signals corrupted by intense motion artifacts. *IEEE Trans Biomed Eng* 63(3):550–562
34. Biswas A, Roy MS, Gupta R (2019) Motion artifact reduction from finger photoplethysmogram using discrete wavelet transform. In: Bhattacharyya S, Mukherjee A, Bhaumik H, Das S, Yoshida K (eds) Recent trends in signal and image processing. Springer, Singapore, pp 89–98
35. Ye Y, Cheng Y, He W, Hou M, Zhang Z (2016) Combining nonlinear adaptive filtering and signal decomposition for motion artifact removal in wearable photoplethysmography. *IEEE Sens J* 16(19):7133–7141
36. Zhang Z, Pi Z, Liu B (2014) TROIKA: a general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. *IEEE Trans Biomed Eng* 62(2):522–531
37. Chung H, Ko H, Lee H, Lee J (2020) Deep learning for heart rate estimation from reflectance photoplethysmography with acceleration power spectrum and acceleration intensity. *IEEE Access* 8:63390–63402
38. Jarchi D, Casson AJ (2017) Description of a database containing wrist PPG signals recorded during physical exercise with both accelerometer and gyroscope measures of motion. *Data* 2(1):1
39. Lee H, Chung H, Lee J (2018) Motion artifact cancellation in wearable photoplethysmography using gyroscope. *IEEE Sens J* 19(3):1166–1175
40. Reiss A, Indlekofer I, Schmidt P, Van Laerhoven K (2019) Deep PPG: large-scale heart rate estimation with convolutional neural networks. *Sensors* 19(14):3079
41. Zhang Z (2015) Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction. *IEEE Trans Biomed Eng* 62(8):1902–1910
42. Murthy NKL, Madhusudana PC, Suresha P, Periyasamy V, Ghosh PK (2015) Multiple spectral peak tracking for heart rate monitoring from photoplethysmography signal during intensive physical exercise. *IEEE Signal Process Lett* 22(12):2391–2395

43. Sun B, Zhang Z (2015) Photoplethysmography-based heart rate monitoring using asymmetric least squares spectrum subtraction and bayesian decision theory. *IEEE Sens J* 15(12):7161–7168
44. Zhang Y, Liu B, Zhang Z (2015) Combining ensemble empirical mode decomposition with spectrum subtraction technique for heart rate monitoring using wrist-type photoplethysmography. *Biomed Signal Process Control* 21:119–125
45. Fallet S, Vesin JM (2015) Adaptive frequency tracking for robust heart rate estimation using wrist-type photoplethysmographic signals during physical exercise. In: 2015 computing in cardiology conference (CinC). IEEE, pp 925–928
46. Salehizadeh S, Dao D, Bolkhovskiy J, Cho C, Mendelson Y, Chon KH (2016) A novel time-varying spectral filtering algorithm for reconstruction of motion artifact corrupted heart rate signals during intense physical activities using a wearable photoplethysmogram sensor. *Sensors* 16(1):10
47. Mashhadi MB, Asadi E, Eskandari M, Kiani S, Marvasti F (2015) Heart rate tracking using wrist-type photoplethysmographic (PPG) signals during physical exercise with simultaneous accelerometry. *IEEE Signal Process Lett* 23(2):227–231
48. Fujita Y, Hiromoto M, Sato T (2017) PARHELIA: particle filter-based heart rate estimation from photoplethysmographic signals during physical exercise. *IEEE Trans Biomed Eng* 65(1):189–198
49. Chowdhury SS, Hyder R, Hafiz MSB, Haque MA (2016) Realtime robust heart rate estimation from wrist-type PPG signals using multiple reference adaptive noise cancellation. *IEEE J Biomed Health Inform* 22(2):450–459
50. Dubey H, Kumaresan R, Mankodiya K (2018) Harmonic sum-based method for heart rate estimation using PPG signals affected with motion artifacts. *J Ambient Intell Humaniz Comput* 9(1):137–150
51. Dao D, Salehizadeh SM, Noh Y, Chong JW, Cho CH, McManus D et al (2016) A robust motion artifact detection algorithm for accurate detection of heart rates from photoplethysmographic signals using time–frequency spectral features. *IEEE J Biomed Health Inform* 21(5):1242–1253
52. Farhadi M, Mashhadi MB, Essalat M, Marvasti F (2016) Real-Time Heart Rate Monitoring Using photoplethysmographic (PPG) signals during intensive physical exercises. *bioRxiv*, 092627
53. Temko A (2017) Accurate heart rate monitoring during physical exercises using PPG. *IEEE Trans Biomed Eng* 64(9):2016–2024
54. Zhao D, Sun Y, Wan S, Wang F (2017) SFST: a robust framework for heart rate monitoring from photoplethysmography signals during physical activities. *Biomed Signal Process Control* 33:316–324
55. Islam MT, Zabir I, Ahmed ST, Yasar MT, Shahnaz C, Fattah SA (2017) A time-frequency domain approach of heart rate estimation from photoplethysmographic (PPG) signal. *Biomed Signal Process Control* 36:146–154
56. Galli A, Narduzzi C, Giorgi G (2017) Measuring heart rate during physical exercise by subspace decomposition and Kalman smoothing. *IEEE Trans Instrum Meas* 67(5):1102–1110
57. Islam MS, Shifat-E-Rabbi M, Dobaie AMA, Hasan MK (2017) PREHEAT: Precision heart rate monitoring from intense motion artifact corrupted PPG signals using constrained RLS and wavelets. *Biomed Signal Process Control* 38:212–223
58. Ye Y, He W, Cheng Y, Huang W, Zhang Z (2017) A robust random forest-based approach for heart rate monitoring using photoplethysmography signal contaminated by intense motion artifacts. *Sensors* 17(2):385
59. Nathan V, Jafari R (2017) Particle filtering and sensor fusion for robust heart rate monitoring using wearable sensors. *IEEE J Biomed Health Inform* 22(6):1834–1846
60. Chung H, Lee H, Lee J (2018) Finite state machine framework for instantaneous heart rate validation using wearable photoplethysmography during intensive exercise. *IEEE J Biomed Health Inform* 23(4):1595–1606
61. Islam MT, Ahmed ST, Zabir I, Shahnaz C, Fattah SA (2018) Cascade and parallel combination (CPC) of adaptive filters for estimating heart rate during intensive physical exercise from photoplethysmographic signal. *Healthc Technol Lett* 5(1):18–24
62. Biagetti G, Crippa P, Falaschetti L, Orcioni S, Turchetti C (2019) Reduced complexity algorithm for heart rate monitoring from PPG signals using automatic activity intensity classifier. *Biomed Signal Process Control* 52:293–301
63. Wang M, Li Z, Zhang Q, Wang G (2019) Removal of motion artifacts in photoplethysmograph sensors during intensive exercise for accurate heart rate calculation based on frequency estimation and notch filtering. *Sensors* 19(15):3312
64. Motin MA, Karmakar CK, Palaniswami M (2019) PPG derived heart rate estimation during intensive physical exercise. *IEEE Access* 7:56062–56069
65. Chung H, Lee H, Lee J (2019) State-dependent Gaussian kernel-based power spectrum modification for accurate instantaneous heart rate estimation. *PLoS ONE* 14(4):e0215014
66. Roy B, Gupta R (2020) MoDTRAP: improved heart rate tracking and preprocessing of motion-corrupted photoplethysmographic data for personalized healthcare. *Biomed Signal Process Control* 56:101676
67. Arunkumar KR, Bhaskar M (2020) Heart rate estimation from wrist-type photoplethysmography signals during physical exercise. *Biomed Signal Process Control* 57:101790
68. Arunkumar KR, Bhaskar M (2020) CASINOR: combination of adaptive filters using single noise reference signal for heart rate estimation from PPG signals. *Signal Image Video Process* 14:1507–1515
69. Yousefi R, Nourani M, Ostadabbas S, Panahi I (2013) A motion-tolerant adaptive algorithm for wearable photoplethysmographic biosensors. *IEEE J Biomed Health Inform* 18(2):670–681
70. Pang B, Liu M, Zhang X, Li P, Yao Z, Hu X et al (2016) Advanced EMD method using variance characterization for PPG with motion artifact. In: 2016 IEEE biomedical circuits and systems conference (BioCAS). IEEE, pp 196–199
71. Arenas-Garcia J, Azpicueta-Ruiz LA, Silva MT, Nascimento VH, Sayed AH (2015) Combinations of adaptive filters: performance and convergence properties. *IEEE Signal Process Mag* 33(1):120–140
72. Arunkumar KR, Bhaskar M (2019) Heart rate estimation from photoplethysmography signal for wearable health monitoring devices. *Biomed Signal Process Control* 50:1–9
73. Sharma H (2019) Heart rate extraction from PPG signals using variational mode decomposition. *Biocybern Biomed Eng* 39(1):75–86
74. Biswas D, Everson L, Liu M, Panwar M, Verhoef BE, Patki S et al (2019) CorNET: deep learning framework for PPG-based heart rate estimation and biometric identification in ambulant environment. *IEEE Trans Biomed Circuits Syst* 13(2):282–291
75. Roy MS, Gupta R, Chandra JK, Sharma KD, Talukdar A (2018) Improving photoplethysmographic measurements under motion artifacts using artificial neural network for personal healthcare. *IEEE Trans Instrum Meas* 67(12):2820–2829
76. Lei R, Ling BWK, Feng P, Chen J (2020) Estimation of heart rate and respiratory rate from PPG signal using complementary ensemble empirical mode decomposition with both independent component analysis and non-negative matrix factorization. *Sensors* 20(11):3238

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

Actionable strategy framework for digital transformation in AECO industry

Digital transformation in AECO industry

Sanjay Bhattacharya

*University School of Architecture and Planning,
Guru Gobind Singh Indraprastha University, New Delhi, India, and*

K.S. Momaya

*Shailesh J. Mehta School of Management, Indian Institute of Technology Bombay,
Mumbai, India*

Received 31 July 2020
Revised 13 December 2020
14 March 2021
2 April 2021
Accepted 3 April 2021

Abstract

Purpose – The Fourth Industrial Revolution (4IR) holds the potential to improve capabilities- and technology-based innovation, which will enable breakout for architectural, engineering, construction and operation and maintenance (AECO) companies, for international competitiveness. Though the top management of such companies is convinced on the utility of the applications, they are unsure on the strategy of implementing the same. The objective of this research is to suggest a strategy framework for digital transformation of the AECO value chain.

Design/methodology/approach – The nascent level of research on 4IR in construction necessitated the adoption of the integrative review methodology for the study. Extensive literature review of research on strategy and 4IR has been utilized to establish the validity of the first two pillars, namely “a strategy of simple rules in a complex environment; and deployment of dynamic capabilities.” The validation of a construct for the third pillar of “confluence of change and continuity forces” has been achieved via hypothesis testing of data obtained through a questionnaire survey.

Findings – The present study has integrated three diverse ideas of strategy, named as the pillars, to facilitate sustainable digital transformation. Within the third pillar, top three continuity forces which offer resistance to change are organization culture, existing delivery processes and networks, and existing standard operating procedures. On the other hand, the leading drivers of change are needs of competitiveness; global industry trends and the advent of new technologies/innovations.

Research limitations/implications – This provides a practical approach to operationalize digital transformation of the AECO at an organization level. The validation relied on opinion and perspectives of a sample frame in the Indian context, which was its limitation.

Originality/value – This paper suggests a strategy framework of three pillars to help address specific strategy dilemmas during implementation of digital transformation of particular organizations in AECO. The study contributes to both theory and practice by helping leaders of AECO companies, associations, policymakers and the academia to strategize transformations successfully.

Keywords 4IR, Strategy as process and practice (SAPP), Digital transformation strategy, AECO competitiveness, Comprehensive/integrating framework, Continuity and change

Paper type Conceptual paper

Introduction to 4IR context

The Fourth Industrial Revolution (4IR) has the promise and potential to improve capabilities- and technology-based innovation, which will enable breakout for architectural, engineering, construction and operation (AECO) companies, for international competitiveness (Momaya, 2014). While 4IR is the buzzword these days (Deloitte, 2019), its potential to contribute to competitiveness should be assessed critically. The active use of digitalization, automation and the widening use of information and communications technology (ICT) across industries, via the use of technologies of cyber-physical systems, Internet of Things (IoT), cloud computing and cognitive computing, is described as 4IR. The term Industry 4.0 was first coined by the German association “Industrie 4.0” in 2011. The association, consisting of



executives, scholars and policymakers, suggested a future based on the digitalization of firm processes (Kagermann *et al.*, 2013). The core idea 4IR revolves around running businesses by adopting digital technologies that can help companies create connections between their machinery, supply systems, production facilities, final products and customers to gather and share information or data on a real-time basis. The revolution opens possibilities for modern techniques to support many components within the industry and with limitless potential. The 4IR aims for viable and sustainable production systems, involving a higher level of complexity by integrating the production, product and service processes (Lasi *et al.*, 2014; Lee and Lee 2015; Bahrin *et al.*, 2016).

McKinsey estimates that switching to automated 4IR can boost productivity in technical professions by 45–55% (Caylar, 2016). IoT-assisted production has already been deployed by companies like Airbus, Cisco, Siemens and several other leaders in the 4IR space. The quantum of savings has, however, still not been captured officially. The changes that are powered by these emerging technologies are expected to offer a better way to organize and manage all standard processes (prototyping, development, production, logistics, supply, etc.) across industries. The technology initiatives are often referred to as exponential technologies because their deployment in each period has the potential to double its productivity performance. In other words, it looks to progressively halve the cost in each period elapsed. It enables a price performance that makes it possible to solve contemporary business problems in ways hitherto unknown or not possible previously.

Likewise the manufacturing industry, AECO performance can also be enhanced through 4IR. The implementation of 4IR in AECO can give rise to the scenario where every mechanized automation would be interconnected through technologies to operate and share information and eliminate human intervention to increase efficiencies (Axelsson *et al.*, 2019). It is appreciated that there are several complexities within AECO, which hinder the easy adoption and compatibility of the technologies. Oesterreich and Teuteberg (2016) have highlighted barriers including complexity, uncertainty, fragmented supply chain, short-term thinking and organization culture. The AECO projects are complex in nature due to the involvement of several stakeholders in a project. Each project itself is unique, and the level of risks and uncertainty in a project adds to the complications. Adding to this, the temporary and short-term nature of projects is a major hurdle to progress and innovative processes. The companies have to constantly deal with troubles recruiting a talented workforce, with the right talent, networking with contractors and suppliers and inadequate transfer of knowledge across projects or even within the industry.

The culture within the industry is known for its reluctant and suspecting nature. While other industries have adopted product and process innovations into the core of their operations, the engineering and construction sector has not kept its pace to adopt technological opportunities (Chan and Ejohwomu, 2018; Hasan *et al.*, 2018). As a result, there has been a predominant stagnation of productivity and efficiencies. Adopting 4IR can be a rare opportunity to achieve inflection in the productivity curves within the industry. Conversely, a reluctance in implementing 4IR technologies can prove to be the nemesis for laggards. It is the companies that compete within an industry, and unless they take proactive measures to adopt the impending changes, they will be rendered uncompetitive, and new players will replace them. The business world is full of examples like the Xerox/Canon or Caterpillar/ Komatsu stories.

Potential uses of 4IR in AECO

Tools such as three-dimensional (3D) scanning, building information modeling (BIM), drones and augmented reality have a potential for extensive use in AECO. By incorporating these innovations, companies can increase productivity level, safety and quality of projects.

BIM has become the single largest central integrating tool at all stages in the project value chain (Mzyece *et al.*, 2019; Ji *et al.*, 2020; Gerrish *et al.*, 2017; Bazjanac, 2006). It creates a possibility to interact and collaborate on a real-time basis throughout the project life cycle. It helps all stakeholders identify potential lacunae in design, construction or operational issues (Ejohwomu *et al.*, 2017; Azhar, 2011). Support from augmented reality, virtual reality or mixed reality can increase customers' understanding of the final product early in the design phase, to avoid changes during project execution (Juan *et al.*, 2017). Proactive use of BIM can improve building quality by the timely discovery of problems. It also enables the concept of integrated project delivery into a collaborative process of consultants and other stakeholders, to reduce waste and optimize efficiency through all phases (Okedara *et al.*, 2020; Glick and Guggemos, 2009).

The IoT is a network of Internet-connected objects that can collect and exchange data on a real-time basis. It comes functional with cyber-physical systems, which allow humans to monitor the processes in real time without physical presence. It can find applications in increasing productivity and monitoring, maintenance, safety and security including wearables, unmanned aerial vehicles like drones, quality control, optimization and creating digital twins, among many others. Besides, IoT devices and sensors can collect job site data in a more affordable, efficient and effective way than previously imaginable (Rane and Narvel, 2019). Improved work safety on-site can be achieved through IoT devices, given the industry's high risk of workplace injuries and accidents.

Construction is currently known to be predominantly driven by manual labor. The productivity and the quality of work produced vary hugely even within the same context (Ellis, 2019). Potentially, robots are capable of working longer, faster and harder. Hence, construction labor is a prime candidate for automation. Reduced labor costs can also be affected through the use of robotics and automatic workflows, be it brickwork or plastering. Automatic tracking of equipment and materials (through the use of embedded sensors like radio-frequency identification [RFID]) can reduce inventory handling costs (Dallasega *et al.*, 2018; Ejohwomu and Hughes, 2019). Project time can be saved by using concepts like prefabrication, 3D printing and additive manufacturing (in an offsite mode), rather than the conventional brick and mortar construction (Moon *et al.*, 2020; Liu and Xu, 2017; Tao *et al.*, 2019).

Cloud computing in combination with BIM-based platforms or social media applications can efficiently improve collaboration among companies and help in streamlining the supply chain management. Big data analytics can support project managers in enhanced decision-making through increased access to accurate and real-time information (Qian and Papadonikolaki, 2020). Predictive simulation can offer insight into modeling factors such as resource utilization, queuing length, sensitivity analysis and what-if scenarios in construction processes. Simulation models can represent a complete portrayal of any system. It may also allow various attributes within a model to be investigated during the experimentation stage of a project, e.g. risk analysis and assessment (Rane *et al.*, 2019). Sensitivity analyses may include activity durations, machine breakdowns, material quantity variations, weather and resource configurations, just to name a few. The benefits of adopting the above technologies have been elicited in detail by Oesterreich and Teuteberg (2016) in their research.

The emerging trends in the global construction industry also reveal affinity toward 4IR technologies as summarized in Table 1.

The statements indicate the rapid adoption of 4IR in construction also. Mahidhar and Davenport (2018) argue that no one can afford to ignore these general-purpose technologies, which will soon transform the landscape of business and society. The companies which ignore would soon be obsolete, and there will be no scope of catching up later. To realize the

Trend data/statement	Surrogate implication	Source
The construction industry is one of the least digitized industries	Urgent need for digitalization	Koeleman <i>et al.</i> (2019)
95% of all data captured in construction and engineering industry go unused	Need for data management systems	Snyder <i>et al.</i> (2018)
Global spending in construction is expected to touch US\$ 17.5 trillion, with China, the USA and India leading the way and accounting for 57% of all global growth	Prospective growth of Indian companies; need for breakout in international competitiveness	Valente (2019), Bhattacharya <i>et al.</i> (2012), Momaya (2001, 2014)
6.5% compound annual growth rate (CAGR) in modular construction by 2026 is predicted	Shift to new technologies	Fortune Business Insights (2019)
About 90% of firms using prefabrication report improved productivity, improved quality and increased schedule certainty compare to traditional stick-built construction	Productivity through new technologies	Momaya (2001), Dodge Data and Analytics (2020)
14% of trades report prefabricating more than 50% of their work in the shop versus field	Increased use of mechanization	Dodge Data and Analytics and Autodesk (2018)
29% of firms are putting longer completion times into their bids for new work because of the lack of workers; 44% of firms report increasing construction prices due to labor shortages	Moving away from manual labor; initiatives to enhance productivity	AGC News (August) (2019), Momaya (2001)
52% of rework is caused by poor project data and miscommunication; US\$ 31.3bn in rework was caused by poor project data and miscommunication in the USA alone in 2018	Better coordination and management through digital means	Young (2019)
29% of firms report investing in technology to supplement worker duties	Use of technology to remove human intervention	Brown (2019)
60% of general contractors see problems with coordination and communication between project team members and issues with the quality of contract documents as the key contributors to decreased labor productivity	Better coordination and management through digital means	Dodge Data and Analytics and Autodesk (2018)
50% variation in productivity of two groups of workers doing identical jobs on the same site and at the same time. This gap in productivity was found to vary by 500% at different sites	Extensively varying productivity standards of manual interventions an irritant	Ellis (2019)
75% of construction companies use cloud storage	Use of digital data management	Matthews (2018)
4% increase in safety application usage in 2019; 63% of contractors are currently using drones on their projects; 37% of contractors expect to adopt equipment tagging by 2022; 33% of contractors expect to use wearable technology in the next three years	Need for digitalization and new technologies	Ellis (2019)

Table 1.
Snapshots on trends and changes expected in construction industry

comprehensive benefits of 4IR, it is also necessary to integrate the technologies across the entire value chain.

Most companies across the construction industry are still struggling to evolve a successful strategy for digital transformation to adopt 4IR technologies, despite the evident benefits they offer. Thus, the *need for an emergent strategy framework* for digital transformation

across operations, technology, personnel, regulation and other resources in the industry emerges as the prime objective of this study.

Research methodology

This study uses the integrative review methodology to synthesize a framework based on three pillars of strategy. This research methodology is particular suitable for an area which is nascent in its stages of development and has to draw from seminal and extant literature review papers. This enhances the rigor of combining diverse methodologies, which can combine both empirical and theoretical sources in an integrative review. Integrative review method holds the potential to allow for diverse primary research methods to contribute to evidence-based practice initiatives (Marabelli and Newell 2014). The purpose of using an integrative review method is to overview the knowledge base, to critically review and potentially reconceptualize, and to expand on the theoretical foundation of specific topics (Webster and Watson, 2002). Torraco (2005) recommends the integrative approach where the purpose is to assess, critique and synthesize from available seminal literature, thus enabling new theoretical frameworks and emerging perspectives.

The first two pillars in this study have been drawn and supported by seminal extant literature review on strategy. For newly emerging topics, the intention is rather to create initial credibility on new conceptual frameworks and theoretical models. This type of review can be identified in various business literature reviews (e.g. Covington, 2000; Gross, 1998; Mazumdar *et al.*, 2005). Synder (2019), too, has strongly argued the use of literature review in support of such a methodology. The third pillar proposed also utilizes the support of past literature on change and continuity, though its construct needed to be validated in the specific research context. This has been addressed through hypothesis testing of data obtained in a questionnaire survey on a five-point Likert scale, as elaborated in a subsequent section of this paper.

Development of framework for transformation strategy in AEEO

There have been several suggested models for digital transformation like the “BUILD” model proposed by Herbert (2017). Such models are generic and cannot be applied to the construction industry keeping their unique and complex nature in mind. Shaughnessy (2018) has suggested strategies based on an agile framework. Others (Mugge *et al.*, 2020; Brunetti *et al.*, 2020) have also argued for practical and tailor-made strategies. Verhoef *et al.* (2021) suggest an appropriate combination of organization structure and allied metric calibrations.

Also, the current study adopts the term architecture, engineering, construction and operations (AEEO) for discussing the integrated impact of 4IR across the entire industry value chain. As has been mentioned earlier, most companies across the AEEO value chain are still struggling to evolve a successful strategy for digital transformation to adopt 4IR technologies, despite the evident benefits they offer. The dilemma remains where to start and what to address first? Every company needs to figure out its strategy based on its nature, requirement and stage of maturity.

Having established already the “why,” i.e. need for adoption of 4IR in AEEO and the “who,” i.e. the stakeholders in the AEEO value chain, the focus needs to be on “what” and “how” of the strategic planning (refer Figure 1). For any company which is already operating in the value chain (and more so successfully), all proposed changes amount to disruption of business activities, which have the potential of stunting productivity and financial performance. This causes a major dilemma for the top management of the company translating into a lot of reluctance and hesitancy. Hence, the “what” and “how” of strategic planning shall depend solely on the company’s standing, skills, capabilities, the current line of

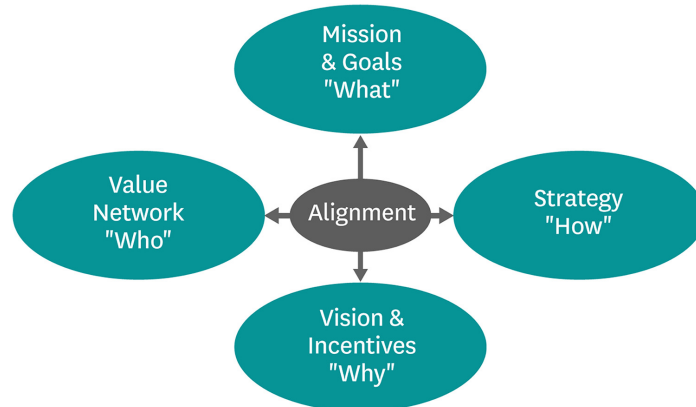


Figure 1.
Crafting a strategy

businesses and maturity of the company. Any transformation should take care of the existing line of businesses and retain existing value already created within the company. A well-crafted strategy provides a clear road map, consisting of a set of guiding principles or rules, that defines the actions people in the business should take and the things they should prioritize to achieve desired objectives. Knowledge and understanding of the 4IR landscape and strategic positioning will facilitate and enhance its effective implementation (Alade and Windpo, 2020).

Strategy as practice attempts to explain how managerial actors perform the work of strategy, both through their social interactions with other actors and with recourse to the specific practices present within a context. Hence, practices must be used to mold the context of the activity, leverage a new pattern of activities and to reconceptualize the rationale in which activities occur. As a new pattern of activities arises, this may create friction with the old practices, leading to their modification or alteration. The prevailing processes and practices within an organization are, therefore, expected to affect and conversely get affected by the changing patterns of activity. However, if planned judiciously, these inherited practices may be used to mediate between the proponents in modifying or leveraging new patterns of strategic activity (Hendry, 2000; Whittington, 1996, 2002). In the current scenario of digital transformation too, this interpretation should hold good. Given this understanding and the typical AECO context, it is suggested to craft a *strategy framework pillared on three key concepts*.

The first pillar: simple, actionable and agile strategy

Though the AECO value chain environment may not be very Volatile, it is Uncertain, Complex and Ambiguous, thereby having three attributes of the VUCA environment. Researchers have shunned complex strategy making in such environments (Sull and Eisenhardt, 2012). Here strategies that are formulated require to be simple, uncomplicated, flexible to respond quickly, accommodate options and what-if solutions. The managers too often root for simple and implementable action plans rather than complex strategizing, for achieving growth targets (Jarzabkowski and Whittington, 2008). They need to craft a handful of simple rules. Agile processes would need to be adopted and consistently applied, to provide an effective approach to address the constant change expected to be encountered (Sushil, 2005). Companies must be able to gauge the current and future levels of consumer-driven change, couple that with existing project and program management capabilities, and develop an action plan to deliver agile project management in the true sense. Simultaneously, there is a need to conceptualize the underlying strategic and organizational problems enough, for taking appropriate and effective action in the said situations. Agility would again be relevant

across the value chain with a purpose to respond to customers and relevant stakeholders quickly, on a real-time basis.

Embedding of an *AEEO Capability Centre (Transformation Command)* in the company structure and the value chain would be critical to facilitate the first pillar of simple, actionable and agile strategy as shown in Figure 2. This suggested transformation command will be similar to the concept of global capability centers that are common in the information technology (IT) industry (Ahuja, 2020). The transformation and integration of the complete AEEO value chain in alignment with 4IR would necessarily require it to be heavily dependent on digital capabilities. The initiative should be led by a group of persons who understand the vision of the company and its intricacies for structuring system. A team of professionals from diverse backgrounds, namely strategists, program and project managers, cognitive and systems thinkers, data analysts and data scientists, digital operatives and robotic programmers have to come together and constitute this group. This team would be capable of spotting new projects while increasing visibility and transparency across the organization. The business environment which is expected to be full of competitive and economic uncertainties has to be led with enterprise-wide capabilities. This group would also lead the business and strategic initiatives within the organization, like a central core and guidance group.

This is suggested because the transformation of the business will mandate a structural shift in work scope, methods, talent/ skill needs and benefits realized. The AEEO capability centers will be responsible to strategize, ideate and create road maps on the implementation of technologies related to 4IR across the entire value chain. The operations, executive functions and processes, however, will continue to be decentralized.

Attention has to be centered on four major areas, namely identifying immediate objectives, managing risk, ensuring coordinated and team approach, and getting results across the entire value chain. The command would also identify the priority areas and phases of the penetration of the digital initiatives. The role and expertise of the command can extend to hand-holding, mentoring, evolving new practices and tools and techniques. The immediate areas of intervention in the value chain shall be identified on basis of competencies available

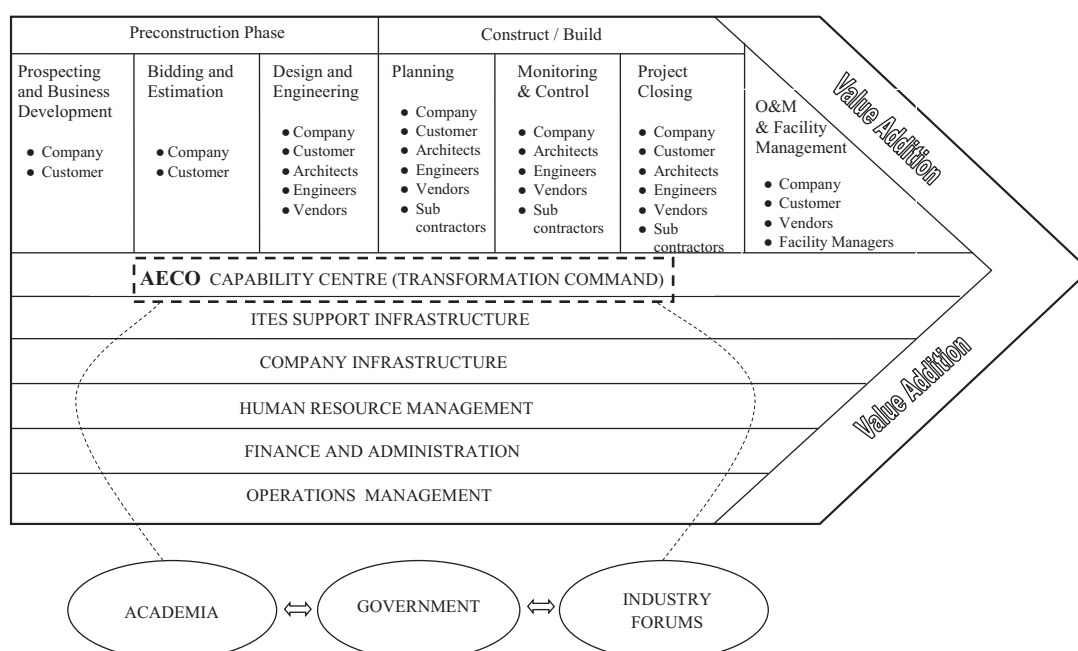


Figure 2.
AEEO industry value
chain and ecosystem

and the urgency to record quick wins and celebrate (Kotter, 2007). The transformations have to be ongoing, continuous and everlasting, thereby charting new territories and new opportunities on a self-sustaining basis. It should not be limited to a few isolated successes and rather look to slowly but steadily integrate all stages of the value chain. It is also essential for all the initiatives and successes to be replicable and scalable.

The transformation command itself is also expected to grow and mature over a while. For this, the group needs to evolve partnerships across the business ecosystems which should include academia, government agencies and various industry forums. This will enable them to evolve, keep a pulse on the market trends, new technologies and customer requirements. This would also catalyze the development of additional new capabilities. Our proposition for the first pillar of digital transformation for the AECO value chain is as follows:

Proposition 1. The strategy for digital transformation for the AECO value chain shall be simple, flexible, agile and actionable in nature.

An appropriate structure of an organization is required to render flexibility and agility in its responses. For achieving the same, a sub-proposition is as follows:

Proposition 1a. The digital transformation initiatives shall be crafted, led and guided by an interdisciplinary capability center also known as the transformation command.

Before crafting the strategy, some of the key questions that the transformation command would need to answer would be as follows:

- (1) Which 4IR applications/technologies does the company already use?
- (2) What are the focus areas in which the company is technologically and culturally ready to adopt 4IR concepts? This will help in evaluating the routes, priorities and phasing.
- (3) Which 4IR applications/technologies would bring quick and quantum benefits vs minimal efforts/resources to give quick wins?
- (4) Which 4IR applications/technologies are expected to bring long-term sustainable competitive advantage and differentiate it from competitors?
- (5) Which current and prospective projects are opportunities to leverage or skill up the 4IR initiatives within the company?

To obtain specific answers to the above questions, it would be required to map the dynamic capabilities of the company, which forms the second pillar of the strategy framework.

Second pillar: strategy based on dynamic capabilities

Dynamic capabilities refer to a subset of capabilities directed toward strategic change, both at the organizational and individual unit level. These enable companies to create, extend and modify their business models, including those through alterations in resources, operating capabilities, scale and scope of businesses, products, customers, ecosystems and other features of their external environments (Teece, 2018; Helfat and Winter, 2011; Zollo and Winter, 2002). Dynamic capabilities of an organization include the sensing, seizing and transforming needed to craft business transformations. Teece (2017) suggests that they can be categorized according to three general types of functions, namely sensing new opportunities and threats, seizing new opportunities through business model design and strategic investments, and transforming or reconfiguring existing business models and strategies. Dynamic capabilities are linked in part by organizational routines and processes,

the gradual evolution of which is punctuated by nonroutine managerial interventions that may become necessary from time to time. For the digital transformation, this may well become one of the most critical features.

In practical terms, business transformations, especially those involving a novel field of technology; a different customer base; organizational reengineering; some combinations of these and other disruptive changes within an existing business are unlikely to succeed without major financial resources and strong commitment. Business model transitions that fit comfortably with the existing business are observed to be far easier to implement (Teece, 2018). It is also suggested that small transitions can enhance value capture, which is something that matches the idea of judicious confluence of continuity and change, which would be discussed subsequently.

Dynamic capabilities would also enable an enterprise to upgrade its ordinary capabilities and the capabilities of partners and collaborators, toward high-payoff endeavors. This is completely in alignment with the vision of digital transformation in the AEEO value chain. This requires developing and coordinating the company's and partner's resources to address and shape changes in the business environment. All this, however, may affect the timeline of implementation. This is because the strength of any company's dynamic capabilities determines its speed and degree and associated costs of aligning its resources. It logically includes aligning its business model(s) with customer needs and aspirations. To achieve this, companies must be able to continuously sense and seize opportunities. This may involve a phase-wise transformation of the organization and culture to proactively address novel threats and opportunities, as and when they arise. Hence, our second proposition:

Proposition 2. The digital transformation strategy shall be crafted based on the dynamic capabilities of a company.

Third pillar: strategy of transformation on the move: a mix of change and continuity

The arrival of new general-purpose technology (as in the case of 4IR) opens opportunities for radically new business models, to which corporate strategy is required to respond. A new wave of business model innovation leads to the emergence of new services and ways of doing business. Once in place, a business model shapes strategy and vice versa, in a reciprocal relationship. They constrain some actions while facilitating some others. In the event of a conflict between strategy and the business model, the onus is on top management to determine which of the two should change. It often takes time for the business model innovation to catch up to technological possibilities because business models are more context dependent as compared to technology. All these may cause conflicts between the old and new order within the company.

Contradictions are grounded within the internal dynamics of the organization, arising from dilemmas over past and projected future range of activities. There is always a need to accommodate and mediate between constituents to promote a more collective capacity for change (Jarzabkowski, 2003). Since contradictions and mediation are important components of change, distributed and participative approaches to resolve these differences act as levers of change.

Theories of change and change management have long been topics of research in the field of management and to a limited extent in strategic thinking. The globalization process of the 1990s and more recently the advent of 4IR and digitalization has made the business environments highly turbulent. Rapid change has generated immense interest from strategic thinkers and practitioners. Over the years, diverse theories, such as crafting strategy, strategic flexibility, complexity and chaos, strategic change and transformation, blue ocean strategy, etc. have evolved in the ever-dynamic business environment. Management literature has extensive frameworks and models, in several books and journals on change,

change management and organizational change (Carter and Varney, 2018; Rosenbaum *et al.*, 2018; Varsos and Assimakopulos, 2016; Burke, 2013; Bamford and Forrester, 2003; Washington and Hacker, 2005; Oakland and Tanner, 2007). However, the record of success in change management has hardly been encouraging so far (Hughes, 2011; Burness, 2011; Beer and Nohria, 2000; Sturdy and Grey, 2003). Several researchers and theorists have suggested multiple novel ways of effecting change outcome also (Gondo *et al.*, 2013; Pettigrew, 2000; Beer and Nohria, 2000; Tsoukas and Chia, 2002).

It is, however, interesting to note that crafting any strategy involves synthesizing one which is suitable to the context. Such a strategy should look to accommodate the existing objectives of a live/running organization, along with the intended strategy and/or a reactive/adaptive strategy in harmony with the change in the business environment (Mintzberg, 1988). Usually, a company that is bidding for transformation looks to continuing with the existing business successfully too. The bigger the inertia of motion (continuity), the stronger the continuity momentum. It would be extremely difficult to first stop a moving vehicle and then change its course. This would result in a waste of effort and resources. This logically forms the origin of the school of thought, proposing a judicious mix of continuity and change to provide stability and dynamism simultaneously. Following this line of thought, companies require to create a strategy road map to address the so-called “confluence” of continuity and change at multiple levels (Sushil, 2012).

Nasim and Sushil (2011) argue that “managing change is invariably managing paradoxes.” They treat the balance of *continuity* in terms of alignment orientation, rigor and discipline and *change* in terms of adaptive orientation, flexibility and agility. The mix is viewed as analogous to a “flowing stream” which reinforces the concept of natural growth and development. The concept of balancing change with continuity has to gain prominence and that too logically, in business environments that evolve continuously (Sushil, 2005, 2012; Gupta, 2016; Sutherland and Smith, 2011; Malhotra and Hinings, 2012; Brown and Eisenhardt, 1997; Leana and Barry, 2000). Mintzberg *et al.* (1998) have also highlighted the need for balancing change with continuity, as per business objectives. In his discourse on planned vs emergent strategy, Mintzberg (1988) established that strategy making is both “deliberate and emergent” and hence needs to be crafted rather than just planned. According to him, a fundamental dilemma of strategy making is the need to reconcile the often conflicting forces of stability and change. While there is a need to focus efforts and gain operating efficiencies, on the one hand, adapting and maintaining pace with a changing external environment needs to be taken care of on the other hand (Mintzberg, 1988, p. 82). Internal continuity has to be maintained to competencies and organization culture, while externally, it should cater to the new opportunities and customer needs (Pettigrew, 2000; Drucker, 1999). Collins and Porras (1994), too, have argued on preserving the core while changing continuously. The integration of two opposing forces changes and continuity is extremely challenging but would be worthwhile in terms of payoffs. Makinen (2017) concludes in his study in the specific context that process is initially incremental rather than transformative, it constructs the foundation which is not a deterministic, carefully preplanned project, but it is rather highly emergent and iterative in nature.

Customer requirements need to be the basis for a sound framework for identifying the areas of continuity and change and helping the company integrate upfront for effective strategy formulation. Such logical alignment would result in higher customer satisfaction and thereby, competitive advantage. Conversely, expanding the pool of business and range of services also provides an avenue for better growth.

A construct incorporating the change and continuity forces in the context of digitalization in the AECO value chain, along with the motivations/takeaways and challenges/downsides has been adapted/collated from the literature on related works of various researchers in strategy and 4IR (Bhattacharya *et al.*, 2020; Oesterreich and Teuteberg, 2016; Wirtz *et al.*, 2018;

Adetunji *et al.*, 2008; Ahuja *et al.*, 2017; Nasim and Sushil, 2011; Yoon and Chae, 2009; Dawes, 2009; Momaya, 2011; Riley, 2007). These have been exhibited in the construct in Figure 3. A few prominent literature review learnings from the references on continuity and change cited above have been discussed in the following paragraphs. Extensive discussions of the motivations and challenges have been avoided, which can be a topic of further in-depth research.

In the context of the AEEO value chain, there can be many forces that contribute to the inertia of continuity, as shown in the construct. Large size of customer base may already exist, whom a company needs to continue serving during and even after implementing the transformation. But these products or services may be required to be delivered differently and more efficiently. In business and commercial organizations, inertia may creep in due to the fear of losing a *large customer base*. It acts as a major deterrent for change and thus becomes a strong continuity force.

Most companies may have varying extents of already established *huge physical infrastructure*, which could be in danger of becoming redundant due to transformation. Thus, the bigger the physical set-up already in place, the larger would be the force of continuity. The *technologies, equipment and hardware* being utilized also become a continuity factor, especially when there is a chance of them being rendered redundant. *Legacy processes or existing processes* of service delivery are identified as another major continuity force in implementation. Such a preexisting network of supply chain/ delivery processes, standard operating procedures or the actors involved therein contribute to greater continuity forces in an existing company.

Core competencies rooted in the old system can be major continuity forces resisting change within a company. These may have a major and a valid claim within the company of earning its bread and butter, especially when they are currently delivering *superior financial results*. The *culture* of any organization is the major driving force for maintaining continuity. However, a positive attitude, and progressive mindset inculcated as a legacy in the culture, can also facilitate the transformation. Some *groups or lobbies* associated with the company can still be expected to be most resistive to change, indicating a higher level of continuity on the culture front. Inherently, there is a need to build a culture that supports the organization's strategy and continues to integrate the work processes with company culture. Companies focusing on organization culture are expected to be five times more successful in digital transformation as compared to others (BCG, 2020).

However, as the transformation progresses, over some time, the influence of such forces is expected to diminish, other than ones involving *competencies and organization culture*. A new order is expected to replace most of them consequent to the transformation. Therefore, the implementation needs to be gradual and not abrupt.

On the other hand, some other forces push a company to change, as is exhibited in the construct (Figure 3). *Information Technology (IT)-enabled processes* in general have already recorded significant successes within various domains and even promised to disrupt how business is done. It also lends the characters of efficiency, responsiveness, accountability and democratization to the work processes (Moon, 2002). As a result, businesses have undergone a rapid and continual change (Stojanovic *et al.*, 2006).

The process of *globalization* is pervasive. Globalization has placed strong pressures on companies to compete for trade flows, investments and resources (Butzbach *et al.*, 2020; Parente *et al.*, 2018; Rodrik, 2018) and is a strong force outside, especially given the industry trends internationally (Verbeke *et al.*, 2018). It has created possibilities for interactive initiatives, putting companies worldwide under pressure to change how to conduct their business. This challenges all the players to set the bar for higher thresholds of competitiveness and to sustain competitive advantage.

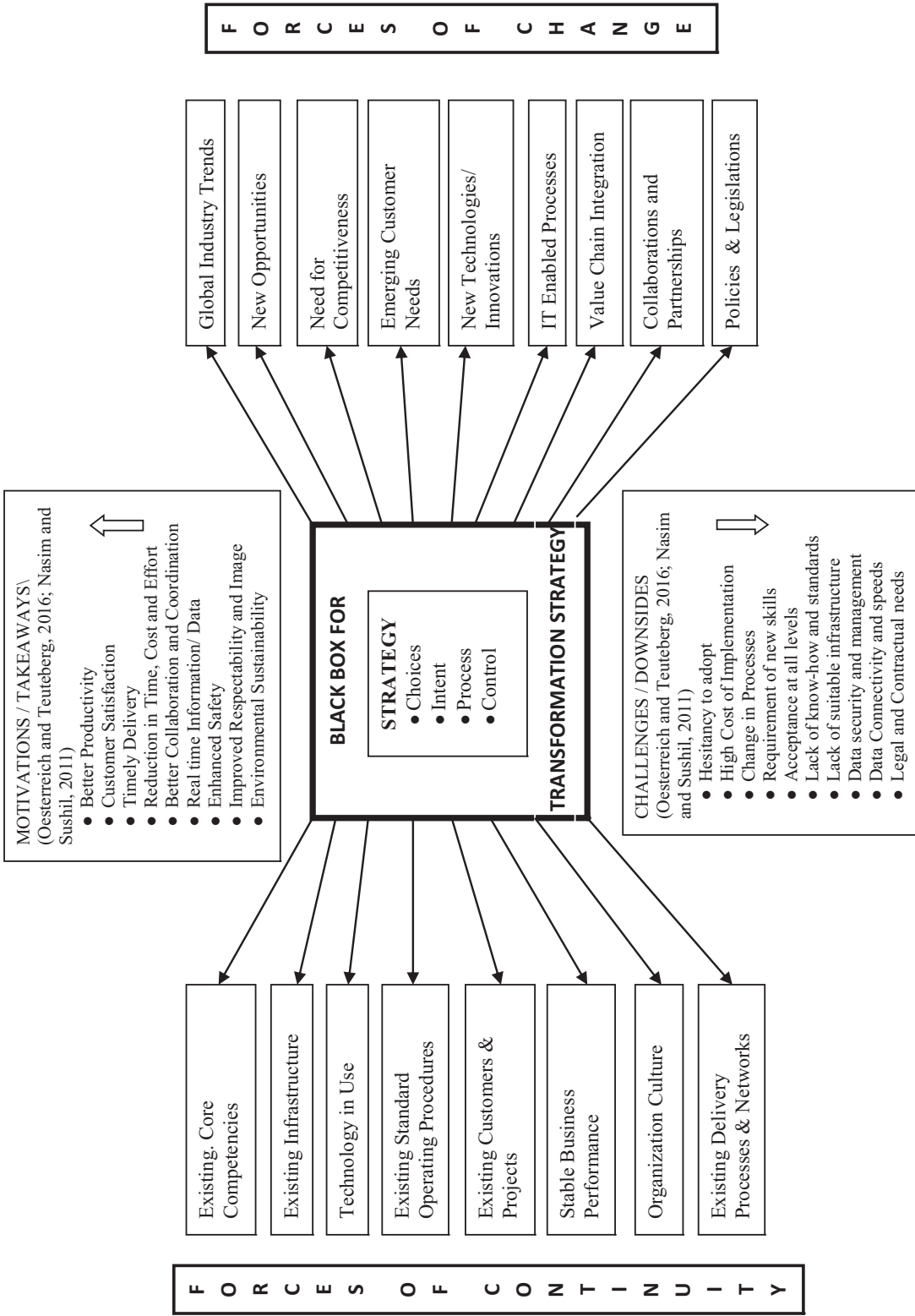


Figure 3.
Construct for change
vs continuity for digital
transformation in
AECO industry
value chain

Despite all localization forces, the *customer's expectation* for better service deliveries owing to rapid technological changes has been a major driver to adopt digitalization. *New opportunities for business* with a novel range of services are expected to be a major harbinger of change. A larger pool of work scope instead of individual products and services can help to tap new opportunities and changing customer requirements. A new customer requirement is likely to lead to further development of internal capabilities and skills.

Integration of the entire value chain unfolds a plethora of advantages in terms of democratization of information and synergies in operations (Lee et al., 2018). While the vendors/suppliers may benefit from the improved, transparent and efficient system, the company itself would gain in terms of cost efficiencies and customer satisfaction by delivering as per exact needs. For all the stakeholders, an integrated platform promises a wide array of collaborations, which can be mutually beneficial. *Collaborations* can be of many kinds including sharing of resources and infrastructure. This will contribute significantly toward faster growth and sustainability. Services running on the cloud, using big data, or artificial intelligence have the potential for a multitude of possibilities in digital entrepreneurship (Giones and Brem, 2017).

New technologies have been the prime force of change across all domains (Mahardika et al., 2019; Preece, 1988; Hattori and Tanaka, 2016). The adoption of cyber-based technologies to conduct business and deliver services has become a global driver of change. Technology is expected to facilitate, enable and empower. Industry experts and researchers believe that technology will continue to unfold, evolve and drive programs till eternity.

Government policies and legislation can be yet another major force for change. By creating a favorable business environment for adopting 4IR at the national level, the government/statutory bodies can ensure and catalyze the change. Institutional as well as *cyber-infrastructure*, along with essential electricity and data connectivity would be critical success factors for a national-level change facilitator.

Therefore, our third proposition based on the forces of change and continuity is as follows:

Proposition 3. The digital transformation shall be successfully implemented by managing a judicious confluence of change and continuity forces.

Thus, the proposed framework for digital transformation in the AEEO value chain based on the three pillars identified and discussed above can be represented below as in Figure 4.

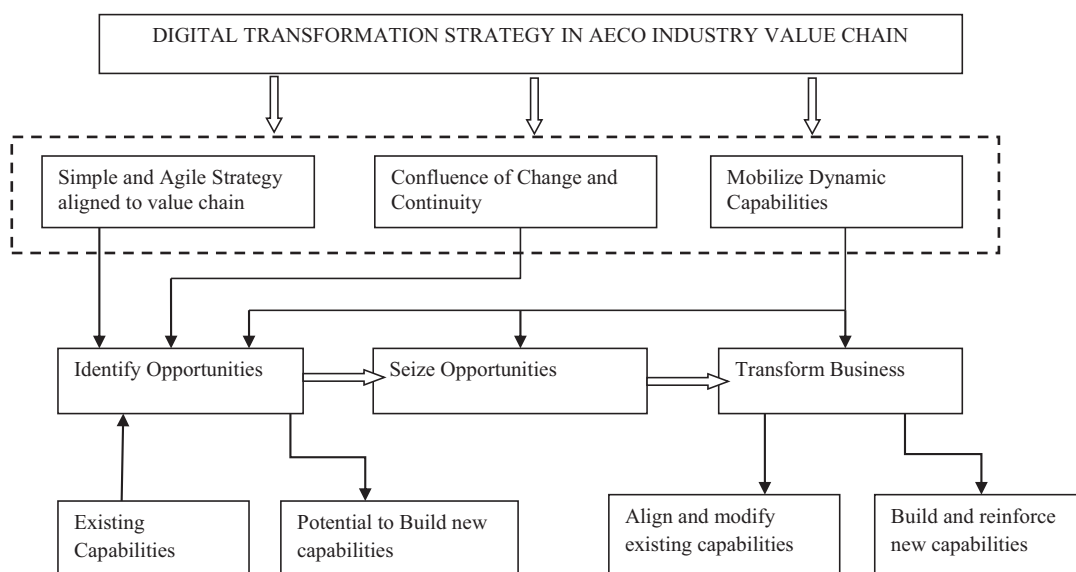


Figure 4.
Proposed strategy
framework for
transformation

Validation of the proposed framework

Among the three suggested pillars of the framework, the first two are strongly supported by the existing seminal strategy literature, as cited earlier. The third pillar which was extremely AEEO context specific required validation checks. The continuity and change forces identified from diverse literature reviews have been validated through a questionnaire survey. The survey was administered to senior and mid-senior AEEO industry professionals to respond to the themes that emerged out of the literature review. They were asked to rate their responses on issues on a five-point Likert scale. They were also provided with ready theoretical definitions of various terms for better understanding. Around 125 questionnaires were distributed, and 42 valid responses were received.

Descriptive statistics like a high mean score (more than three) in conjunction with median or mode (four and above) of the response distribution endorsed the significance of the element issues under analysis and hence provided a fair basis of acceptance. To further authenticate the survey results, a single tail “z” statistic test has been used to compare the mean value of each of the elements with a specified constant mean test value of 3, at a 2.5% level of significance. Since the questionnaire responses ranged from strongly disagree (1) to strongly agree (5), a mean value of more than 3 was assumed to be a reasonable test value for hypothesis testing.

The major conclusions that can be drawn from the survey have been elicited in the following sections.

Use of 4IR applications in AEEO:

The first set of hypotheses was to establish the immediately relevant 4IR technology applications and their utility in the AEEO value chain. They had been enumerated as follows:

Null hypothesis (H0): (4IR application name) does not have a significant role in the digital transformation of the AEEO value chain.

Alternate hypothesis (H1): (4IR application name) has a significant role in the digital transformation of the AEEO value chain.

An application would be assumed to be having a significant and valid role to play if the value of the z-statistic exceeded 1.96 (significance level lesser than 0.025 or more than 97.5% confidence level). The results of the analysis have been summarized in [Table 2](#).

In general, all respondents were upbeat about the use of 4IR technologies in the AEEO value chain. They view the transformation to be *inevitable and would positively impact productivity* in a significant way. Around 50% of respondents feel that cost of implementation is a factor currently. Few have suggested that over time, these technologies will become economical. The suggested list of technologies in AEEO for hypothesis testing was collated from relevant past research ([Oesterreich and Teuteberg, 2016](#); [Woodhead et al., 2018](#); [Wirtz et al., 2018](#); [Deloitte, 2015](#); [Büyükoçkan and Göçer, 2018](#)). Among the technologies available *Virtual/Augmented Reality; Simulations; Cloud Computing; Internet of Things; Data Analytics; Machine Learning or Artificial Intelligence; Cyber Security; Automation/ Robotics; Building Information Modeling and Drones/Sensors/Wearable devices* have all emerged as significant applications through the hypothesis testing of the responses. On an immediate basis, *Building Information Modeling; Augmented Reality; Machine Learning and Drones/ Sensors* are the top four identified applications for immediate deployment, as per the survey carried out.

The significance of continuity and change forces

On basis of the literature review cited earlier, eight forces of continuity and nine forces of change had been identified. Hypothesis testing of the data collected in the questionnaire

Applications	Mean score (\bar{x})	Median value	Mode value	Std. dev. (s)	Calculated z-statistic	Null hypothesis Status H0	Contributes significantly
Advanced manufacturing/3D printing	3.48	4	4	1.77	1.75	Accepted	No
Additive manufacturing	3.29	3	3	1.42	1.30	Accepted	No
Augmented/ virtual reality	4.10	4	4	1.18	6.04	Rejected	Yes
Simulation/ predictive modeling	3.90	4	5	1.48	3.97	Rejected	Yes
Cloud computing	3.95	4	4	1.30	4.74	Rejected	Yes
Internet of Things	3.90	4	4/5	1.48	3.97	Rejected	Yes
Blockchain	3.43	3	3	1.45	1.91	Accepted	No
Data analytics	3.81	4	3/5	1.46	3.60	Rejected	Yes
Machine learning	4.00	4	4	1.10	5.92	Rejected	Yes
Cyber security	4.00	4	5	1.41	4.58	Rejected	Yes
Automation/ robotics	3.95	4	4	1.14	5.42	Rejected	Yes
Building information modeling	4.52	5	5	1.31	7.52	Rejected	Yes
Drone sensors, wearable devices	4.29	5	5	1.42	5.85	Rejected	Yes

Note(s): Mean test value = 3; cut off “z” statistic value > 1.96 at 2.5% level of significance

Table 2.
Summary of perceived
significance of
applications of 4IR
in AECO

survey was carried out, on both these sets of forces. The hypotheses had been enumerated as follows:

Null hypothesis (H0): (The name of change/continuity force) does not significantly impact the 4IR digital transformation of the AECO value chain.

Alternate hypothesis (H1): (The name of change/continuity force) significantly impacts the 4IR digital transformation of the AECO value chain.

Thus, a force would be assumed to be having a significant and valid impact if the value of the z-statistic exceeded 1.96 (significance level lesser than 0.025 or more than 97.5% confidence level), indicating a higher level of effect. The results have been summarized in [Tables 3 and 4](#), as exhibited below.

All the continuity forces listed, except *stable business performance*, tested were significant in the current context of the AECO value chain. The possible interpretation for this is that a company doing well would be expected to have spare resources to invest in digital transformation. Conversely, a company not performing well may turn to digital transformation to change its fortunes.

The top three continuity forces which offer resistance to change are *organization culture*, *existing delivery processes and networks*, and *existing standard operating procedures*. All these are behavioral aspects involving stakeholders, who have become comfortable within the existing system. Individuals are often led to deep-seated fears or insecurities about their skills and abilities, fear of the unknown and new environment or fear of losing control, which discourages them to favor change ([Kegan and Lahey, 2001](#)). The human mind also gets accustomed to a way of working and in a comfort zone over some time. Several researchers have highlighted deep-rooted prejudices associated with established

ECAM

Table 3.
Summary of perceived
significance of
continuity forces
in AEEO

Applications	Mean score (\bar{x})	Median value	Mode value	Std. dev. (s)	Calculated z-statistic	Null hypothesis Status H0	Contributes significantly
Existing core competencies	3.43	4	3	1.16	2.38	Rejected	Yes
Existing infrastructure	3.38	3	4	1.20	2.05	Rejected	Yes
Technologies in use/deployed	3.43	4	3/4	1.33	2.10	Rejected	Yes
Existing SOPs	3.67	4	5	1.28	3.38	Rejected	Yes
Stable business performance	3.05	3	3	1.20	0.26	Accepted	No
Organization culture	3.81	4	4	1.03	5.09	Rejected	Yes
Existing delivery processes and networks	3.71	4	5	1.19	3.89	Rejected	Yes

Note(s): Mean test value = 3; cut off “z” statistic value > 1.96 at 2.5% level of significance

Table 4.
Summary of perceived
significance of change
forces in AEEO

Applications	Mean score (\bar{x})	Median value	Mode value	Std. dev. (s)	Calculated z-statistic	Null hypothesis Status H0	Contributes significantly
Global industry trends	4.00	4	5	1.14	5.68	Rejected	Yes
New opportunities	3.81	4	4	1.08	4.87	Rejected	Yes
Emerging customer needs	3.52	3	4	1.08	3.15	Rejected	Yes
New technologies/innovations	3.90	4	4	1.09	5.37	Rejected	Yes
Needs of competitiveness	4.05	4	5	1.02	6.63	Rejected	Yes
IT-enabled processes	3.62	4	4	0.92	4.36	Rejected	Yes
Value chain integration	3.67	4	4	1.06	4.06	Rejected	Yes
Collaborations and partnerships	3.33	3	3	1.15	1.87	Accepted	No
Policies and legislations	3.38	4	4	1.20	2.05	Rejected	Yes

Note(s): Mean test value = 3; cut off “z” statistic value > 1.96 at 2.5% level of significance

organization practices and culture (Hellriegel and Slocum, 2011; McLaughlin *et al.*, 2008; Senarathna *et al.*, 2014; Sethi, 2003). These require a substantial realignment of attitudes and mindset.

The change forces too, other than *collaborations and partnerships*, were all hypothesized to have a significant impact. This again may be specific to the context of AEEO because it still is a very fragmented value chain. Collaborations and partnerships may at a later point in time become significant, once the integration of the value chain has been effected. The leading drivers of change are *needs of competitiveness*; *global industry trends* and the advent of *new technologies/innovations*. Any company in the globalized and business ecosystem can survive

and succeed, only by staying ahead of peer competition. These top three change forces are critical toward ensuring this sustainable competitive advantage in the AECO value chain, as has been duly underscored by prior research also (Bhattacharya *et al.*, 2020; Ambastha and Momaya, 2004; Momaya and Ambastha, 2005).

Conclusions and research implications

4IR involves running businesses by adopting digital technologies that can help companies create integration between their machinery, supply systems, production facilities, final products, employees and customers to gather and share information or data on a real-time basis. The revolution opens possibilities for modern techniques to support key components within the industry. The AECO industry too should not remain insulated from these developments. In general, all respondents of the questionnaire survey are upbeat about the use of 4IR technologies in the AECO value chain and industry value system. They view the transformation to be inevitable and would positively impact productivity, strategic flexibility and international competitiveness in a significant way. Around 50% of respondents feel that cost of implementation is a key barrier currently. Few have suggested that over a while, these technologies will become economical. However, currently, the research on applications of 4IR in AECO is still at a nascent stage.

The AECO value chain is typically characterized by problems of complexity, uncertainty, fragmented supply chain, short-term thinking and conservative culture. The projects too are complex in nature due to the involvement of several stakeholders. Given this unique nature, a suitable integrated framework should ideally address all the identified issues and bottlenecks.

The conceptual framework presented in the present study has integrated three diverse ideas of strategy, named as pillars, to deliver digital transformation in the AECO value chain. These ideas are based on *creating a strategy of simple rules; a strategy based on dynamic capabilities and a strategy having a confluence of change and continuity*.

Strategy of simple rules is less complicated, flexible responses to any circumstance or context, while accommodating multiple and what-if solutions. These are expected to provide an effective approach to address constant change in a transient business environment. For this purpose, *AECO Capability Centers or Transformation Commands* have been proposed. Strategists; program and project managers; cognitive and systems thinkers; data analysts and data scientists; digital operatives and robotic programmers have to come together and constitute this group. The group will be responsible to strategize, ideate and create road maps on the deployment of technologies related to 4IR across the entire AECO value chain and system.

Dynamic capabilities refer to a subset of capabilities directed toward strategic change, both at the organizational and individual unit level. Mobilizing the dynamic capabilities of a company can *enable the creation, extension and modification of business models*, through alterations in its resources; operating capabilities; scale and scope of businesses; products; customers; ecosystems and other features of their external environment.

A confluence of change and continuity provides a feasible and practical path to carry out the 4IR digital transformation, without disrupting the current business revenues. In the quest for transformation, the internal operations and current sources of revenues get ignored. There is a need to ensure that the strategy crafted and implemented is suitable to the context and maturity of the company in question. The transition journey while *judiciously balancing the conflicting forces of continuity and change* is expected to be smooth, seamless and successful. These conflicting forces may be both internal and external to any organization. Empirical studies carried out on the change continuity construct evidence that the top three continuity forces which offer resistance to change are *organization culture, existing delivery*

processes and networks, and existing standard operating procedures. The leading drivers of change are *needs of competitiveness; industry trends worldwide* and the advent of *new technologies or innovations*.

An exact strategy based on the “three-pillar framework” would need to be crafted for particular companies based on the detailed analysis carried out as per the proposed framework. The framework can help guide a company through its various stages of transformation maturity based on diagnostics of trends in competitiveness (Momaya, 2001, 2011). It will give a company the option to choose the areas of intervention based on its priorities, preferences, choice and comfort. As a downside, the flexibility offered can also result in complacency creeping into the organization. The company has to be wary and alert to this aspect. As a mitigating measure, there is a need to create a very strong vision, which needs to be reinforced regularly in no uncertain terms within the organization, by the capability centers proposed. Depending on the performance goals, there would be a need to fix metrics for the transformation. Discussion on the same has currently been kept beyond the scope of this study. Likewise any strategy, these have to be quantifiable, objectively stated and fixed time frames (also referred to as SMART objectives) to create a sense of urgency. There should be an attempt to record quick wins and celebrate them as a motivator for the workforce. Successful performances and target achievements need to be rewarded with more delegation of responsibilities. This will positively strengthen the organization culture and encourage rapid adaptation of the new processes. Similar to any other quality management process, the motto needs to be continuous improvement, albeit incremental (like Kaizen).

It is expected that slow enhancements in international competitiveness (Momaya, 2011) of the AECO industry would be a vexing problem for many developing countries such as India (Manda and Dhaou, 2019; Bhattacharya *et al.*, 2012, 2021) as they are still investing significantly in infrastructure. The potential of 4IR initiatives being leveraged to address problems like the following can have very contentious and useful implications.

- (1) Massive deficits on international trade (e.g. reflected in net forex losses) due to excessive dependence on imported inputs in machinery, plant, etc. is an issue in several countries. 4IR should look to address them and overcome these hurdles by enhancing productivity exponentially, without getting weighed down by the barriers.
- (2) For several reasons known for decades, most AECO companies in developing nations have been less capable of investing in innovation and leveraging capabilities for exports (Bhat and Momaya, 2020). This mindset needs to be addressed.
- (3) To mitigate problems of international competitiveness (Momaya, 2001), early exposure and experiences for young engineers of AECO firms become very critical. Education and training infrastructure and support from technological institutions are crucial to this need.
- (4) Specific to digital transformation in AECO, there is a need to integrate efforts across the value chain; compilation of success stories; realization of lean objectives or innovative models for new business and start-ups. This will help create knowledge resources on a prospective basis and realize benefits in the long term.

Limitations and topics for further research

This study, however, has a few limitations. The dynamics within the project portfolio have not been kept within the ambit of this study. Given the multidimensional nature of the constructs and framework involved, theoretical research is required to be carried out to strengthen and further refine them. The propositions too would need further testing to be

extended and replicated in various contexts. The interplay of the continuity and change forces may not have been completely revealed in the hypothesis testing carried out as these may tend to counter each other's effects. This would require advanced statistical analysis to gain valuable insights.

The current validation relied on the opinion and perspectives of respondents. The responses of practicing professionals were limited to the Indian context. Hence, their opinion is expected to be influenced by their domain and geography of experience.

Despite the above limitations, this perspective paper indicates several high potential topics to revive international competitiveness (IC) of the AECO companies and industry value chain by leveraging not only 4IR but more sustainable concepts such as flexibility for *Society 5.0* (Keidanren Policy and Action, 2016) that aged societies in countries such as Japan are piloting. The record of firms of emerging countries may not be remarkable, but the future can be shaped for bigger contributions if revolutionary innovations such as 4IR, digital platforms and their drivers such as "Sharing Economy" are adapted strategically for better future international competitiveness, prosperity and sustainability (Momaya, 2019).

References

- Adetunji, I., Price, A.D.F. and Fleming, P. (2008), "Achieving sustainability in the construction supply chain", *Proceedings of the ICE - Engineering Sustainability*, Vol. 161, pp. 161-172.
- AGC News (August) (2019), "Eighty percent of contractors report difficulty finding qualified craft workers to hire as firms give low marks to quality of new worker pipeline", available at: <https://www.agc.org/news/2019/08/27/eighty-percent-contractors-report-difficulty-finding-qualified-craft-workers-hire-0> (accessed on 25 January 2020).
- Ahuja, L. (2020), "Accelerating transformation", *Innovation Black Book Enterprise 4.0*, Wiley Innovation Advisory Council, pp. 421-430.
- Ahuja, R., Sawhney, A. and Arif, M. (2017), "Prioritizing BIM capabilities of an organization: an interpretive structural modeling analysis", *International Journal of Sustainable Built Environment*, Vol. 6 No. 1, pp. 69-80.
- Alade, K. and Windapo, A. (2020), "4IR leadership effectiveness and practical implications for construction business organisations", in Aigbavboa, C. and Thwala, W. (Eds), *The Construction Industry in the Fourth Industrial Revolution*, Springer, Cham, pp. 62-70, doi: [10.1007/978-3-030-26528-1_7](https://doi.org/10.1007/978-3-030-26528-1_7), CIDB 2019.
- Ambastha, A. and Momaya, K. (2004), "Competitiveness of firms: review of frameworks and models", *Singapore Management Review*, Vol. 6 No. 1, pp. 45-61.
- Axelsson, J., Fröberg, J. and Eriksson, P. (2019), "Architecting systems-of-systems and their constituents: a case study applying Industry 4.0 in the construction", *Systems Engineering*, Vol. 22 No. 6, pp. 455-470, doi: [10.1002/sys.21516](https://doi.org/10.1002/sys.21516).
- Azhar, S. (2011), "Building information modeling (BIM): trends, benefits, risks, and challenges for the AEC industry", *Leadership and Management in Engineering*, Vol. 11 No. 3, pp. 241-252.
- Bahrin, M.A.K., Othman, M.F., Azli, N.H.N. and Talib, M.F. (2016), "Industry 4.0: a review on industrial automation and robotic", *Jurnal Teknologi*, Vol. 78, pp. 137-43.
- Bamford, D.R. and Forrester, P.L. (2003), "Managing planned and emergent change within an operations management environment", *International Journal of Operations and Production Management*, Vol. 23 No. 5, pp. 546-64.
- Bazjanac, V. (2006), "Building information models and their relevance to building construction", *Clients Driving Innovation: Moving Ideas into Practice*, Cooperative Research Centre (CRC) for Construction Innovation.

- BCG (2020), "How to drive digital transformation: culture is the key", available at: <https://www.bcg.com/capabilities/digital-technology-data/digital-transformation/how-to-drive-digital-culture> (accessed 12 March 2021).
- Beer, M. and Nohria, N. (2000), "Cracking the code of change", *Harvard Business Review*, Vol. 78 No. 3, pp. 133-141.
- Bhat, S. and Momaya, K.S. (2020), "Innovation capabilities, market characteristics and export performance of EMNEs from India", *European Business Review*, Vol. 32 No. 5, pp. 801-822, doi: [10.1108/EBR-08-2019-0175](https://doi.org/10.1108/EBR-08-2019-0175).
- Bhattacharya, S., Momaya, K.S. and Iyer, K.C. (2012), "Strategic change for growth: a case of construction company in India", *Global Journal of Flexible Systems Management*, Vol. 13 No. 4, pp. 195-205, doi: [10.1007/s40171-013-0020-2](https://doi.org/10.1007/s40171-013-0020-2).
- Bhattacharya, S., Momaya, K.S. and Iyer, K.C. (2020), "Benchmarking enablers to achieve growth performance: a conceptual framework", *Benchmarking: An International Journal*, Vol. 27 No. 4, pp. 1475-1501, doi: [10.1108/BIJ-08-2019-0376](https://doi.org/10.1108/BIJ-08-2019-0376).
- Bhattacharya, S., Momaya, K.S. and Iyer, K.C. (2021), "Bridging the gaps for business growth among Indian construction companies", *Built Environment Project and Asset Management*, Vol. ahead-of-print No. ahead-of-print, doi: [10.1108/BEPAM-08-2020-0135](https://doi.org/10.1108/BEPAM-08-2020-0135).
- Brown, A. (2019), "Construction skills shortage in US 'threat to the industry'", available at: <https://www.khl.com/international-construction/construction-skills-shortage-in-us-threat-to-the-industry/139858.article> (accessed 25 January 2020).
- Brown, S.L. and Eisenhardt, K. (1997), "The art of continuous change: linking complexity theory and time-paced evolution in relentlessly shifting organizations", *Administrative Science Quarterly*, Vol. 42 No. 1, pp. 1-34.
- Brunetti, F., Matt, D.T., Bonfanti, A., De Longhi, A., Pedrini, G. and Orzes, G. (2020), "Digital transformation challenges: strategies emerging from a multi-stakeholder approach", *The TQM Journal*, Vol. 32 No. 4, pp. 697-724, doi: [10.1108/TQM-12-2019-0309](https://doi.org/10.1108/TQM-12-2019-0309).
- Büyükoçkan, G. and Göçer, F. (2018), "Digital Supply Chain: literature review and a proposed framework for future research", *Computers in Industry*, Vol. 97, pp. 157-177, doi: [10.1016/j.compind.2018.02.010](https://doi.org/10.1016/j.compind.2018.02.010).
- Burke, W.W. (2013), *Organization Change: Theory and Practice*, Sage Publications, New Delhi.
- Burnes, B. (2011), "Introduction: why does change fail, and what can we do about it?", *Journal of Change Management*, Vol. 11 No. 4, pp. 445-450.
- Butzbach, O., Fuller, D.B. and Schnyder, H. (2020), "Manufacturing discontent: national institutions, multinational firm strategies, and anti-globalization backlash in advanced economies", *Global Strategy Journal*, Vol. 10, pp. 67-93, doi: [10.1002/gsj.1369](https://doi.org/10.1002/gsj.1369).
- Carter, A. and Varney, S. (2018), "Change capability in the agile organisation", *IES Perspectives on HR 2018*.
- Caylar, P.-L., Noterdaeme, O. and Naik, K. (2016), "Digital in industry: from buzzword to value creation", available at: <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/digital-in-industry-from-buzzword-to-value-creation> (accessed 18 July 2020).
- Chan, P.W. and Ejohwomu, O. (2018), "How does project management relate to productivity?", *Association for Project Management*, Princes Risborough, Report, 2018.
- Collins, J.R. and Porras, J.I. (1994), *Built to Last*, Harper Collins, New York, NY.
- Covington, M.V. (2000), "Goal theory, motivation, and school achievement: an integrative review", *Annual Review of Psychology*, Vol. 51, pp. 171-200, doi: [10.1146/annurev.psych.51.1.171](https://doi.org/10.1146/annurev.psych.51.1.171).
- Dallasega, P., Rauch, E. and Linder, C. (2018), "Industry 4.0 as an enabler of proximity for construction supply chains: a systematic literature review", *Computers in Industry*, Vol. 99, pp. 205-225.
- Dawes, S.S. (2009), "Governance in the digital age: a research and action framework for an uncertain future", *Government Information Quarterly*, Vol. 26 No. 2, pp. 257-64.

-
- Deloitte (2015), "Industry 4.0: challenges and solutions for the digital transformation and use exponential technologies", available at: <https://www2.deloitte.com/content/dam/Deloitte/ch/Documents/manufacturing/ch-en-manufacturingindustry-4-0-24102014.pdf> (accessed 25 January 2020).
- Deloitte (2019), "2018 global construction industry overview: insight into strategies, trends, and market size", available at: <https://www2.deloitte.com/us/en/pages/energy-and-resources/articles/global-construction-industry-overview.html> (accessed 25 January 2020).
- Dodge Data and Analytics (2020), "Prefabrication and modular construction 2020", available at: <https://www.construction.com/toolkit/reports/prefabrication-modular-construction-2020> (accessed 21 May 2020).
- Dodge Data and Analytics and Autodesk (2018), "The key performance indicators of construction", available at: <https://damassets.autodesk.net/content/dam/autodesk/drafr/4794/KPIs-of-Construction-Report.pdf> (accessed 25 January 2020).
- Drucker, P.F. (1999), "Knowledge-Worker productivity: the biggest challenge", *California Management Review*, Vol. 41 No. 2, pp. 79-94, doi: [10.2307/41165987](https://doi.org/10.2307/41165987).
- Ejohwomu, O.A. and Hughes, W. (2019), "Incentivization and innovation in construction supply chains", *Construction Industry Development Board Postgraduate Research Conference CIDB 2019, The Construction Industry in the Fourth Industrial Revolution*, pp. 577-587.
- Ejohwomu, O.A., Oshodi, O.S. and Lam, K.C. (2017), "Nigeria's construction industry: barriers to effective communication", *Engineering Construction and Architectural Management*, Vol. 24 No. 4, pp. 652-667, doi: [10.1108/ECAM-01-2016-0003](https://doi.org/10.1108/ECAM-01-2016-0003).
- Ellis, G. (2019), "Top 6 construction risks — and how to reduce them", available at: <https://blog.plangrid.com/2019/09/top-6-construction-risks/> (accessed 25 January 2020).
- Fortune Business Insights (2019), "Market research report", available at: <https://www.fortunebusinessinsights.com/industry-reports/toc/modular-construction-market-101662> (accessed 25 January 2020).
- Gerrish, T., Ruikar, K., Cook, M., Johnson, M. and Phillip, M. (2017), "Using BIM capabilities to improve existing building energy modelling practices", *Engineering Construction and Architectural Management*, Vol. 24 No. 2, pp. 190-208, doi: [10.1108/ECAM-11-2015-0181](https://doi.org/10.1108/ECAM-11-2015-0181).
- Giones, F. and Brem, A. (2017), "From toys to tools: the co-evolution of technological and entrepreneurial developments in the drone industry", *Business Horizons*, Vol. 60 No. 6, pp. 875-884, doi: [10.1016/j.bushor.2017.08.001](https://doi.org/10.1016/j.bushor.2017.08.001).
- Glick, S. and Guggemos, A. (2009), "IPD and BIM: benefits and opportunities for regulatory agencies", *Proceedings of 45th Associated Schools of Construction National Conference*, Gainesville, FL.
- Gondo, M., Patterson, K.D. and Palacios, S.T. (2013), "Mindfulness and the development of a readiness for change", *Journal of Change Management*, Vol. 13 No. 1, pp. 36-51.
- Gross, J.J. (1998), "The emerging field of emotion regulation: an integrative review", *Review of General Psychology*, Vol. 2 No. 3, pp. 271-299, doi: [10.1037/1089-2680.2.3.271](https://doi.org/10.1037/1089-2680.2.3.271).
- Gupta, V.K. (2016), "Strategic framework for managing forces of continuity and change in innovation and risk management in service sector: a study of service industry in India", *International Journal of Services and Operations Management*, Vol. 23 No. 1, pp. 1-17, doi: [10.1504/IJSOM.2016.073285](https://doi.org/10.1504/IJSOM.2016.073285).
- Hasan, A., Baroudi, B., Elmualim, A. and Rameezdeen, R. (2018), "Factors affecting construction productivity: a 30 year systematic review", *Engineering Construction and Architectural Management*, Vol. 25 No. 7, pp. 916-937.
- Hattori, M. and Tanaka, Y. (2016), "Subsidizing new technology adoption in a stackelberg duopoly: cases of substitutes and complements", *Italian Economic Journal*, Vol. 2 No. 2, pp. 197-215, doi: [10.1007/s40797-016-0031-1](https://doi.org/10.1007/s40797-016-0031-1).

- Helfat, C.E. and Winter, S.G. (2011), "Untangling dynamic and operational capabilities: strategy for the (N)ever-Changing world", *Strategic Management Journal*, Vol. 32 No. 11, pp. 1243-1250, doi: [10.1002/smj.955](https://doi.org/10.1002/smj.955).
- Hellriegel, D. and Slocum, J. (2011), *Organizational Behavior, South Western*, 13th ed., Mason, Ohio Thomson, USA, Cengage Learning.
- Hendry, J. (2000), "Strategic decision making, discourse, and strategy as social practice", *Journal of Management Studies*, Vol. 37 No. 7, pp. 955-77, doi: [10.1111/1467-6486.00212](https://doi.org/10.1111/1467-6486.00212).
- Herbert, L. (2017), *Digital Transformation: Build Your Organisation's Future for the Innovation Age*, Bloomsbury Publishing, London.
- Hughes, M. (2011), "Do 70 percent of all organizational change initiatives really fail?", *Journal of Change Management*, Vol. 11 No. 4, pp. 451-464.
- Jarzabkowski, P. (2003), "Strategic practices: an activity theory perspective on continuity and change", *Journal of Management Studies*, Vol. 40 No. 1, pp. 23-55, doi: [10.1111/1467-6486.t01-1-00003](https://doi.org/10.1111/1467-6486.t01-1-00003).
- Jarzabkowski, P. and Whittington, R. (2008), "A strategy as practice approach to strategy research and education", *Journal of Management Enquiry*, Vol. 17 No. 4, pp. 282-286.
- Ji, Y., Qi, K., Qi, Y., Li, Y., Li, H.X., Lei, Z. and Liu, Y. (2020), "BIM-based life-cycle environmental assessment of prefabricated buildings", *Engineering Construction and Architectural Management*, Vol. 27 No. 8, pp. 1703-1725, doi: [10.1108/ECAM-01-2020-0017](https://doi.org/10.1108/ECAM-01-2020-0017).
- Juan, Y.K., Lai, W. and Shih, S.G. (2017), "Building information modeling acceptance and readiness assessment in Taiwanese architectural firms", *Journal of Civil Engineering and Management*, Vol. 23 No. 3, pp. 356-367, doi: [10.1080/17452007.2020.1793721](https://doi.org/10.1080/17452007.2020.1793721).
- Kagermann, H., Helbig, J., Hellinger, A. and Wahlster, W. (2013), "Recommendations for implementing the strategic initiative INDUSTRIE 4.0: securing the future of German manufacturing industry", *Final Report of the Industrie 4.0 Working Group*, Acatech, Munich.
- Kegan, R. and Lahey, L.L. (2001), "The real reason people won't change", *Harvard Business Review*, Vol. 79 No. 10, pp. 85-92.
- Keidanren Policy and Action (2016), "Toward realization of the new economy and society", available at: http://www.keidanren.or.jp/en/policy/2016/029_outline.pdf (accessed 18 July 2020).
- Koeleman, J., Maria João Ribeirinho, M.J., Rockhill, D., Sjödin, E. and Strube, G. (2019), "Decoding digital transformation in construction", [Online] available at: <https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/decoding-digital-transformation-in-construction> (accessed 20 February 2020).
- Kotter, J.P. (2007), "Leading change: why transformation efforts", *Harvard Business Review*, Vol. 86 No. 2, pp. 97-103.
- Lasi, H., Fettke, P., Kemper, H.G., Feld, T. and Hoffmann, M. (2014), "Industry 4.0", *Business Information System Engineering*, Vol. 6 No. 4, pp. 239-242, doi: [10.1007/s12599-014-0334-4](https://doi.org/10.1007/s12599-014-0334-4).
- Leana, C.R. and Barry, B. (2000), "Stability and change as simultaneous experiences in organizational life", *Academy of Management Review*, Vol. 25 No. 4, pp. 753-759.
- Lee, I. and Lee, K. (2015), "The Internet of things (IoT): applications, investments, and challenges for enterprises", *Business Horizons*, Vol. 58 No. 4, pp. 431-440.
- Lee, J.W., Park, S.H., Oh, J.G. and Yeo, G.T. (2018), "Efficiency analysis of construction project freight forwarding companies using DEA analysis", *Digital Convergence Research*, Korean Digital Policy Association, Vol. 16 No. 6, pp. 75-84, doi: [10.14400/JDC.2018.16.6.075](https://doi.org/10.14400/JDC.2018.16.6.075).
- Liu, Y. and Xu, X. (2017), "Industry 4.0 and cloud manufacturing: a comparative analysis", *Journal of Manufacturing Science and Engineering*, Vol. 139 No. 3, 034701, doi: [10.1115/1.4034667](https://doi.org/10.1115/1.4034667).
- Mahardika, H., Thomas, D., Ewing, M.T. and Japutra, A. (2019), "Experience and facilitating conditions as impediments to consumers' new technology adoption", *International Review of Retail Distribution & Consumer Research*, Vol. 29 No. 1, pp. 79-98, doi: [10.1080/09593969.2018.1556181](https://doi.org/10.1080/09593969.2018.1556181).

-
- Mahidhar, V. and Davenport, T.H. (2018), "Why companies that wait to adopt AI may never catch up", *Harvard Business Review*, available at: <https://hbr.org/2018/12/why-companies-that-wait-to-adopt-ai-may-never-catch-up> (accessed 20 February 2020).
- Mäkinen, T. (2017), "Strategizing for digital transformation: a case study of digital transformation process in the construction industry", *Thesis: Master's Programme in Industrial Engineering and Management*, available at: https://aaltodoc.aalto.fi/bitstream/handle/123456789/29030/master_M%c3%a4kinen_Tim_o_2017.pdf?sequence=2&isAllowed=y (accessed 29 October 2020).
- Malhotra, N. and Hinings, B. (2012), "Unpacking continuity and change as a process of radical transformation", *Paper Presented at the 2012 Academy of Management Boston*.
- Manda, M.I. and Dhaou, S.B. (2019), "Responding to the challenges and opportunities in the 4th industrial revolution in developing countries", *ICEGOV2019, Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance*, April, pp. 244-253, doi: [10.1145/3326365.3326398](https://doi.org/10.1145/3326365.3326398).
- Marabelli, M. and Newell, S. (2014), "Knowing, power and materiality: a critical review and reconceptualization of absorptive capacity", *International Journal of Management Reviews*, Vol. 16 No. 4, pp. 479-499, doi: [10.1111/ijmr.12031](https://doi.org/10.1111/ijmr.12031).
- Matthews, K. (2018), "How the cloud has changed the construction industry", available at: <https://cloudtweaks.com/2018/07/cloud-construction-industry/> (accessed 25 January 2020).
- Mazumdar, T., Raj, S.P. and Sinha, I. (2005), "Reference price research review and propositions", *Journal of Marketing*, Vol. 69 No. 4, pp. 84-102.
- McLaughlin, P., Bessant, J. and Smart, P. (2008), "Developing an organisation culture to facilitate radical innovation", *International Journal of Technology Management*, Vol. 44 Nos 3/4, doi: [10.1504/IJTM.2008.021041](https://doi.org/10.1504/IJTM.2008.021041).
- Mintzberg, H. (1988), "Crafting strategy", *The McKinsey Quarterly*, Summer, pp. 71-90.
- Mintzberg, H., Ahlstrand, B. and Lampel, J. (1998), *Strategy Safari: The Complete Guide through the Wilds of Strategic Management*, Pearson Education, Singapore.
- Momaya, K. (2001), *International Competitiveness: Evaluation and Enhancement*, Hindustan Publishing Corporation, New Delhi.
- Momaya, K. (2011), "Cooperation for competitiveness of emerging countries: learning from a case of nanotechnology", *Competitiveness Review: An International Business Journal*, Vol. 21 No. 2, pp. 152-170, doi: [10.1108/10595421111117443](https://doi.org/10.1108/10595421111117443).
- Momaya, K.S. (2014), "Break-out for Competitiveness of Indian firms: context, need and opportunities", *International Journal of Global Business and Competitiveness*, Vol. 9 No. 1, pp. 3-7.
- Momaya, K.S. (2019), "The past and the future of competitiveness research: a review in an emerging context of innovation and EMNEs", *International Journal of Global Business and Competitiveness*, Vol. 14 No. 1, pp. 1-10, doi: [10.1007/s42943-019-00002-3](https://doi.org/10.1007/s42943-019-00002-3).
- Momaya, K. and Ambastha, A. (2005), "Technology management and competitiveness: is there any relationship?", *International Journal of Technology Transfer and Commercialisation*, Vol. 4 No. 4, pp. 518-524.
- Moon, M.J. (2002), "The evolution of e-government among municipalities: rhetoric or reality?", *Public Administration Review*, Vol. 62 No. 4, pp. 424-33.
- Moon, S., Ham, N., Kim, S., Hou, L., Kim, J.H. and Kim, J.-J. (2020), "Fourth industrialization- oriented offsite construction: case study of an application to an irregular commercial building", *Engineering Construction and Architectural Management*, Vol. 27 No. 9, pp. 2271-2286, doi: [10.1108/ECAM-07-2018-0312](https://doi.org/10.1108/ECAM-07-2018-0312).
- Mugge, P., Abbu, H., Michaelis, T.L., Kwiatkowski, A. and Gudergan, G. (2020), "Patterns of digitization, research-technology", *Management*, Vol. 63 No. 2, pp. 27-35, doi: [10.1080/08956308.2020.1707003](https://doi.org/10.1080/08956308.2020.1707003).

- Mzyece, D., Ndekugri, I.E. and Ankrah, N.A. (2019), "Building information modelling (BIM) and the CDM regulations interoperability framework", *Engineering Construction and Architectural Management*, Vol. 26 No. 11, pp. 2682-2704, doi: [10.1108/ECAM-10-2018-0429](https://doi.org/10.1108/ECAM-10-2018-0429).
- Nasim, S. and Sushil (2011), "Revisiting organizational change: exploring the paradox of managing continuity and change", *Journal of Change Management*, Vol. 11 No. 2, pp. 185-206.
- Oakland, J.S. and Tanner, S.J. (2007), "A new framework for managing change", *The TQM Magazine*, Vol. 19 No. 6, pp. 572-89.
- Oesterreich, T.D. and Teuteberg, F. (2016), "Understanding the implications of digitisation and automation in the context of Industry 4.0: a triangulation approach and elements of a research agenda for the construction industry", *Computers in Industry*, Vol. 83, pp. 121-139.
- Okedara, K., Ejohwomu, O. and Chan, P. (2020), "Ethics and stakeholder engagement for industry/construction 4.0: a systematic review", in Aigbavboa, C. and Thwala, W. (Eds), *The Construction Industry in the Fourth Industrial Revolution, CIDB 2019*, Springer, Cham.
- Parente, R.C., Geleilate, J.M.G. and Rong, K. (2018), "The sharing economy globalization phenomenon: a research agenda", *Journal of International Management*, Vol. 24 No. 1, pp. 52-64, doi: [10.1016/j.intman.2017.10.001](https://doi.org/10.1016/j.intman.2017.10.001).
- Pettigrew, A.M. (2000), "Linking change processes to outcomes", in Beer, M. and Nohria, N. (Eds), *Breaking the Code of Change*, HBS Press, Boston, MA, pp. 243-265.
- Preece, D.A. (1988), "Managing the adoption of new technology", *Management Research News*, Vol. 11 Nos 1/2, pp. 50-52, doi: [10.1108/eb027964](https://doi.org/10.1108/eb027964).
- Qian, X. and Papadonikolaki, E. (2020), "Shifting trust in construction supply chains through blockchain technology", *Engineering Construction and Architectural Management*, Vol. 28 No. 2, pp. 584-602, doi: [10.1108/ECAM-12-2019-0676](https://doi.org/10.1108/ECAM-12-2019-0676).
- Rane, S.B. and Narvel, Y.A.M. (2019), "Re-designing the business organization using disruptive innovations based on blockchain-IoT integrated architecture for improving agility in future Industry 4.0", *Benchmarking: An International Journal*, Vol. ahead-of-print No. ahead-of-print, doi: [10.1108/BIJ-12-2018-0445](https://doi.org/10.1108/BIJ-12-2018-0445).
- Rane, S.B., Potdar, P.R. and Rane, S. (2019), "Development of project risk management framework based on industry 4.0 technologies", *Benchmarking: An International Journal*, Vol. ahead-of-print No. ahead-of-print, doi: [10.1108/BIJ-03-2019-0123](https://doi.org/10.1108/BIJ-03-2019-0123).
- Riley, T.B. (2007), *The 'e' in Government Projects: Basic Issues*, The Riley Report, Riley Information Services, Ottawa.
- Rodrik, D. (2018), "Populism and the economics of globalization", *Journal of International Business Policy*, Vol. 1, pp. 12-33, doi: [10.1057/s42214-018-0001-4](https://doi.org/10.1057/s42214-018-0001-4).
- Rosenbaum, D., More, E. and Steane, P. (2018), "Planned organisational change management: forward to the past? An exploratory literature review", *Journal of Organizational Change Management*, Vol. 31 No. 2, pp. 286-303, doi: [10.1108/JOCM-06-2015-0089](https://doi.org/10.1108/JOCM-06-2015-0089).
- Senarathna, I., Warren, M., Yeoh, W. and Salzman, S. (2014), "The influence of organisation culture on E-commerce adoption", *Industrial Management and Data Systems*, Vol. 114 No. 7, pp. 1007-1021, doi: [10.1108/IMDS-03-2014-0076](https://doi.org/10.1108/IMDS-03-2014-0076).
- Sethi, R. (2003), "Managing change: transforming organizations and culture", in Gupta, M.P. (Ed.), *Towards E-Government: Management Challenges*, GIFT, New Delhi.
- Shaughnessy, H. (2018), "Creating digital transformation: strategies and steps", *Strategy and Leadership*, Vol. 46 No. 2, pp. 19-25, doi: [10.1108/SL-12-2017-0126](https://doi.org/10.1108/SL-12-2017-0126).
- Stojanovic, L., Stojanovic, N. and Apostolou, D. (2006), "Change management in e- government: OntoGov case study", *Electronic Government: An International Journal*, Vol. 3 No. 1, pp. 74-92.
- Sturdy, A. and Grey, C. (2003), "Beneath and beyond organizational change management: exploring alternatives", *Organization*, Vol. 10 No. 4, pp. 651-662.

-
- Sull, D. and Eisenhardt, K.M. (2012), "Simple rules for a complex world", *Harvard Business Review*, Vol. 90 No. 9, pp. 68-74.
- Sushil (2005), "A flexible strategy framework for managing continuity and change", *International Journal of Global Business and Competitiveness*, Vol. 1 No. 1, pp. 22-32.
- Sushil (2012), "Flowing stream strategy: managing confluence of continuity and change", *Journal of Enterprise Transformation*, Vol. 2 No. 1, pp. 26-49, doi: [10.1080/19488289.2011.650280](https://doi.org/10.1080/19488289.2011.650280).
- Sutherland, F. and Smith, A.C. (2011), "Duality theory and the management of the change– stability paradox", *Journal of Management and Organization*, Vol. 17 No. 4, pp. 534-547.
- Synder, H. (2019), "Literature review as a research methodology: an overview and guidelines", *Journal of Business Research*, Vol. 104, pp. 333-339, doi: [10.1016/j.jbusres.2019.07.039](https://doi.org/10.1016/j.jbusres.2019.07.039).
- Snyder, J., Menard, A. and Spare, N. (2018), "Big data = big questions for the engineering and construction industry", available at: https://www.fminet.com/wp-content/uploads/2018/11/FMI_BigDataReport.pdf (accessed 25 January 2020).
- Tao, F., Qi, Q., Wang, L. and Nee, A.Y.C. (2019), "Digital twins and cyber–physical systems toward smart manufacturing and industry 4.0: correlation and comparison", *Engineering*, Vol. 5 No. 4, pp. 653-661, doi: [10.1016/j.eng.2019.01.014](https://doi.org/10.1016/j.eng.2019.01.014).
- Teece, D.J. (2017), "Business models and dynamic capabilities", *Long Range Planning*, Vol. 51 No. 1, pp. 40-49.
- Teece, D.J. (2018), "Dynamic capabilities as (workable) management systems theory", *Journal of Management and Organization*, Vol. 24 No. 3, pp. 1-10.
- Torraco, R.J. (2005), "Work design theory: a review and critique with implications for human resource development", *Human Resource Development Quarterly*, Vol. 16 No. 1, pp. 85-109.
- Tsoukas, H. and Chia, R. (2002), "On organizational becoming: rethinking organizational change", *Organization Science*, Vol. 13 No. 5, pp. 567-582.
- Valente, F. (2019), "Global spending in construction to reach \$17.5 trillion by 2030", available at: <https://ww2.frost.com/news/press-releases/global-spending-in-construction-to-reach-17-5-trillion-by-2030-finds-frost-sullivan/#:~:text=By%202030%2C%20global%20spending%20in,57%25%20of%20all%20global%20growth> (accessed 25 January 2020).
- Varsos, D.S. and Assimakopoulos, N.A. (2016), "A systems approach to alternative paradigms for organisation and organisational change", *International Journal of Applied Systemic Studies*, Vol. 6 No. 4, pp. 302-326.
- Verbeke, A., Coeurderoy, R. and Matt, T. (2018), "The future of international business research on corporate globalization that never was...", *Journal of International Business Studies*, Vol. 49 No. 9, pp. 1101-1112, doi: [10.1057/s41267-018-0192-2](https://doi.org/10.1057/s41267-018-0192-2).
- Verhoef, P.C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Dong, J.Q., Fabian, N. and Haelelin, M. (2021), "Digital transformation: a multidisciplinary reflection and research agenda", *Journal of Business Research*, Vol. 122 January 2021, pp. 889-901, doi: [10.1016/j.jbusres.2019.09.022](https://doi.org/10.1016/j.jbusres.2019.09.022).
- Washington, M. and Hacker, M. (2005), "Why change fails: knowledge counts", *Leadership and Organization Development Journal*, Vol. 6 No. 5, pp. 400-411.
- Webster, J. and Watson, R.T. (2002), "Analyzing the past to prepare for the future: writing a literature review", *MIS Quarterly*, Vol. 26 No. 2, pp. 13-23.
- Whittington, R. (1996), "Strategy as practice", *Long Range Planning*, Vol. 29 No. 5, pp. 731-35.
- Whittington, R. (2002), "Theories of strategy", in Mazzucato, M. (Ed.), *Strategy for Business*, Sage Publications, London, pp. 32-58.
- Wirtz, B.W., Weyerer, J.C. and Geyer, C. (2018), "Artificial intelligence and the public sector - applications and challenges", *International Journal of Public Administration*, Vol. 42 No. 7, pp. 596-615, doi: [10.1080/01900692.2018.1498103](https://doi.org/10.1080/01900692.2018.1498103).

- Woodhead, R., Stephenson, P. and Morrey, D. (2018), "Digital construction: from point solutions to IoT ecosystem", *Automation in Construction*, Vol. 93, pp. 35-46, doi: [10.1016/j.autcon.2018.05.004](https://doi.org/10.1016/j.autcon.2018.05.004).
- Yoon, J. and Chae, M. (2009), "Varying criticality of key critical success factors national e- strategy along the status of economic development of nations", *Government Information Quarterly*, Vol. 26, pp. 25-34.
- Young, R. (2019), "Picking the right technology for your E&C firm", available at: <https://www.fminet.com/fmi-quarterly/article/2019/06/picking-the-right-technology-for-your-ec-firm/> (accessed 25 January 2020).
- Zollo, M. and Winter, S.G. (2002), "Deliberate learning and the evolution of dynamic capabilities", *Organisation Science*, Vol. 13 No. 3, pp. 339-351.

About the authors

Dr. Sanjay Bhattacharya is instructor, practitioner, researcher and trainer specializing in strategy and competitiveness. He gained his PhD degree from the Department of Management Studies at Indian Institute of Technology (IIT) Delhi. He is actively involved in teaching postgraduate business management subjects in the built environment, at Delhi Technological University, RICS School of Built Environment and Jaipuria Institute of Management. He holds a master's degree in construction management as well as bachelor's degree in architecture both from reputed IITs. Further, he has a rich industry experience of over two decades, involving building projects design and execution in India and Middle East countries. Sanjay Bhattacharya is the corresponding author and can be contacted at: sanjbhat66@yahoo.co.in

Dr. K.S. Momaya is professor of competitiveness at Shailesh J. Mehta School of Management, Indian Institute of Technology Bombay, Mumbai, India. His research interests include role of business excellence, management of technology and innovation (MoT), and collaborations for competitiveness. He worked with the Department of Management Studies, IIT Delhi as a core faculty for more than a decade and made several unique contributions. He has done some challenging projects in India, Canada and Japan including one at the Institute of Innovation Research, Hitotsubashi University, Tokyo, and has also worked with Shimizu Corporation.



Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Advancements in steam distillation system for oil extraction from peppermint leaves

Ravi Kant^a, Anil Kumar^{a,b,*}

^a Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India

^b Centre for Energy and Environment, Delhi Technological University, Delhi 110042, India

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Peppermint
Distillation
Solar
Steam
Herb

ABSTRACT

The numerous advancements conducted experimentally to improve the efficiency of steam desalination systems to extract the essential oil from the peppermint plant are presented in this article. Peppermint oil is important for human life because of its health benefits. It is used in the medical sector and food industry as an herb and fragrance, respectively. In this study, two types of distillation systems, namely the solar distillation system and the electrical energy-based distillation system, are discussed. The major factors that lead to increased system efficiency, such as the mass flow rate of heat transfer fluid, inlet water flow rate to boiler and batch size of peppermint, are addressed.

© 2021 Elsevier Ltd. All rights reserved.

Selection and peer-review under responsibility of the scientific committee of the Technology Innovation in Mechanical Engineering-2021.

1. Introduction

Approximately 80% of the population relies on herbal medicines to meet their health needs. As a result, medicinal and aromatic plants are important because the products derived from plants are used to make herbal medicines. According to WHO reports, about 21,000 species are used as medicinal plants. Distillation process is used to extract the essential oil (EO) from these aromatic plants [1]. The EO obtained from these plants is utilized as a fragrance in the cosmetic industry and for flavouring in the food industry. The medical industry also used the peppermint oil for its functional use [1,2].

The Attention of food and medical industries has attracted the peppermint (*Mentha peperita* L.), a medicinal herb, because of its health advantages. The use of Peppermint oil as health benefits are shown in Fig. 1 [3].

Peppermint oil is extracted from peppermint plant leaves through the steam distillation process. The boiler, distillery, and condenser are the main components of the distillation system. Steam distillation is a process in which water is heated up to boiling point by a heat source and converted into steam in the boiler. Distillery consists of peppermint leaves and the steam generated in the boiler is passed through the leaves. The oil is evaporated from the leaves at a temperature range of 250–300 °C. The evaporated

steam then passed to the condenser through the connection pipe and condensed. The oil and water then separated in a Florentine flask. The steam distillation system is shown in Fig. 2 [3].

This paper discusses technical advancements in steam distillation systems for peppermint oil extraction. The paper aims to provide complete information for designing an energy-efficient peppermint oil extraction system, which will help researchers, developers, and the people associated with the field, leading to further futuristic development in the concerned area.

2. Technical advancement in steam distillation systems for peppermint oil extraction

Radwan et al. (2020) designed the conventional and solar energy based steam distillation system to extract the peppermint oil from leaves. The setups for conventional and solar energy based distillation are shown in Fig. 3 and Fig. 4, respectively. An electrical heat source was used for evaporation of water in a conventional distillation system. The solar distillery consisted of a parabolic solar disc as a heat source. The study concluded that the productivity, essential oil yield, and extraction efficiency of the system are affected by variation in the boiler inlet mass flow rate and batch size of peppermint. The productivity and essential oil yield increase with an increase in batch size and boiler inlet flow rate, and however, the requirement of energy for distillation reduce [1,2].

* Corresponding author.

E-mail address: anilkumar76@dtu.ac.in (A. Kumar).

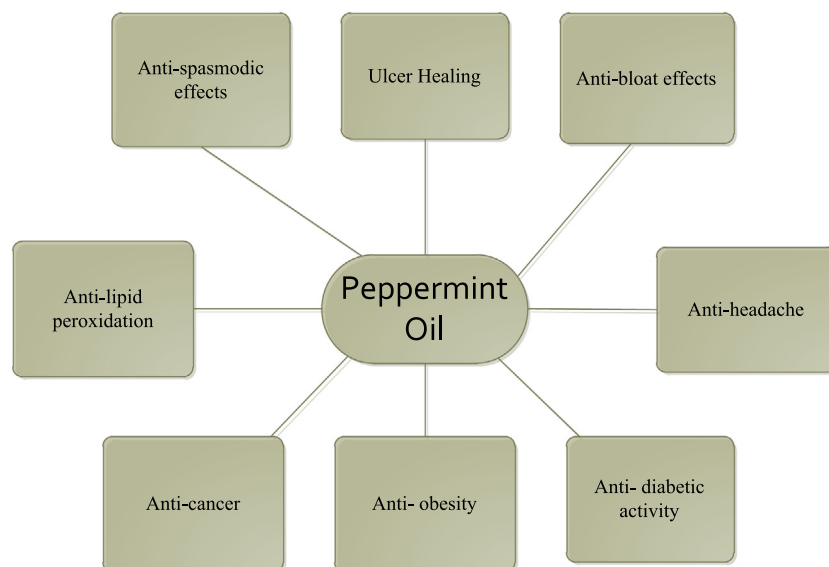


Fig. 1. Health benefits of Peppermint oil.

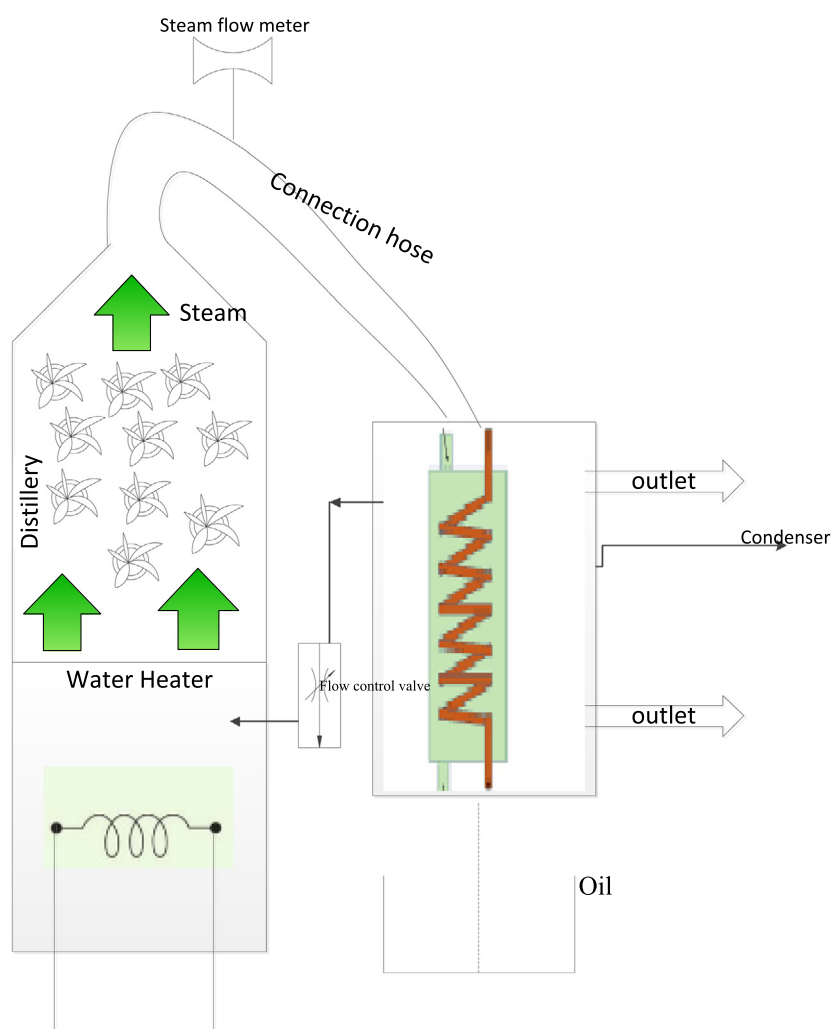


Fig.2. Steam distillation system for oil extraction from peppermint leaves.

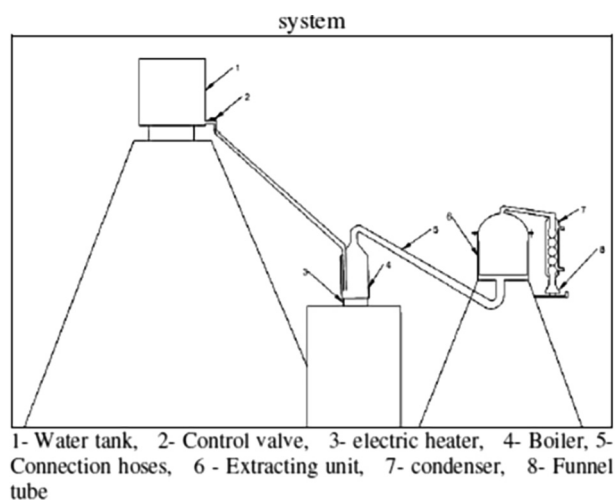


Fig.3. Conventional steam distillation [2].



Fig.4. Solar distillation system [1].

Gavahian and Chu (2018) designed and developed an ohmic accelerated steam distillation system (OASD) to extract lavender oil. The OASD system is illustrated in Fig. 5. Ohmic electrodes were used for heating the water. In this study, the yield and energy consumed for distillation of the proposed system have been calculated and compared with the conventional distillation system (SD). The results indicated that the productivity of the OASD system was 3.3–3.8% more than the SD system; however, the energy consumption to extract 1 ml essential oil was 55.55% lower than SD system [4].

Chen and Spiro (1994) studied a microwave heating-based system to extract the essential oil from peppermint plants. This study aims to predict the factors that evaluate the microwave heating rate of plants and solvent mixture. The power output affects the rate of extraction of essential oil from the peppermint plant. The extraction rate of essential oil increased with an increase in the power output. The oil extraction system is illustrated as in Fig. 6 [5].

Gavahian and Farahnaky (2017) presented a review on ohmic assisted hydro distillation (OAHD) systems for Peppermint oil extraction. This study discusses the concepts of OAHD and its recent implementations. OAHD is an innovative application of ohmic heating that takes advantage of volumetric heating to fix the drawbacks of conventional distillation system. The processing time, energy consumption, and operational cost of the proposed system were lower compared to SD systems. The OAHD system is illustrated in Fig. 7 [6].

Kulturel and Tarhan (2016) developed a solar energy based distillation system to extract the essential oil. The seven compound parabolic collectors (CPCs) attached in series were used as heat source to heat the oil. Heat transfer oil was circulated through the circulation pump to heat the water in the boiler. The experimental setup is shown in Fig. 8. An experimental study was performed to determine the performance of the system for various ambient and operating conditions [7].

Munir et al. (2014) designed and fabricated solar distillation system to extract essential oil from the peppermint plant. Thermal energy requirements for the distillation process and system effi-

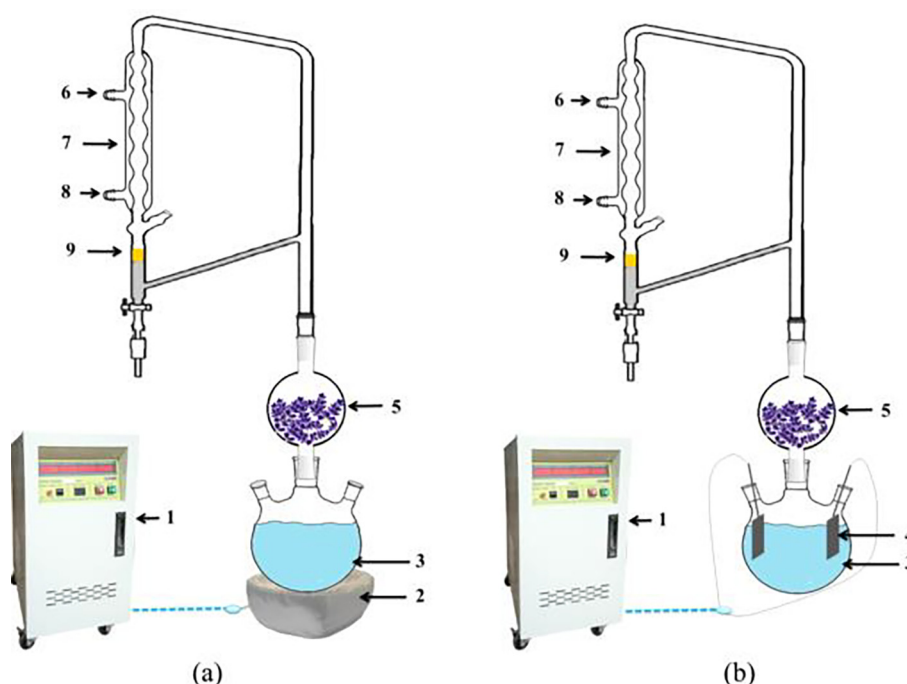


Fig.5. (a) SD; (b) OASD system [4].

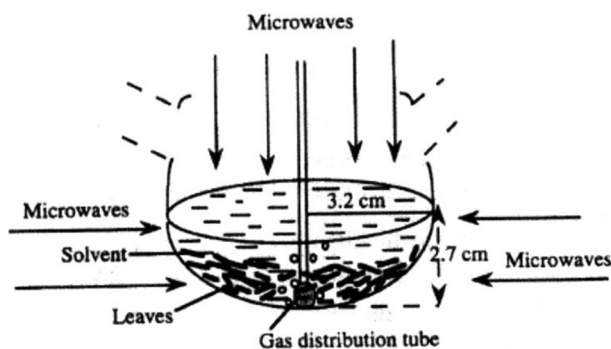


Fig.6. Oil extraction system [5].

ciency were evaluated in this study. The proposed system is shown in Fig. 9. Solar disc, distillery and condenser are the main components of the system. Tracking system was also coupled to receive maximum solar radiation. A secondary receiver was used to focus the radiation received from disc to boiler. The efficiency of the system was achieved to 33.21% and 1.548 kW thermal powers available for the distillation process. An average 3.5 kWh energy was consumed for the processing of 10 kg batch [8].

Afjal et al. (2017) fabricated a hybrid solar distillation system for the essential oil extraction from the peppermint and evaluated the essential oil yield and thermal energy requirement for distillation (Fig. 10). The distillation system consisted of a solar disc, distillery, and coil condenser. A supplementary biomass arrangement was also attached with the distillation system to supplement the

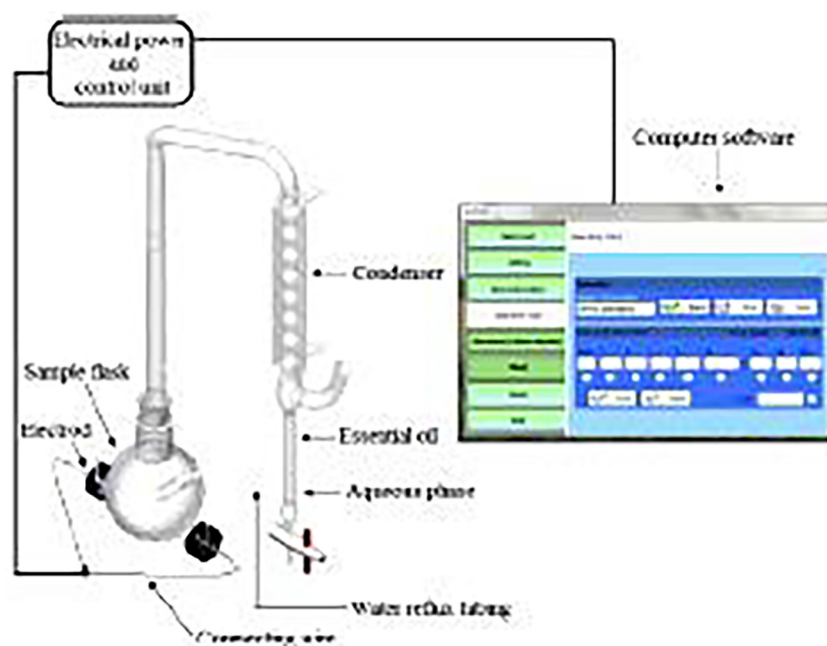


Fig.7. OAH system to extract the oil [6].

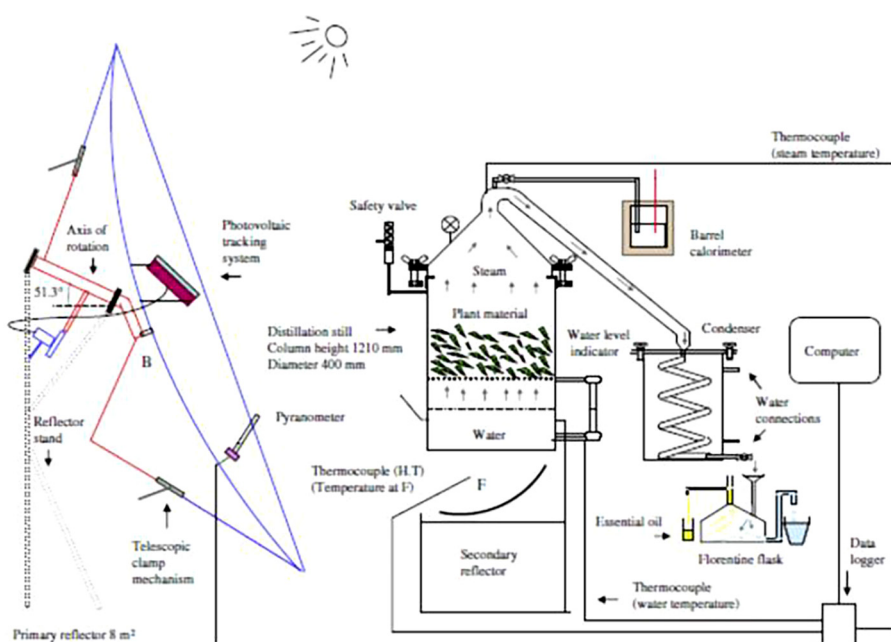


Fig.8. Solar distillation system to extract the oil [7].



Fig.9. Solar distillation system for peppermint oil extraction [8].

system for the period of severe weather conditions or seasonal climatic conditions in a typical year [9].

3. Conclusions

The present study concentrated on numerous technical advancements conducted to improve the performance of steam distillation system for oil extraction from peppermint leaves. Following conclusions were reported based on the present study (Table 1):

- The productivity of the system is affected by the change in heat transfer oil mass flow rate, inlet water mass flow rate, and batch size of peppermint.
- The yield and essential oil yield of the system increased with an increase in batch size and inlet water flow rate.

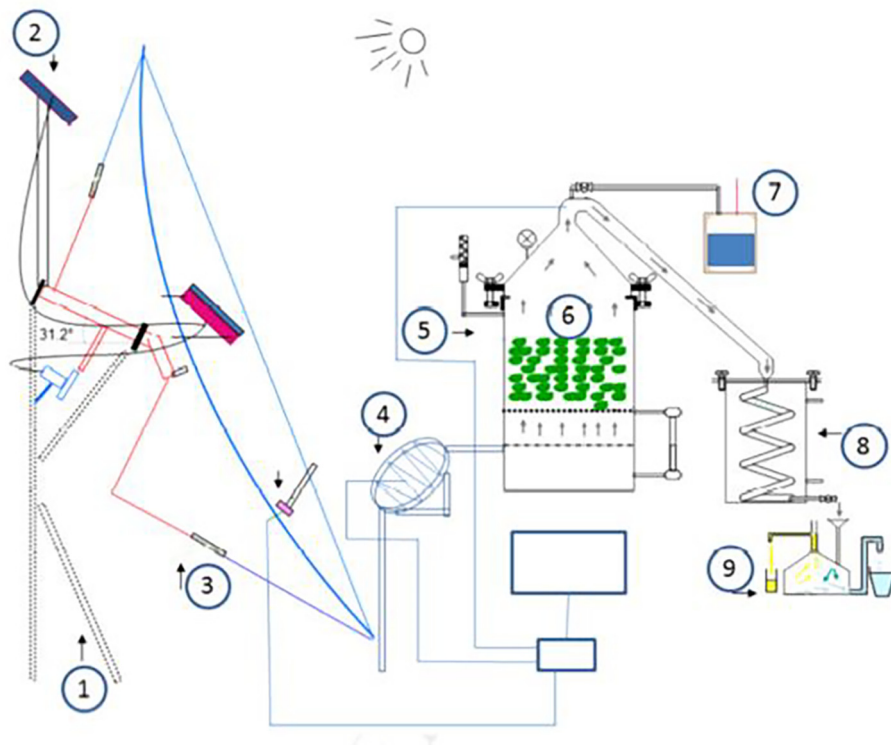


Fig.10. Schematic solar distillery [9].

Table 1

. Conclusion of numerous steam distillation systems for peppermint oil.

S. No.	Authors	Classification	Heat Source	Purpose	Conclusions
1	Radwan et al. (2020) [1]	Solar energy	Solar disc	Water heating	<ul style="list-style-type: none"> Performance of the system was evaluated by calculating the yield, essential oil yield and extraction efficiency. The performance parameters were affected by the inlet boiler flow rate and batch size of peppermint.
2	Radwan et al. (2020) [2]	Electrical energy	Electrical heater	Water heating	<ul style="list-style-type: none"> The study concluded that the productivity, essential oil yield, and extraction efficiency of the system are affected by variation in the boiler inlet mass flow rate and batch size of peppermint. The productivity and essential oil yield increased with an increase in batch size and boiler inlet flow rate and however, a requirement of energy for distillation reduces.
3	Gavahian and Chu (2018) [4]	Electrical energy	Electrical electrodes	Water heating	<ul style="list-style-type: none"> In the study, the yield and energy consumed for distillation of the proposed system have been calculated and compared with the biomass conventional distillation system (SD). The results indicated that the productivity of the OASD system is 3.3–3.8% more than the SD system. However, the energy consumption to extract 1 ml of EO is 55.55% lower than SD system.
4	Chen and Spiro (1994) [5]	Electrical energy	Microwave heating	Water heating	<ul style="list-style-type: none"> The result concluded that power output affects the rate of extraction of essential oil from peppermint plants. The extraction rate of essential oil increased with an increase in the power output.
5	Gavahian and Farahnaky (2017) [6]	Electrical energy	Electrical heating	Water heating	<ul style="list-style-type: none"> OAHD is an innovative application of ohmic heating that takes advantage of volumetric heating to fix the drawbacks of a conventional distillation system. The processing time, energy consumption and operational cost of the proposed system were lower compared to SD systems.
6	Kulturel and Tarhan (2016) [7]	Solar energy	Parabolic trough collector	Heat transfer oil heating	<ul style="list-style-type: none"> An experimental study was performed to determine the performance of the system for various ambient and operating conditions. The maximum solar utilization efficiency was evaluated as 80%.
7	Munir et al. (2014) [8]	Solar energy	Solar disc	Water heating	<ul style="list-style-type: none"> Thermal energy requirements for the distillation process and system efficiency were evaluated in this study. The efficiency of the system was achieved to 33.21% and 1.548 kW thermal powers available for the distillation process. An average 3.5 kWh energy was consumed for the processing of 10 kg batch size of peppermint.
8	Afjal et al. (2017) [9]	Hybrid (solar & conventional)	Solar disc and Biomass	Water heating	<ul style="list-style-type: none"> Thermal energy for the distillation process and essential oil yield were evaluated in this study. The results indicated that essential oils from peppermint and eucalyptus leaves were extracted as 0.40% and 0.59% w/w, respectively.

- The energy requirement for the distillation process was lower in the solar and electrical energy-based distillation system compared to the conventional distillation system.

CRediT authorship contribution statement

Ravi Kant: Methodology, Visualization, Investigation, Writing - original draft. **Anil Kumar:** Conceptualization, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Delhi Technological University, Delhi, for awarding the fellowship for pursuing the Ph.D. in the Mechanical Engineering Department and gratefully acknowledging Centre for Energy and Environment to support the present work.

References

- [1] M.N. Radwan, M.M. Morad, M.M. Ali, K.I. Wasfy, A solar steam distillation system for extracting lavender volatile oil, *Energy Rep.* 6 (2020) 3080–3087, <https://doi.org/10.1016/j.egy.2020.11.034>.
- [2] M.N. Radwan, M.M. Morad, M.M. Ali, K.I. Wasfy, Extraction of Peppermint volatile using a sample steam distillation system, *plant Arch.* 20 (2020) 1487–1491.
- [3] M. Loolae, N. Moasefi, H. Rasouli, H. Adibi, Peppermint and Its Functionality : A Review, *Arch. Clin. Microbiol.* 8 (2017) 1–16, <https://doi.org/10.4172/1989-8436.100053>.
- [4] M. Gavahian, Y. Chu, Ohmic accelerated steam distillation of essential oil from lavender in comparison with conventional steam distillation, *Innov. Food Sci. Emerg. Technol.* 50 (September) (2018) 34–41, <https://doi.org/10.1016/j.ifset.2018.10.006>.
- [5] S.S. Chen, M. Spiro, Study of Microwave Extraction of Essential Oil Constituents from Plant Materials, *J. Microw. Power Electromagn. Energy.* 29 (4) (1994) 231–241, <https://doi.org/10.1080/08327823.1994.11688251>.
- [6] M. Gavahian, A. Farahnaky, Ohmic-assisted hydrodistillation technology: A review, *Trends Food Sci. Technol.* 72 (2018) 153–161, <https://doi.org/10.1016/j.tifs.2017.12.014>.
- [7] Y. Kulturel, S. Tarhan, Performance of a solar distillery of essential oils with compound parabolic solar collectors, *J. Sci. Ind. Res.* 75 (2016) 691–696.
- [8] A. Munir, O. Hensel, W. Scheffler, H. Hoedt, W. Amjad, A. Ghafoor, Design, development and experimental results of a solar distillery for the essential oils extraction from medicinal and aromatic plants, *Sol. Energy* 108 (2014) 548–559, <https://doi.org/10.1016/j.solener.2014.07.028>.
- [9] A. Afzal, A. Munir, A. Ghafoor, J.L. Alvarado, Development of hybrid solar distillation system for essential oil extraction, *Renew. Energy* (2017), <https://doi.org/10.1016/j.renene.2017.05.027>.

An Adaptive Master-Slave Technique using Converter Current Modulation in VSC-based MTDC System

Radheshyam Saha

Department of Electrical Engineering,
Delhi Technological University,
New Delhi, India
rshahacno@yahoo.com

Madhusudan Singh

Department of Electrical Engineering,
Delhi Technological University,
New Delhi, India

Ashima Taneja

Department of Electrical Engineering,
Delhi Technological University,
New Delhi, India
tanejaashima.1988@gmail.com

Abstract—The VSC-based multi terminal HVDC technology is gaining popularity in power system to integrate and efficiently operate multi-area AC systems of same or asynchronous frequency as one grid. Nevertheless, this technology has enabled to enhance controllability and stability of AC-DC integrated system. Depending upon the amount of power transfer between AC and DC systems, control strategy adopted for MTDC system varies. Master-Slave, Voltage Margin and Voltage Droop control are classical DC voltage control strategies for VSC-based MTDC systems. Although Voltage Droop technique is reliable but it suffers from drawbacks like inability to deliver constant active power to critical AC systems, inability to maintain DC voltage to a constant value, risk of irreversible switching from droop control mode to constant power control mode, etc. Voltage Margin technique is reliable than Master Slave control but suffers from power oscillations and large DC voltage overshoot when DC voltage control is transferred from main master controller to reserve controller. In this paper, an adaptive Master-Slave control technique with modulation of converter currents in Master terminal is proposed where direct-axis converter current is allowed to increase well up to maximum IGBT current carrying ability limit. Thus, augmented Master control scheme is able to compensate for more unbalanced power in DC grid than the conventional one thereby preventing severe DC overvoltage or arresting DC voltage from running down. The proposed control has enabled to improve the reliability and stability of AC-DC grid and is found to be more flexible than conventional Master-Slave control.

Keywords—Vector control, Voltage source converters, Multiterminal DC systems, Master-slave control.

I. INTRODUCTION

The extensive use of HVDC transmission technologies is well justified because of many benefits offered by the DC transmission over AC transmission like unconstrained bulk power transmission over long distance, reduced power losses, flexible power flow control, effective conductor utilization, reduced RoWs, enhanced AC-DC power system stability, evacuation of power from remotely located RESs, islanding operation and many more [1,2]. Also, with advent of HVDC technology, it has been possible to not only interconnect two asynchronous systems stably but also an AC area can be divided into a multi-area system. The Voltage Source Converter (VSC) based HVDC technologies and topologies have ushered a new dimension to enhance the stability and security of the power system of large AC-DC network integration. Besides the various benefits indicated, few other attributes of these technologies are rapid power

reversal, feasible multi-terminal systems, improved power quality, decoupled power flow control and dynamic reactive power compensation, etc. [3].

As regards the anatomy of the various control techniques that are employed in HVDC system, the vector current control scheme is used for control of VSC-based HVDC transmission in which generation of direct-axis reference current is done either by DC voltage control or by constant active power control. Similarly, generation of quadrature-axis reference current is done either by AC voltage control or by constant reactive power control. In a point to point VSC-HVDC system, one converter exercises DC voltage control while other is transferring constant active power [2].

Multi-terminal (MTDC) systems are formed by interconnecting DC grids of various converter terminals to a common DC voltage [1]. It is necessary to have proper DC voltage control and active power control [4]. At least one converter in MTDC grid is responsible for regulation of DC voltage so that balanced power exchange takes place in between AC and DC system [5]. Basically, MTDC control system is based upon how the generation of direct-axis reference current is done. Master-Slave (MS) control, Voltage Margin (VM) control and Voltage Droop (VD) control are three basic control techniques for VSC-MTDC systems whose features are depicted in Table I [1-3].

MS control is not so reliable due to its dependency on single converter for critical task of DC voltage regulation. Also, in case of increase in unbalanced power in DC grid, Master may lose its capability of maintaining stable DC voltage. As a result, overvoltage or undervoltage can occur due to lack of further DC voltage regulation [1,3]. For larger Voltage Margin controlled MTDC systems, more reserve masters are required, thus, making it cumbersome to define reference DC voltages for each converter station [1, 3-5]. In certain critical AC systems and in passive AC networks, it is desired to have a constant active power output from VSCs [3]. But voltage source converter stations controlled by VD technique are unable to deliver constant active power, if required [4]. With power losses & deviation in DC voltage neglected, this control cannot provide desired power flow without deviation from allotted power references [1]. In fact, such system is unable to regulate DC voltage to a fixed value rather only keeps the DC voltage within a permissible range [4]. Also, while transferring unbalanced power in grid, it does not take care of power margin left with the converter. Therefore, converter control may switch from droop control to constant power control mode which cannot be reversed [3]. Unlike droop controlled MTDC systems, DC voltage

regulator in Master Slave-MTDC systems keeps DC voltage constant and compensates for power balance in DC grid including power losses in DC grid [1].

TABLE I. DC VOLTAGE CONTROL IN MTDC SYSTEMS

Criterion	Master-Slave	Voltage Margin	Voltage Droop
DC Voltage Regulation	Single.	Done by more than one but one at a time.	Two, more than two or all at a time.
Automatic Power Sharing	Unbalanced grid power transferred by Master to hold DC voltage constant.	DC voltage control transfers to the next priority converter when DC voltage matches its setpoint.	Unbalanced grid power is shared as per droop constants.
Stress on single converter(s)	Yes.	Yes.	No.
Power Quality	Good.	Oscillations are produced during transfer of DC voltage control.	Free from oscillations.
Control Structure	Simplest	Complex	Simpler
Ability to keep DC Voltage Constant	Yes.	DC voltage overshoot is observed while control is shifted.	Steady state error.

In this paper, an adaptive Master Slave control strategy for VSC-based MTDC systems has been proposed. It is interesting to note that the proposed scheme is based on a Converter Current Modulation (CCM) technique [2] employed by modifying vector current control scheme for VSC-HVDC converters. This control scheme enables a VSC terminal to be acting as Master in order to have improved power transferring capabilities than the conventional Master converter, thereby offering reliable and secure MTDC system in comparison to conventional Master Slave MTDC system. Further, reliability of Master Slave control system can be improved by incorporating a backup Master into MTDC system which can take over DC voltage regulation from Master if the later goes into outage. This is achieved by sending a communication signal to the backup Master about the Master terminal outage.

This paper is divided into six sections. Section II describes VSC-based HVDC transmission systems and conventional vector current control scheme for converter control. Section III explains MTDC systems and their control. The proposed scheme to improve Master Slave control based on Converter Current Modulation scheme is explained Section IV. Simulation and Results are presented in Section V to validate performance of proposed scheme in a four-terminal MTDC system by introducing a power variation in DC grid. Section VI provides the conclusions.

II. VSC-BASED HVDC TRANSMISSION SYSTEM AND VECTOR CURRENT CONTROL SCHEME

A. Description

A typical VSC-HVDC system consists of transformers, AC filters, phase reactors, DC capacitors, three-phase converter bridges and DC lines and/or cables. Fig. 1 shows a VSC-HVDC transmission system having symmetrical monopolar configuration using IGBT valves as switching devices. The DC capacitors keep DC voltage of the link

constant by their energy storing capacity. In addition, they offer DC harmonic filtering to avoid undesirable harmonic flow in AC systems. The phase reactors aid in the active and reactive power flow in between AC/DC system. AC Filters are also used to block flow of any harmonics from entering into the respective AC systems. Pulse Width Modulation (PWM) technique is used by VSC system to produce AC voltage waveform of desired magnitude, phase angle and frequency. The PWM employment reduces the filtering requirements of VSC-HVDC system with respect to conventional HVDC system [6].

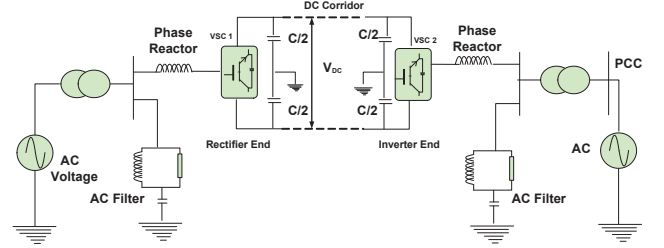


Fig. 1. VSC-HVDC transmission system having symmetrical monopolar configuration.

B. Vector Current Control Scheme for Voltage Source Converter

This control technique uses synchronous dq-reference frame approach for control of VSC to exercise independent control of active power and reactive power. From Fig. 2,

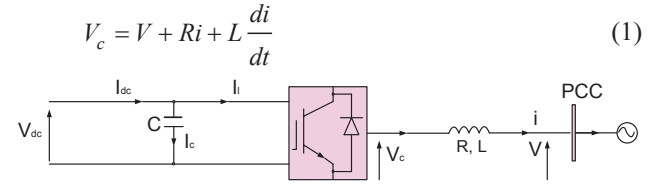


Fig. 2. A grid-connected VSC.

Performing three-phase to dq transformation on (1),

$$V_{c,dq} = V_{dq} + Ri_{dq} + L \frac{di_{dq}}{dt} - L\omega i_{qd} \quad (2)$$

To synchronize the converter output voltage, V_c with the grid voltage, V at PCC, a Phase Locked Loop (PLL) is employed. The output voltage of the converter is aligned with direct-axis component of the grid voltage by the PLL. Assuming, PCC voltage is constant and balanced, thus its quadrature component becomes zero [7]. Therefore,

$$P_{AC} = \frac{3}{2} V_d i_d \quad (3)$$

$$Q_{AC} = -\frac{3}{2} V_d i_q \quad (4)$$

Thus, by modifying d-axis current and q-axis current, the decoupled control of active and reactive power is achieved. This control technique uses Outer Control (OC) loop and Inner Current Control (ICC) loop as shown in Fig. 3.

The ICC loop obtains reference currents, i_d^* and i_q^* from the OC loop. Measured values of i_d and i_q are compared with their respective reference values. The error obtained is fed into PI controllers. Using (2),

$$V_{c,dq}^* = (k_p + \frac{k_i}{s})(i_{dq}^* - i_{dq}) + R i_{dq} - L \omega i_{dq} + V_{dq} \quad (5)$$

The reference converter output voltage, $V_{c,abc}^*$ is generated by performing inverse Park's Transformation on the obtained value of $V_{c,dq}^*$ from (5).

The conventional decoupled vector control strategy is shown in Fig. 3. The OC Loop generates the direct and quadrature axes reference currents, i_d^* and i_q^* for supplying to inner loop. The outer loop has two inside layers. The first one is having linear PI controllers and is known as PI Control Layer while the second layer is Current Generation layer as shown in Fig. 3. It can be noted that, for constant DC voltage and constant AC voltage control, i_d^* as well i_q^* are directly obtained from the PI control layer. However, if the VSC is operated in constant Active Power control mode and/or constant Reactive Power control mode, the reference current components, i_d^* and i_q^* are obtained from the Current Generation Layer.

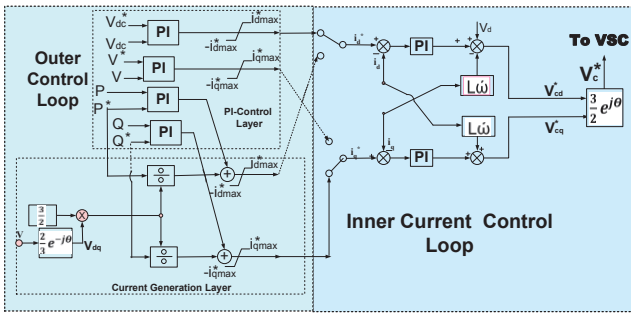


Fig. 3. Vector current control scheme for voltage source converter.

The PI Control layer consists of four different PI controllers (Fig. 3); each of which process error of reference and actual value of DC voltage, AC voltage, active power and reactive power respectively. Depending upon the application for which VSC is put to use, only two out of four PI controllers are in operation at a time. Depending upon the VSC control objective, only one out of DC voltage error processing PI controller or active power error processing PI controller operates at a time. Similarly, one out of AC voltage PI controller or reactive power PI controller is acting at a time. The reference current generation, i_{dq}^* is done by using following four equations [7] corresponding to the four control modes of VSC-HVDC transmission system:

$$i_d^* = (K_p + \frac{K_i}{s})(P_{AC}^* - P_{AC}) + \frac{2P_{AC}^*}{3V_d} \quad (6)$$

$$i_d^* = (K_p + \frac{K_i}{s})(V_{DC}^* - V_{DC}) \quad (7)$$

$$i_q^* = (K_p + \frac{K_i}{s})(Q_{AC}^* - Q_{AC}) + \frac{2Q_{AC}^*}{3V_d} \quad (8)$$

$$i_q^* = (K_p + \frac{K_i}{s})(V^* - V) \quad (9)$$

The first component in all the four equations as stated above is obtained from the PI control layer while the second component for (6) and (8) is obtained from the Current Generation layer in the Outer Control loop of the vector control scheme of VSC-HVDC system. For protection of IGBT valves, it is necessary to prevent overcurrent flowing via VSC. Thus, limits are applied on direct-axis and quadrature axis currents, i_{dmax}^* and i_{qmax}^* [7].

III. MULTITERMINAL HVDC SYSTEMS AND THEIR CONTROL

A VSC-based MTDC grid is formed by connecting more than two voltage source converters at their high voltage DC sides [1,8]. In such systems, one or more than one VSC may be transferring power to same AC system. Otherwise, each VSC will be inverting/rectifying into/from different synchronous/asynchronous AC systems. The VSC technology makes it easier to increase size of MTDC grid by just adding more voltage source converters as DC voltage is always same. A typical four-terminal MTDC network topology is shown in Fig. 6a, Section-V.

Like any point to point VSC-based HVDC system, a MTDC system should have at least one converter dedicated for control of DC voltage in the grid. Irrespective of power flowing via this converter, DC voltage should be maintained constant across its terminals or at least should be held within the permissible limits. A MTDC system where one & only one single terminal is dedicated to control DC voltage at a time is known as single node DC voltage control technique [3]. In Master Slave technology of VSC-based MTDC system, the converter terminal dedicated for DC voltage control is referred to as Master converter while rest converters are Slaves which are required to have constant power flowing through them. This DC voltage controlling Master controller is desired to exhibit following features:

- 1) It would regulate DC voltage of MTDC grid by compensating for unbalanced power in the grid due to a converter outage or change in active power delivered by any other slave terminal. [1].
- 2) It would be able to operate both in rectifier as well as inverter mode as per the power requirements of the DC grid [9].
- 3) It can be tied to a stronger AC network which can supply power to the DC grid during contingency and vice versa to enable minimal frequency variations [10-13].
- 4) While operating as rectifier, it would be able to supply losses in the DC grid.

Master converter will regulate DC voltage by compensating for unbalanced power in DC grid but is constrained by its design limits [10, 14]. In a Master-Slave regulated MTDC system, increase in DC grid voltage is avoided by Master converter absorbing more power from the grid while a drop in DC grid voltage will authorize master converter to supply scarce power into the DC grid [9, 13, 15-16]. Fig. 4 shows characteristics of master converter. It is depicted that Master terminal can balance power in the DC grid up to a maximum value of P_{max} , depending upon its converter rating.

In case of outage of master converter, DC voltage of MTDC grid will collapse due to absence of any other DC voltage controlling terminal. Such converter outage can take place due to some DC fault, etc. To increase reliability of Master Slave technology, a Backup Master converter terminal is often used. This backup terminal is chosen to be one among the slave terminals in the MTDC system. The controller of this converter terminal switches its action from constant power control mode to DC voltage control mode upon receiving a communication signal about the failure of Master terminal to control DC voltage further [3].

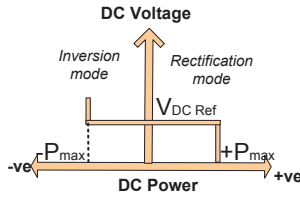


Fig. 4. DC-voltage power characteristics of master terminal of Master-Slave control.

DC voltage of grid can also become unstable if active power handling capacity of master converter reaches its rated value, P_{max} . A Master controller hitting its capacity limits, $\pm P_{max}$ will push the entire MTDC system into an unstable state resulting either in to overvoltage or to undervoltage [9,10]. Although, a converter outage due to a DC fault can't be avoided but it is possible to increase active power handling capacity of Master converter terminal so that DC voltage can be controlled stably by it up to maximum possible extent, thereby resulting in a strong MTDC grid operation. This can be done by Converter Current Modulation scheme explained in next section.

IV. PROPOSED CONVERTER CURRENT MODULATION SCHEME FOR MASTER CONVERTER OF MASTER-SLAVE MTDC SYSTEM

In order to obtain maximum active power support from VSC-HVDC transmission system during a frequency disturbance, Converter Current Modulation (CCM) Technique is employed in [2]. In present paper, this CCM technique has been utilized to increase active power handling capability of Master converter terminal of Master-Slave controlled MTDC system. The proposed scheme is very much similar to the traditional vector current control scheme used for controlling voltage source converters. Except that, here values of direct-axis and quadrature-axis converter reference currents, i_d^* and i_q^* , are determined without applying any constraints. In fact, the value of i_d^* is allowed to increase up to the maximum current that can be carried by IGBT valves used in the converter. This is done by introducing a Converter Current Modulation layer in the vector current control as shown in Fig. 5.

Conventionally, for DC voltage regulating vector current controlled Master terminal of MTDC system, its maximum active power handling capacity is constrained by i_{dmax}^* limit. Similarly, the reactive power compensating capability of voltage source converter is also limited by a similar i_{qmax}^* limit applied on quadrature axis reference currents. These

limits are applied on converter currents because a VSC is responsible to provide not only desired active power support but also reactive power compensation into interconnected AC network. However, if required, for maintaining DC voltage stability of MTDC grid, Master terminal can prioritize DC voltage control and thus, can increase its active power transfer capability to maintain constant DC voltage.

Whenever, DC voltage tries to vary from reference value, the limits applied on the converter currents can be dynamically altered. However, during modulation of converter currents, in response to DC voltage change, the maximum value of i_d^* is prioritized over i_q^* . Also, such improvisation in reference currents is entertained only when DC voltage stability of MTDC grid is at risk. If not so, the values of i_d^* and i_q^* calculated by vector current control scheme in Master converter control system holds valid.

Under normal conditions, when DC voltage of MTDC grid is maintained constant to reference value, the values of i_d^* and i_q^* obtained from PI control layer in OC loop are passed as it is to inner current control loop. However, when DC voltage of the grid starts to increase or decrease beyond the reference value, the values of i_d^* and i_q^* obtained from outer control loop become the input to the newly introduced CCM layer instead of entering ICC loop directly as shown in Fig. 5. However, no constraints are applied on the reference currents before entering CCM layer. But in CCM layer, i_d^* obtained from the OC loop is limited to a value equal to the maximum current that can be carried by IGBT valves, which is usually 1.5 times the rated current [17]. The value of i_q^* is calculated as vector difference between the maximum current carrying capacity of IGBT valves and the obtained direct-axis reference current. Since, i_d^* is free to be increased up to maximum valve current carrying capacity, at the same time, the magnitude of i_q can be reduced to zero.

By employing the CCM technique for controlling the Master terminal, the value of i_d^* (obtained from the PI controlled error between reference and measured DC voltage) can be increased to a greater extent with respect to vector current controlled Master terminal. Due to this increase in value of i_d^* obtained from the OC loop of CCM-controlled Master terminal, the amount of active power inverted or rectified by it into/from the DC grid to keep DC voltage constant is more than the conventional Master Slave MTDC system. Such CCM controlled Master Slave MTDC system, thus, can work with additional stability than the conventional Master Slave MTDC system.

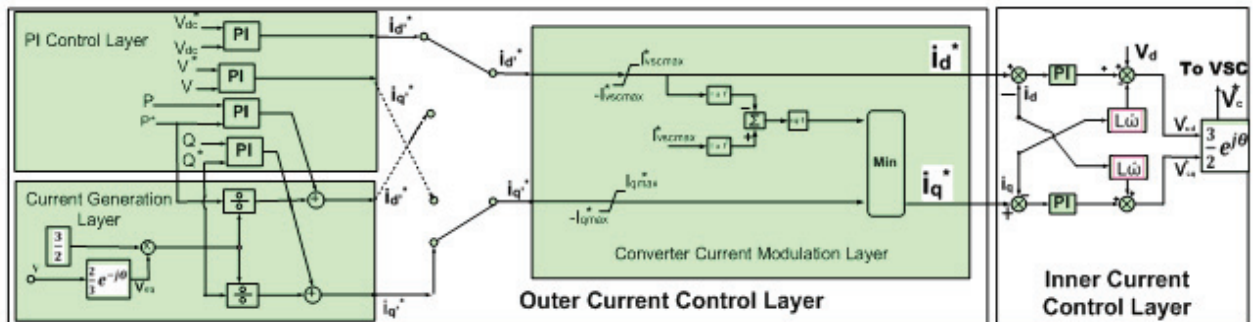


Fig. 5. Proposed scheme to increase active power handling capability of Master Slave controlled VSC-based MTDC system

For fulfilling the objective of increasing active power handling capability of DC voltage controlling terminal of Master-Slave MTDC system, the equations for generating i_d^* and i_q^* used in [2] are worked out as shown in (10) and (11). In these equations, V_{dc} represents DC voltage at the terminals of Master converter in MTDC grid. V_{dc}^* is reference value of DC voltage given to Master. V_{dcmin} and V_{dcmax} are the minimum and maximum allowed values of DC voltages in the MTDC grid. It can be observed that the reference value of reactive current is dependent upon both the DC voltage of grid and as well as upon the magnitude of the direct-axis reference current.

$$i_d^* = \begin{cases} = i_d^*, & \text{if } V_{dc} = V_{dc}^* \text{ or } V_{dcmin} \leq V_{dc} \leq V_{dcmax} \\ \\ = i_d^*, & \text{if } -I_{vscmax} \leq i_d^* \leq I_{vscmax}; \\ & \text{and if} \\ & V_{dcmin} \geq V_{dc} \quad \text{or} \quad V_{dc} \geq V_{dcmax} \\ \\ = I_{vscmax}, & \text{if } i_d^* \geq I_{vscmax} \text{ and if } V_{dcmin} \geq V_{dc}; \\ \\ = -I_{vscmax}, & \text{if } i_d^* \leq -I_{vscmax} \text{ and if } V_{dc} \geq V_{dcmax}; \end{cases} \quad (10)$$

$$i_q^* = \begin{cases} = i_q^*, & \text{if } V_{dc} = V_{dc}^* \\ & \text{or } V_{dcmin} \leq V_{dc} \leq V_{dcmax}; \\ \\ = \sqrt{(i_{vscmax}^*)^2 - (i_d^*)^2}, & \text{if } -I_{vscmax} \leq i_d^* \leq I_{vscmax}; \text{ and if} \\ & V_{dcmin} \geq V_{dc} \quad \text{or} \quad V_{dc} \geq V_{dcmax}; \\ \\ = 0, & \text{if } i_d^* \geq I_{vscmax} \text{ and if } V_{dcmin} \geq V_{dc}; \\ \\ = 0, & \text{if } i_d^* \leq -I_{vscmax} \text{ and if } V_{dc} \geq V_{dcmax}; \end{cases} \quad (11)$$

During a severe DC grid disturbance event when the unbalanced power present in DC grid is beyond the rated capacity of Master converter, by action of Current modulation, as explained above, DC voltage of MTDC grid can be brought back to the rated value. However, during this action of Master converter terminal, it may have to either neglect or reduce its priority for offering reactive power support to the interconnecting AC grid, only during the disturbance period. Once, DC voltage of the grid is brought back to rated value, generally a few seconds, once again, Master converter will continue to provide reactive power support offered by it to the interconnected AC network before the disturbance. However, during the period of disturbance, when lesser or no reactive power support is offered by the Master terminal to interconnected AC grid, the deficient reactive power can be supplied by other reactive power compensating devices in the AC grid. These can be either AVR of synchronous generators, FACTS devices, etc. Thus, not only DC voltage can be held stable but also AC voltage will be maintained within the permissible range.

V. SIMULATION AND RESULTS

A typical configuration for a four-terminal VSC-based MTDC system is shown in Fig. 6a. The present paper is focused upon VSC-based MTDC control systems; therefore, detailed AC grid modelling has been avoided. In fact, three-phase ideal voltage sources are considered to represent AC grids. Average order VSC-HVDC models have been considered over detailed switching models to optimize simulation times. Since, the present work does not require any simulation of converter outage or DC fault, thus, a symmetrical monopolar configuration is chosen for interconnection among the converters on the DC side. Each DC cable connection is represented by an equivalent pi-section model having two cascaded sections. The parameters of AC and MTDC systems are shown in Table II. The base MVA and base voltage values are taken as 200 MVA and 100 kV respectively.

A four-terminal MTDC system is considered in this paper based on representation in Fig. 6a. First it is controlled by conventional Master Slave control technique and is referred as MS-MTDC. In this conventional system, maximum values of i_d^* and i_q^* are considered as 1.1 pu and 0.8 pu respectively, thereby making i_{vscmax} as 1.36 pu.

MS-MTDC mentioned above is compared with another control scheme based on proposed CCM technique. This system is referred here as CCM-MS-MTDC. In Master converter of such system, initially no limits are applied on i_d^* and i_q^* . But the maximum current carrying capacity of IGBT valves is considered as 1.36 pu only. In MS-CCM-MTDC, only Master terminal is controlled by CCM technique, rest converters are controlled by vector current control scheme only. This is because other than master converters, others are operating in constant power mode so their operation via CCM or vector control does not affect the performance of master (whose performance is significant). Simulink representation of CCM-MS-MTDC is shown in Fig. 6b.

TABLE II. SYSTEM PARAMETERS

S.No.	System Components	Ratings
1.	AC System Rating	230kV, 50 Hz
2.	AC System Capacity	2000 MVA
3.	Transformer Ratings	4 no's each of 230 kV / 100 kV, 200 MVA, 50 Hz, R=0.0025 pu, X=0.075 pu.
4.	Phase Reactor	0.15 pu
5.	DC Side Voltage	± 100 kV
6.	DC Side Power	200 MW
7.	DC Cable Parameters	100 kV, r=0.139 mΩ/km, l=15.9 mH/km and c=23.1 μF/km.
8.	Length of DC Cable	4 X (2 X 75 km) symmetrical monopole
9.	DC Capacitance	70 μF

The converter terminals T1, T2 and T4 in both MS-MTDC and CCM-MTDC utilize active power controller while that at T3 is Master terminal and exercises constant DC voltage control having 1 pu as reference DC voltage. It is considered that DC voltage is allowed to vary within ±10% as the permissible limits. The sign convention used for power is positive (+ve) for rectified power into the DC grid and negative (-ve) for power inverted from the DC grid. Also, in MS-MTDC & CCM-MS-MTDC, converter terminals T1 and T2 are rectifying 1.0 pu & 0.97 pu

respectively into the DC grid while terminal T4 is inverting 1 pu from the DC grid. This is shown in Fig. 7. Also, for both of the MTDC systems, Master Terminal T3 is maintaining DC voltage of 1 pu as shown in Fig. 8 by inverting 0.93 pu (depicted by Fig. 7b) from the MTDC grid. Thus, in normal operating conditions, the behavior of both these systems is exactly same.

To illustrate the superior control capabilities of CCM-MS-MTDC system over the MS-MTDC, at $t=7s$, the active power set point of terminal T4 controller is decreased from -1 pu to -0.7 pu. In response to this decrease in active power inversion, the active power flowing via all the four terminals of MS-MTDC and CCM-MS-MTDC are shown in Fig. 9 and Fig. 10 respectively. It is observed that as the controllers at terminals T1 and T2 employ constant power controllers, thus, they do not depict any considerable changes in their output power after applying the active power variation at T4 at $t=7s$.

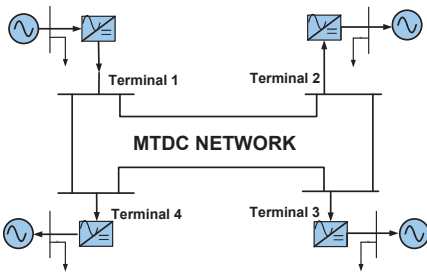


Fig. 6. (a) A typical configuration for a four-terminal VSC-MTDC system.

For DC voltage regulation, Master terminal will invert this extra active power left in DC grid which could not be inverted via terminal T4. As depicted by Fig. 9b, the Master terminal of MS-MTDC system can only increase its active power inversion from -0.929 pu to -1.195 pu. Any more active power inversion is not possible due to the predefined limits applied on the direct-axis reference current of the VSC controller. As the maximum allowed direct-axis reference current is 1.1 pu, therefore, Master terminal of MS-MTDC system is unable to invert active power beyond 1.195 pu from the DC grid. Due to constraints on direct-axis converter current, MS-MTDC system is not able to further invert

unbalanced power in the DC grid and as a result, the DC voltage of the grid rises and becomes more than 1.4 pu in just 1.7s as depicted in Fig. 11 and becomes unstable. Thus, making Master-Slave controlled MTDC system unstable. The direct axis reference current for this system is shown in Fig. 12.

However, in CCM-MS-MTDC system, to invert the additional active power left by terminal T4 into the DC grid, its Master controller terminal, increases its active power inversion from -0.929 pu to -1.22 pu, as seen in Fig. 10b. Due to the elimination of pre-imposed limits on the converter currents of the Master terminal of CCM-MS-MTDC system, i_d^* of this converter have flexibility to increase from 0.89 pu to 1.16 pu as shown in Fig. 13. However, still more active power can be inverted by this Master terminal (if required) as i_d^* is still lesser than i_{vscmax} (1.36 pu).

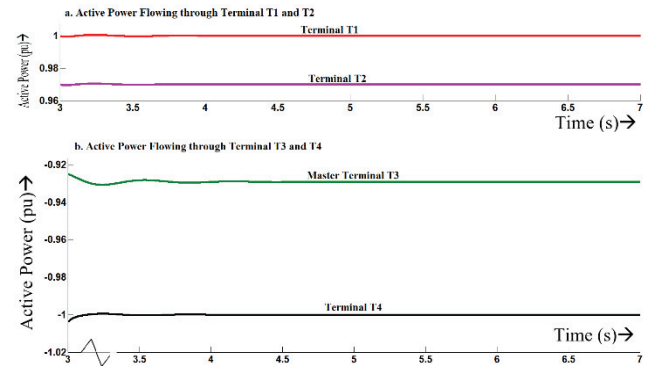


Fig. 7. Active power a. rectified by terminal T1 and terminal T2 b. inverted by Master terminal T3 and terminal T4 in MS-MTDC and CCM-MS-MTDC.

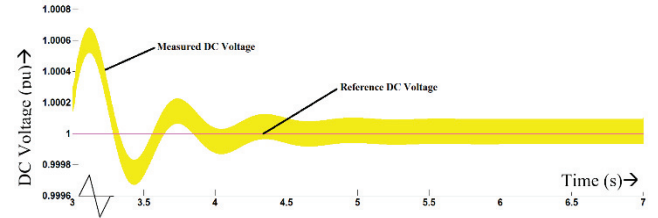


Fig. 8. DC voltage maintained by Master terminal T3 of MS-MTDC and CCM-MS-MTDC.

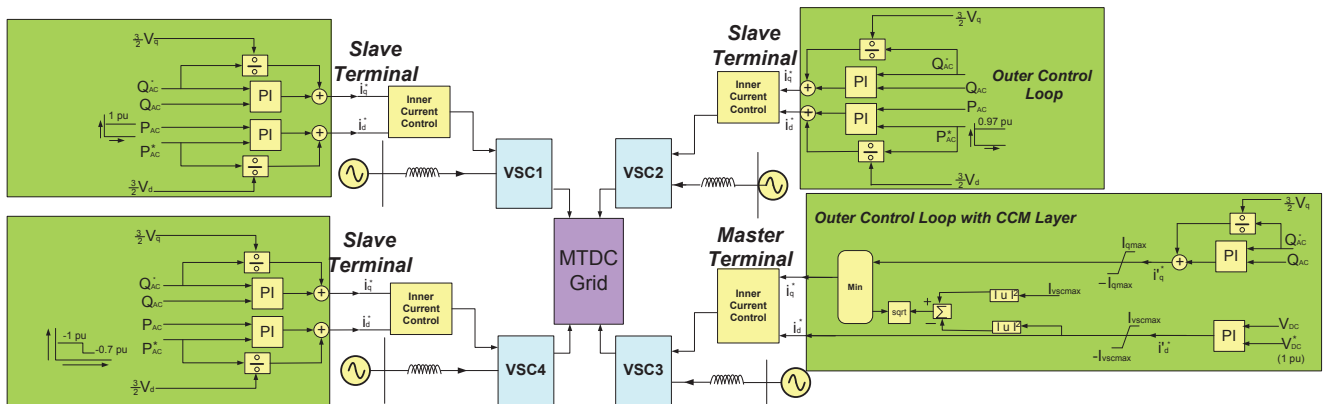


Fig. 6 (b) Converter Current Modulation based Master Slave controlled VSC- based MTDC system in MATLAB/Simulink.

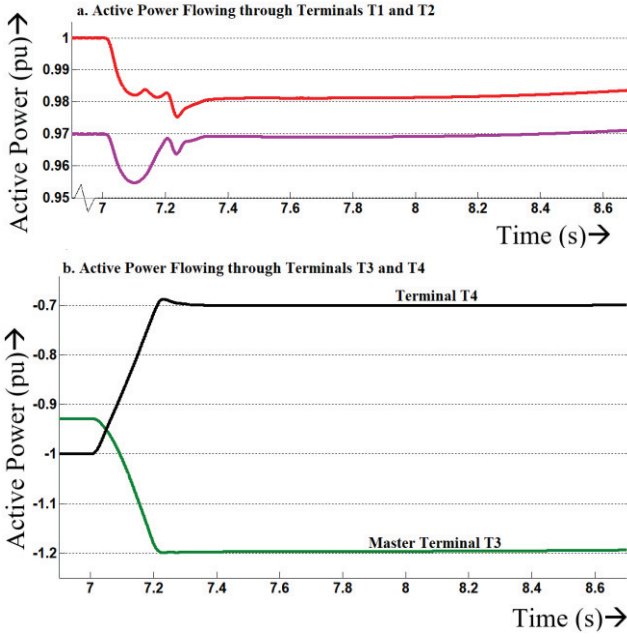


Fig. 9. Active power a. rectified by terminal T1 and terminal T2 b. inverted by Master terminal T3 and terminal T4 in MS-MTDC.

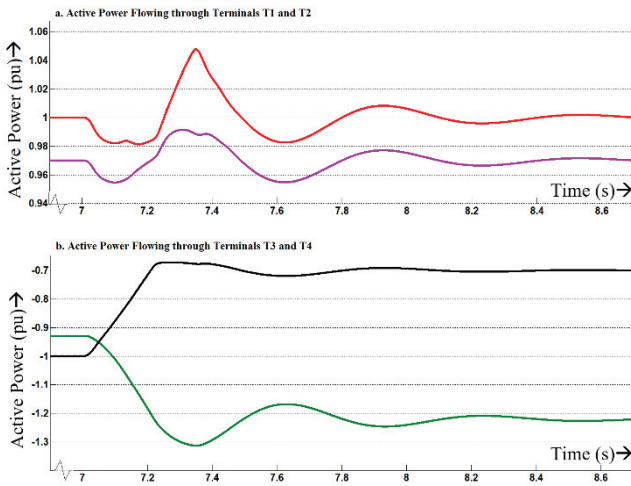


Fig. 10. Active power a. rectified by terminal T1 and terminal T2 b. inverted by Master Terminal T3 and terminal T4 in CCM-MS-MTDC into DC grid.

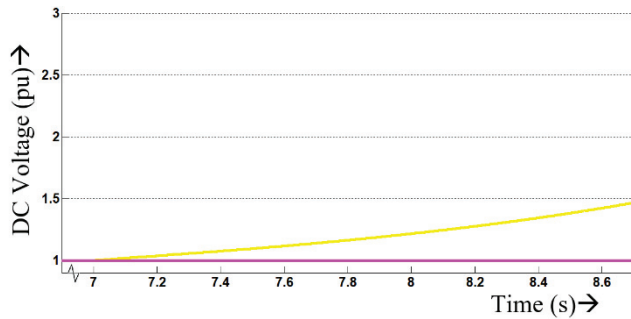


Fig. 11. DC voltage regulated by Master at T3 after applying active power variation at T4 in MS-MTDC.

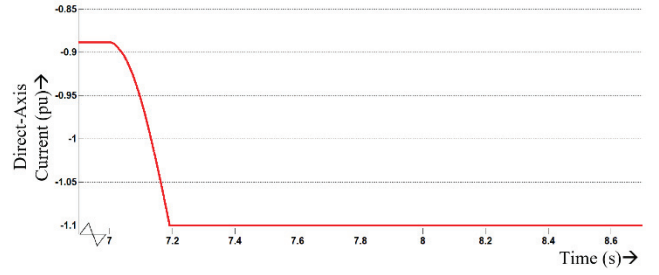


Fig. 12. Direct-axis converter reference current flowing via Master at T3 after applying active power variation at T4 in MS-MTDC.

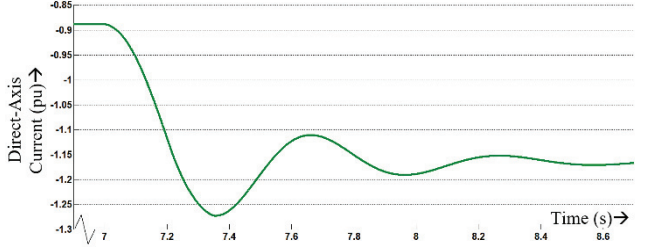


Fig. 13. Direct-axis converter reference current flowing via Master at T3 after applying active power variation at T4 in CCM-MS-MTDC.

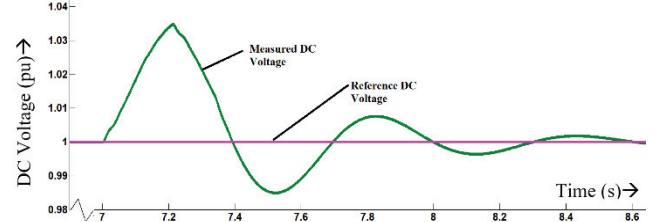


Fig. 14. DC voltage regulated by Master at T3 after applying active power variation at terminal T4 in CCM-MS-MTDC.

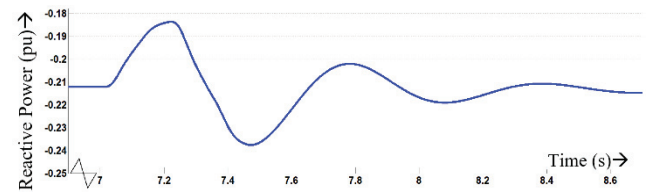


Fig. 15. Reactive power flowing through Master terminal T3 after applying active power variation at terminal T4 in CCM-MS-MTDC.

Initially, due to extra power left in DC grid at $t=7s$, results in increase in DC voltage of CCM-MS-MTDC system to 1.035 pu as illustrated by Fig. 14, but by current modulation ability of CCM-MS-MTDC system, Master terminal now inverts more active power and thus, aids to maintain DC voltage of grid to 1 pu.

It can be observed from Fig. 15 that in CCM-MS-MTDC system, reactive power delivered by T3 into interconnected AC system 3 is almost unaffected by applied active power variation even after application of CCM technique. This is because of two reasons. First, no reactive power demand change is made by interconnecting AC network 3. Secondly, the value of i_q^* (before the applied active power variation at T4) was less than vector difference between i_{vscmax} and i_d^* . So, in this case, increase in value of i_d^* has no effect on value of i_q^* , as the vector sum of i_d^* and i_q^* after the applied variation is still lesser than i_{vscmax} . This can be followed from equations (10) and (11) and also CCM layer shown in Fig. 5.

VI. CONCLUSION

In this paper, a converter current modulation-based DC voltage control scheme has been superimposed in Master-Slave VSC-based MTDC system for DC voltage regulating converter station. In effect, the active power handling capability of the Master terminal performing DC voltage control has been considerably augmented. This is done by allowing direct-axis reference current of Master converter to take value up to the maximum current handling capacity of IGBT valves; whenever DC voltage goes unstable. As a result of which, the DC voltage regulating VSC can compensate for more unbalanced power in MTDC system in comparison to its counterpart utilizing DC voltage control by conventional vector current control scheme, which has been tested in the four-terminal VSC-based MTDC system. In case of unbalanced power in DC grid, the result shows that more DC power can be inverted by the proposed converter current modulation-controlled Master converter than vector current controlled Master converter. The augmented control mechanism promises better DC voltage stability than conventional Master Slave technique used for VSC-based MTDC system thereby preventing severe overvoltage and undervoltage in DC grid.

REFERENCES

- [1] T. M. Haileselassie and K. Uhlen, "Precise Control of Power Flow in Multi-terminal VSC-HVDCs Using DC Voltage Droop Control", IEEE Power and Energy Society General Meeting, 2012, USA, pp. 1-9, 22nd-26th, July.
- [2] A. Taneja, R. Saha and M. Singh, "Frequency Regulation Technique in AC-DC Network using Converter Current Modulation in VSC-HVDC System", accepted in IEEE INDICON, 2020, New Delhi, India, pp. 1-8, 11th-13th, December.
- [3] Z. P. Cheng, Y. F. Wang, Z. W. Li and J. F. Gao, "DC Voltage Margin Adaptive Droop Control Strategy of VSC-MTDC Systems", AC and DC Power Transmission (ACDC 2019), 14th IET International Conference in the Journal of Engineering, 2019, pp. 1783-1787.
- [4] K. Rouzbehi, A. Miranian, J.I. Candela, A. Luna and P. Rodriguez, "A Generalized Voltage Droop Strategy for Control of Multi-terminal DC Grids," IEEE Trans. Ind. Appl., vol. 51, no. 1, pp. 607-618, Jan./Feb. 2015.
- [5] L. Shen, W. Wang and M. Barnes, "The Influence of MTDC Control on DC Power Flow and AC System Dynamic Response", IEEE PES General Meeting, 2014, MD, USA, pp.1-6, 27th-31st July
- [6] M. M. Alharbi and M. L. Crow, "Modelling of Multi-terminal VSC based HVDC System", IEEE Power and Energy Society and General Meeting (PESGM), 2016, Boston, MA, USA, pp. 1-5, 17th-21st July.
- [7] J. Zhu, C. D. Booth, G. P. Adam, A. J. Roscoe and C. G. Bright, "Inertia Emulation Control Strategy for VSC-HVDC Transmission Systems," IEEE Transactions on Power Systems, Vol. 28, No. 2, May 2013, pp 1277-1581.
- [8] Mier, P.G. Casielles, J. Coto and L. Zeni, "Voltage Margin Control for Offshore Multi-use Platform Integration", EA4EPQ International Conference on Renewable Energies and Power Quality (ICREPQ'12), 2012, Santiago de Compostela, Spain, pp. 1-6, 28th - 30th March.
- [9] M. Nazari. Control of DC Voltage in Multi-Terminal HVDC Transmission (MTDC) Systems. PhD. Dissertation Thesis, KTH Royal Institute of Technology, Stockholm, Sweden, June 2014.
- [10] J. Beerten, D. V. Hertem and R. Belmans, "VSC-MTDC Systems with a Distributed DC Voltage Control- A Power Flow Approach", PowerTech (POWERTECH), 2011 IEEE Trondheim, pp.1-6, 19-23 June 2011.
- [11] L. Dewangan and H.J. Bahirat, "Comparison of HVDC Grid Control Strategies", IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC), 2017, Bangalore, India, pp. 1-5, 8th Nov-10th Nov.
- [12] L. Zhang, Y. Zou, J. Yu, J. Qin, V. Vittal, G.G. Karady, D. Shi, and Z. Wang, "Modelling, Control and Protection of Modular Multilevel Converter-based Multi-terminal HVDC Systems: A Review", CSEE Journal of Power and Energy Systems, Vol. 3, No. 4, December 2017.
- [13] F. Torres, S. Martinez, C. Roa and E. Lopez, "Comparison between Voltage Droop and in Association of Automatic Control (ICA-ACCA), 2018, Great Concepcion, Chile, pp. 1-6, 17th-19th October.
- [14] J. Zhu and C. Booth, "Future Multi-terminal HVDC Transmission Systems using Voltage Source Converters", 45th International Universities Power Engineering Conference (UPEC), 2010 IEEE Wales, pp.1-6, 21st Aug-3rd Sept.
- [15] R. T. Pinto, S. F. Rodrigues, P. Bauer and J. Pierik, "Comparison of Direct Voltage Control Methods of Multi-terminal DC (MTDC) networks through Modular Dynamic Models", Proc. of the 14th European Conference on Power Electronic and Applications, 2011, Birmingham, UK, pp.1-10, 30th Aug-1st Sept.
- [16] P. Rodriguez and K. Rouzbehi, "Multi-terminal DC Grids: Challenges and Prospects", Journal of Modern Power Systems and Clean Energy, July 2017, Volume 5, Issue 4, pp. 515-523.
- [17] Y. Liu and Z. Chen, "A Flexible Control Method of VSC-HVDC Link for Enhancement of Effective Short-Circuit Ratio in a Hybrid Multi-Infeed HVDC System", IEEE Transactions. on Power Systems, Vol 28, No. 2, May 2013, pp 1568-1581.

Analysis of COVID-19 Tweets During Lockdown Phases

Prince Tyagi

Department of Mechanical Engineering
Delhi Technological University, India
Delhi, India
princetyagi193@gmail.com

Naman Goyal

Department of Mechanical Engineering
Delhi Technological University, India
Delhi, India
namangoyal99@gmail.com

Trasha Gupta

Department of Applied Mathematics
Delhi Technological University, India
Delhi, India
trashagupta@gmail.com

Abstract—With the advent of the internet among people in recent times, usage of social media and expressing views online has become part of everyone's routine. People are sharing their opinions on social media through text, videos, images, etc. Due to the nature of data shared on social media, it could be used to effectively analyze the emotions of humans, understand and model various events. One such event that happened in recent times is a pandemic due to the Covid-19 virus. Through this paper, we try to compare the emotions and sentiments of people worldwide during four phases of complete and relaxed lockdown through tweets. The four phases of lockdown are defined as Constricted Phase, Semi Constricted Phase, Semi Relaxed Phase, Relaxed Phase. This work will enable the community to provide useful insights and show how people adjusted and how they fought themselves to the pandemic.

Keywords— COVID-19, Twitter, Emotion Analysis, N-grams

I. INTRODUCTION

The pandemic caused by the novel corona virus was reported to have originated from China and in a few months engulfed the entire world. The first confirmed case of this explosive surge was reported in Wuhan, China¹. At the time of writing this document, there were around 82 million confirmed cases of COVID-19 globally, and around 1.8 M people had succumbed to the virus. US is the worst hit country with the maximum number of coronavirus cases followed by India and then Brazil at the time of writing this document². To fight the pandemic, preventive measures were taken by the countries worldwide in a bid to reduce its spread desperately. The major steps among these included: Initiations of country wide lock-downs, stay at home norms, Social distancing, and usage of masks and sanitizers. The lockdowns forced people into staying at their homes which led to huge impacts on economies, businesses, public events and almost every other day to day activities concerned with the human life. The company facing losses due to the pandemic led off many people leading to a large number of people jobless. Being unemployed and also getting infected by the virus spiked high levels of stress in the personal lives of people as well as in the communities. Studies of behavioural economics dictate that emotions and feelings (Happy and Joy, Optimistic, Confident, Depressed, Fear, etc.) of an individual

profoundly affects his/her decision-making skills along with the way they behave socially [1]. Social networking sites such as Facebook, Twitter, Instagram, etc., are perfect epitome of platforms which hold the potential of revealing valuable insights related to human emotions at both personal and community level. Many people turned to such media to express their reaction to the pandemic and an increase of posts and tweets related to COVID-19 was observed. Among these, examining tweets is particularly much more of value during and after COVID-19 pandemic due to its robustness to capture the unprecedented rate of changing reactions of people to the situation during these hard times. Thus, the analysis of twitter data focused on COVID-19 might provide significant insights to comprehend the people behaviour and response to the pandemic. To make the analysis constrained, we have defined and collected the tweets as per the basis of the lockdown phases. We defined the analysis time period during pandemic into 4 phases as shown in Table I.

Through this study, we are trying to understand the sentiment and human emotions through the Lockdown Phase's perspective globally. This study brings out the sentimental exposure and expression of the people in a bounded time frame and study general opinion about covid during this time period. Create these components, incorporating the applicable criteria that follow. The categorization of these lockdown is inspired from Lockdown Phases in India.

TABLE I. LOCKDOWN PHASES CATEGORIZATION

Phase	Time Period	Description
Phase 1	25 th March – 14 th April	Constricted Phase
Phase 2	15 th April – 3 rd May	Semi Constricted Phase
Phase 3	4 th May – 17 th May	Semi Relaxed Phase
Phase 4	18 th May – 31 st May	Relaxed Phase

II. RELATED WORK

Online Social Media like Twitter generates a lot of data daily which could be used for a plethora of research fields like mass communication, engineering, social science, and psychology. Twitter has remained a very popular choice among researchers

¹ <https://health.economictimes.indiatimes.com/news/diagnostics/coronavirus-in-china-may-have-come-from-bats-studies/73923178>

² <https://www.worldometers.info/coronavirus/>

for studying emotions of people, reaction to a particular calamity or disaster. In the past, Twitter has been used to study Effect of Hurricanes on Human Mobility [2], predicting flu trends [3], electoral prediction [4], sentiment analysis [5]. Christian et al [6] analysed corpus of tweets that relate to the COVID-19 pandemic and identified common responses to the pandemic using Text Mining, Natural Language Processing, and Network Analysis to and analysed how responses change with time. Lisa et al [7] used twitter data on information and misinformation to COVID-19. During the time of COVID-19 both information and misinformation are spreading on the Internet space. The author provided initial step to understanding discussions pertaining to COVID-19 via social media. The research concluded with author providing various implications for understanding disease spread, information seeking behaviours during public health crises, and general communication patterns in this unprecedented combination of global pandemic and modern information environment. A lot of researchers around the world are studying the impact of corona virus on the human emotions.

III. DATA DESCRIPTION AND COLLECTION

The data pertaining to Covid Pandemic is collected from twitter using twint library during 25th March to 31st May. We have collected more than 2 million tweets which is around 200 MB of data stored a csv file. We have used COVID-19 related keyword 'covid' to collect the data. We processed various features from the tweets such as Tweet ID, time, Tweet Text, name ,place ,user id, etc. if available. Figure 2 is a high-level description of data along with the features scraped from twitter:

	id	conversation_id	created_at	date	time	timezone	user_id	username
0	1258909545000325122	1258909545000325122	2020-05-08 23:59:59 UTC	2020-05-08	23:59:59	0	82158439	luiboonen
1	1258909544870285313	1258909544870285313	2020-05-08 23:59:59 UTC	2020-05-08	23:59:59	0	935342394	anaimbeychris
2	1258909543783948289	1258909543783948289	2020-05-08 23:59:59 UTC	2020-05-08	23:59:59	0	1292561	sjenkins
3	1258909543624585222	1258909543624585222	2020-05-08 23:59:59 UTC	2020-05-08	23:59:59	0	4835887404	kathin_mariano
4	1258909541900705792	1258909541900705792	2020-05-08 23:59:58 UTC	2020-05-08	23:59:58	0	1256833849067274242	macelosoresentha
5	1258909540050968582	1258909540050968582	2020-05-08 23:59:58 UTC	2020-05-08	23:59:58	0	1168324007986388996	gabriel36053482
6	125890953944864128	125890953944864128	2020-05-08 23:59:58 UTC	2020-05-08	23:59:58	0	222496725	ekwulu
7	1258909538377498624	1258909538377498624	2020-05-08 23:59:58 UTC	2020-05-08	23:59:58	0	104996245	korifeener

Fig. 2. Data description.

IV. DATA PRE-PROCESSING AND CLEANING

The raw tweets data collected for each phase is pre-processed in order to perform the analysis in a much effective and efficient manner. Unprocessed tweets may interrupt all sorts of analysis as it contains numerous stop words and improper words which lead to ambiguous results. The data pre-processing includes all the necessary steps involved for cleaning the data, and also transforming it into something meaningful. This converts the text data into much more comprehensive state enabling machine learning algorithms to run and perform better. The pre-processing is done in the following stages:

Removing of Stop words: Generally, a stop word can be defined as a word which most commonly used in a language. In English examples of these words would be such as “the”, “a”, “an”, “in”, etc. These words do not add much significant meaning in a sentence nor do they change the topic used. Thus, such words can be safely ignored while keeping the meaning of a sentence intact. Removal of these words allows the algorithm to focus more on those words which contribute to the definition of the text.

Tokenization: The next step of pre-processing is dividing of phrases and sentences of the tweets into small units or in form of individual words. Each of this newly obtained smaller unit is referred to as a token and the process is called tokenization. The tokens extracted aid in developing better modelling and representative understanding of the context of the text processed.

Lemmatization: Lemmatization is a process which involves grouping together words with the same root and keep them in a non-inflected form. This enables the inflected forms of the words to be analysed as one single item. This process is very much similar to stemming but has an added advantage of bringing context to the words. To explain in easier terms, lemmatization connects words with almost same meaning to one single word. For E.g.: “help”, “helping”, and “helper” are all linked and reduced to “help”.

V. EXPERIMENTS AND RESULTS

We have done numerous analysis on the data scraped from twitter. During the Lockdown Phases, we have tried to compare

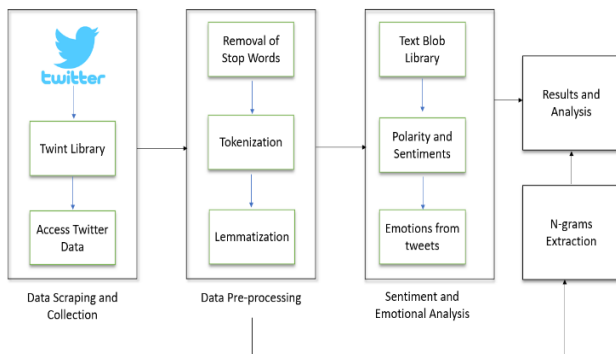


Fig. 1. Methodology.

The processed data is saved in the data repository within the folder named as “data”³. The data folder contains 4 csv files each having data related to each phase. Every file consists of tweets for the particular phase that is specified as the name of that file. The data is divided and collected according to Lockdown Phases as given below: Phase 1: 25th March to 14th April Phase 2: 15th April to 3rd May Phase 3: 4th May to 17th May Phase 4: 18th May to 31st May.

³ https://github.com/princetyagitech/phase_data

A. Frequent Words in each Phase

Phase 1: Figure 3 illustrates the frequently occurring keywords the Phase 1, which are: Covid, coronavirus, pandemic, health, people, test, new case, death, and time. Phase 1 was constricted phase of the lockdown around the world. During this period, people of the world seemed to be talking much about the pandemic and discussing way to deal with it.

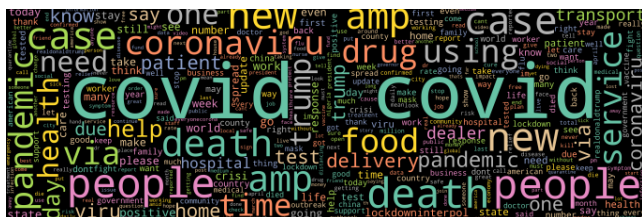


Fig. 3. Phase 1 word cloud



Fig. 4. Phase 2 word cloud

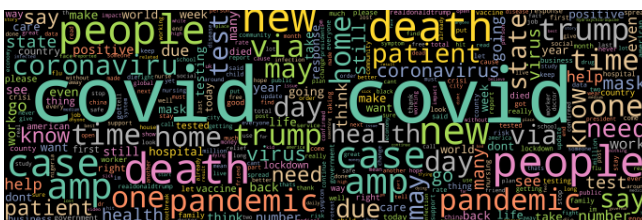


Fig. 5. Phase 3 word cloud

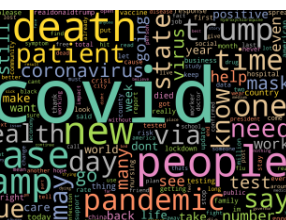


Fig. 6. Phase 4 word cloud

Phase 3: During phase 3 many services were restored to some extent while intra movement of people was controlled. The most frequent words appeared in the tweets are: Covid, People, may, death, pandemic, test, patient, trump, time, home, etc. (Figure 5). During this semi relaxed phase in order to save the

Phase 4: In this relaxed phase the most frequently occurring words are: mask, back, life, death, health, new, case, pandemic, number, may, still, home, etc. (Figure 6). The use of masks is now being discussed along with a still ever-increasing number of deaths and COVID-19 cases.

B. Sentiment (Polarity) during the 4 Phases of Lockdown

With the aim to understand the people's reaction during the COVID-19 pandemic, we performed extensive analysis on the sentiments of the shared tweets and the users in different phases of lockdown. The sentiment analysis was done by using 3 polarities - Positive, Neutral, Negative. We used the Sentiment intensity analyser from python Text Blob library. We have defined ranges to polarity to categorize tweets into three categories- positive, neutral and negative. Table II. provides ranges used for sentiment categorization.

TABLE II. RANGES USED FOR SENTIMENT CATEGORIZATION

Range	Sentiment
0.05 to +1.0	Positive
-0.05 to +0.05	Neutral
-0.05 to -1.0	Negative

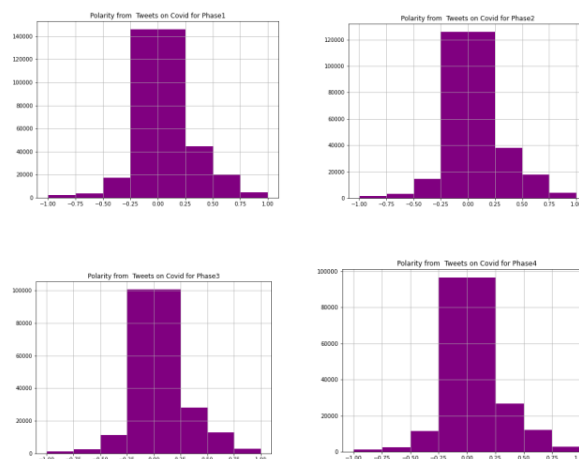


Fig. 7. Polarity distribution during all phases of lockdown

We found that there is good contrast between the content shared by positive and negative users during the different phases. Fig. 6 and 7 below represent the temporal sentiments in the collected tweets. We observed that the polarity of the sentiments is distributed across the scale while it mostly ranges between -0.40 to 0.40. However, in some we can see a spike in positive or negative polarity. Also, we observed that most of the tweets are in neutral zones with a bit of positive polarity. It is quite visible from the Fig. 7 that most of the negative emotions occurs in the range -0.4 to 0.6 compared to positive emotions. The number of positive, negative and neutral tweets are presented in Fig. 7 and

Fig. 8 according to the phases. We can see that the number of tweets with negative polarity is fairly large in the data in the first few phases of lockdown and this number decreases as the lockdown proceeded.

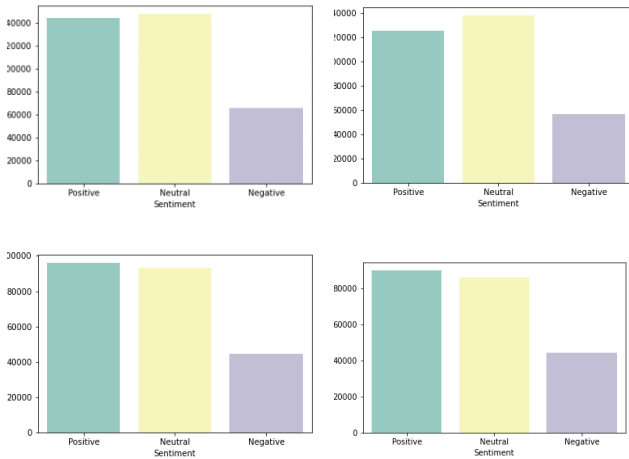


Fig. 8. Sentiment distribution from phase 1 to 4 of lockdown

C. Emotions during the 4 Phases during of Lockdown

The Table III. shows emotion scale [8] used for analysing emotions of people and has been applied to polarity estimation. The highest emotion on the scale is happy and joy with value 1 which denotes better-feeling, while Depressed and Fear is the lowest on the scale with value -1 denoting a much more negative emotional state. The middle emotional state with 0 value represents neutral and relaxed state. The rest of the emotions are distributed across these above-mentioned emotions in a spectrum. From the emotional charts in Figure 9 we find that the majority of reactions throughout all the phases is of neutral and relaxed, hopeful, calm and confident. The minority class is comprised of Discouraged and difficulty, Depressed and Fear, Happy and Joy. As lockdown has proceeded from Phase 1 to Phase 4, negative emotions have decreased while positive emotions have increased as illustrated in the Figure 9.

D. Bigrams

Text analytics provides a powerful way to find the context in the large dataset. It could be used to find useful insights in the data through finding unigrams, bigrams, trigrams or n-grams.

TABLE III. EMOTIONAL SCALE [8]

Scale	Emotion
1	Happy and Joy
0.8	Confident
0.6	Optimistic
0.4	Hopeful
0.2	Calm and Content
0.0	Neutral and Relaxed
-0.2	Relieved
-0.4	Pessimistic and impatient
-0.6	Worry and Boredom
-0.8	Discouraged and difficulty

Scale	Emotion
-1.0	Depressed and Fear

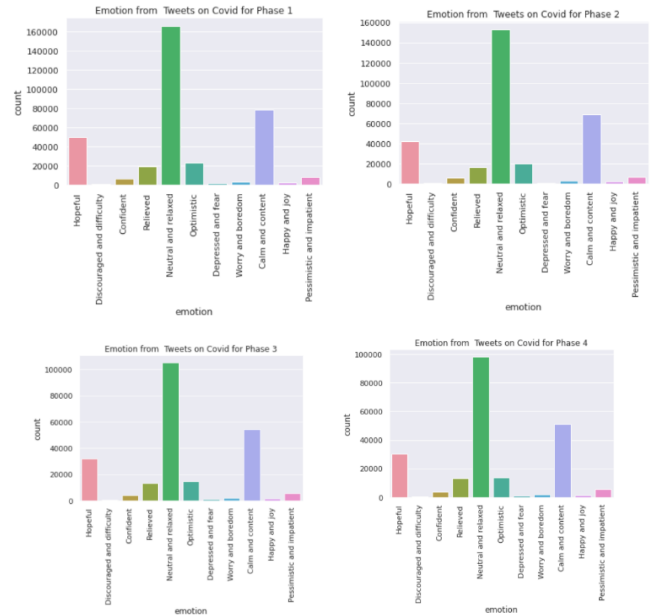


Fig. 9. Emotional distribution during all phases of lockdown.

The most frequently occurring pair of words used together in a document is referred to as a Bigram. In the similar fashion a frequently occurring group of three words used together is called a trigram. Bigrams can be very useful when text is very large and we have to find the occurrence of two words which occur together in the data. To extract these, the data was first made free of stop words. Following the stop word removal, the sentences were tokenized and broken into smaller bits of words. These tokens were then lemmatized and then collected in a single text document. The **nlk** library was then used to extract bigrams from this formerly saved text document of lemmatized tokens. We have shown bigrams for each phase of lockdown. The top bigrams in each phase show what people are going through and believing as the phases are proceeding further. We have shown the top 50 frequent bigrams for each phase individually.

Phase 1: The most common bigrams during this phase are ‘covid case’, ‘covid death’, ‘tested positive’, ‘stay home’, ‘fight covid’, ‘social distancing’ point towards the spread of virus during this time and various measures taken by people to stop its spread through social distancing. During this period a lot of death due to covid happened and people are asked to stay at home.

Phase 2: ‘food delivery’, ‘nursing home’, ‘covid crisis’, ‘covid vaccine’, ‘service transport’ signalling towards the fact that lot of people are ordering food online and people are talking about getting vaccine for covid in future.

Phase 3: A lot of bigrams during this phase are related to death due to covid like ‘death toll’, ‘died covid’, ‘death rate’ and few related to stop spread of covid such as ‘wear mask’, ‘stay home’.

Phase 4: Many of bigrams are related to testing and health during this phase- 'covid test', 'covid testing', 'public health'. During this phase people are staying at home. People are also aware to get themselves tested for covid.

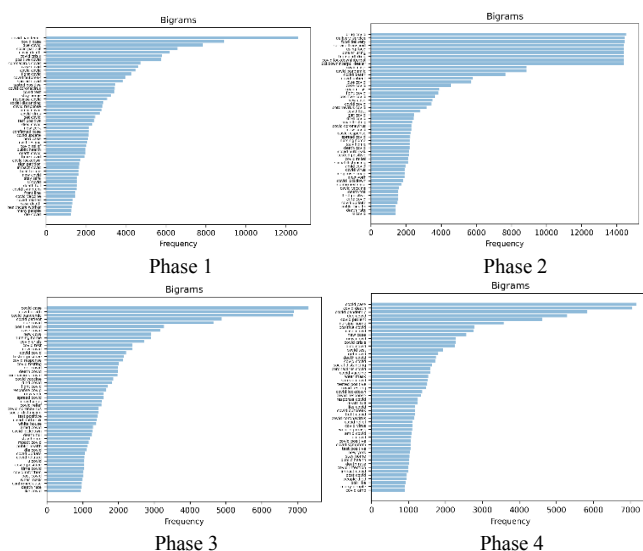


Fig. 10. Bigrams from phase 1 to phase 4

VI. CONCLUSION

This is the study of sentiment and emotional analysis on twitter data related to COVID-19 pandemic. The data has been collected during 25st March 2020 to 31st May 2020 by using **Twint**. We have shared our dataset publicly which further could be used for further research and analysis. We performed extensive sentiment analysis and related that with frequent keywords to find out the reason behind sentiment for better understanding of the human emotions during each phases of lockdown. We have successfully presented all the emotions and sentiment in the research.

ACKNOWLEDGMENT

There is no direct funding for this work. Authors are highly thankful to Delhi Technological University for providing basic facilities for the work. We also like to thank Twint library for

providing open-source code which proved to be useful in collecting the data.

REFERENCES

- [1] Virlics, Agnes. "Emotions in economic decision making: a multidisciplinary approach." *Procedia-Social and Behavioral Sciences* 92 (2013): 1011-1015.
- [2] Ahmouda, Ahmed, Hartwig H. Hochmair, and Sreten Cvetojevic. "Using Twitter to Analyze the Effect of Hurricanes on Human Mobility Patterns." *Urban Science* 3.3 (2019): 87.
- [3] Achrekar, Harshavardhan, et al. "Predicting flu trends using twitter data." 2011 IEEE conference on computer communications workshops(INFOCOM WKSHPs). IEEE, 2011.
- [4] Gayo-Avello, Daniel. "A meta-analysis of state-of-the-art electoral pre-diction from Twitter data." *Social Science Computer Review* 31.6 (2013):649-679.
- [5] Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." *Proceedings of the workshop on language in social media (LSM 2011)*. 2011.
- [6] Lopez, Christian E., Malolan Vasu, and Caleb Gallemore. "Understanding the perception of COVID-19 policies by mining a multilanguage Twitterdataset." *arXiv preprint arXiv:2003.10359* (2020).
- [7] Singh, Lisa, et al. "A first look at COVID-19 information and misinformation sharing on Twitter." *arXiv preprint arXiv:2003.13907* (2020).
- [8] Sharma, Karishma, et al. "Covid-19 on social media: Analyzing misinformation in twitter conversations." *arXiv preprint arXiv:2003.12309* (2020).
- [9] Pokharel, Bishwo Prakash. "Twitter sentiment analysis during covid-19 outbreak in nepal." Available at SSRN 3624719 (2020).
- [10] Kleinberg, Bennett, Isabelle van der Vegt, and Maximilian Mozes. "Measuring emotions in the covid-19 real world worry dataset." *arXiv preprint arXiv:2004.04225* (2020).
- [11] Lwin, May Oo, et al. "Global sentiments surrounding the COVID-19pandemic on Twitter: analysis of Twitter trends." *JMIR public health and surveillance* 6.2 (2020): e19447.
- [12] Kleinberg, Bennett, Isabelle van der Vegt, and Maximilian Mozes."Measuring emotions in the covid-19 real world worry dataset." *arXivpreprint arXiv:2004.04225* (2020)
- [13] E. Kušen, G. Cascavilla, K. Figl, M. Conti and M. Strembeck, "Identifying Emotions in Social Media: Comparison of Word-Emotion Lexicons," 2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), Prague, 2017, pp. 132-137, doi: 10.1109/FiCloudW.2017.75.
- [14] Groh, Georg, and Jan Hauffa. "Characterizing social relations via nlp-based sentiment analysis." *Proceedings of the International AAIL Conference on Web and Social Media*. Vol. 5. No. 1. 2011.

Comparative Study of Different Image Captioning Models

Sahil Takkar

Department of Computer Science
and Engineering
Delhi Technological University
Delhi, India
Email:
sahiltakkar4@gmail.com

Anshul Jain

Department of Computer Science
and Engineering
Delhi Technological University
Delhi, India
Email:
jainanshul3863@gmail.com

Piyush Adlakha

Department of Computer Science
and Engineering
Delhi Technological University
Delhi, India
Email:
piyushadlakha204@gmail.com

Abstract—This paper has compared various deep learning models for generating caption of images gathered from Flickr 8k Dataset. Also, this research work attempts to combine a CNN type encoder for extracting features from images and a Recurrent Neural Network for generating caption for the extracted features. The CNN encoders used are VGG16 and InceptionV3. The extracted features are then passed to a unidirectional or a bidirectional LSTM for generating captions. The proposed model has used beam search as well as greedy algorithms to generate captions from vocabulary. The generated captions are then compared with actual captions with the help of BLEU scores. The Bilingual Evaluation Understudy score (BLEU) is used to compare how close a given sentence is to another sentence. The BLEU score of captions generated using beam search as well as greedy algorithms are analyzed and compared to see which is better.

Keywords—VGG16, InceptionV3, Bidirectional LSTM, BLEU, Beam Search

I. INTRODUCTION

Nowadays, Artificial Intelligence (AI) is a very important part of the innovation sector and hence the basis of our project is also Machine Learning and AI. In the recent history, the sector of Deep Learning[1] has impressed everybody when compared to already famous currently present Machine Learning(ML) methodologies like Decision Trees, Logistic Regression, SVM, Naive Bayes, K Nearest Neighbors, Random Forest, etc. because of its extraordinary results in terms of accuracy as compared to that of already present traditional Machine learning models like KNN, Logistic Regression etc. It is a difficult task to generate a relevant description for an image but once done it can prove to be a great benefit to society. This can help visually impaired people to have better understanding of their surroundings by generating a suitable caption for an environment. It has many other applications like usage in virtual assistants, recommendations in editing applications, for social media etc.

Generating a caption[2] from an image is a notably harder task as compared to that of classifying an image, which has been the centre of attraction for the computer vision community. A description for an image must take into account the relationship between different objects presents in the image. Along with the visual description of the objects in image, the knowledge mentioned above has to be stated in a natural language understandable by humans. It means that, a language model is required, where it not only understands the image but also expresses it in a natural language. The attempts made in the past have all been to use two different models, one for understanding the image[3] and another for

using that understanding to generate a caption and then stitching the two models together.

The proposed method has attempted to combine the two models into one combined model, which consists of a CNN[4] type encoder that aids us in extracting features of image by creating encodings of images. Here, the pre-trained VGG16 and Inception V3 architecture model is used for encoding images. The CNN encoders extract features from the image and store them in the form of numerical encodings which can be easily understood by the machine. These extracted features are then passed to a type of Recurrent Neural Network namely LSTM network. The network architecture of the LSTM network works in almost the same way as that used in natural language machine translators. LSTM is replaced with bidirectional LSTM in order to see which one works better for prediction captions from extracted features.

The proposed project uses the Flickr 8k set of data which consists of eight thousand (8000) images and for each and every image, there are 5 captions respectively. By default, the dataset is splitted into two folders, image folder and text folder. For each image the caption is stored along with the respective ID as there is a distinctive image-id for every image in the set. The images in the dataset are divided into three parts: Training set, development set and Test set. Test and development set consist of 1000 images each whereas the training set consists of 6000 images. The model predicts a caption based on the vocabulary it creates from the tokens of words that it gathers from descriptions of images gathered from the training dataset. The description predicted by our model is then compared with the actual description provided in the dataset via BLEU score.

The upcoming sections of the paper will briefly discuss about the tools, techniques and dataset. Also, this research work attempts to discuss the CNN encoder namely VGG16 and Inception-v3 and the RNN[5] decoders namely LSTM and Bidirectional-LSTM decoder in full length. Also, this research work discusses about algorithm for caption generation, which has used to generate the predicted captions, namely argmax and Beam search. The BLEU score metric is used for comparing the accuracy of the different image captioning models being proposed. BLEU score helps in analyzing the text quality which has been generated by the ML model. BLEU score was among the earliest developed metrics to get such high correlation with actual verdict. The value of BLEU score always lies between 0-1. If BLEU score is zero, it means machine translation is not relevant to actual description at all. On the other hand, a BLEU score of 1 means that the machine translation is equivalent to actual description. BLEU score has also been discussed in detail in coming sections. At the end this paper includes the examples

of some images, which have been used to test the proposed model.

II. RELATED WORK

Caption Recommender System is an integral part of understanding the environment, which has various applications (e.g. - subtitle generation, helping visually impaired people to understand their surroundings, storytelling from albums, search using image, etc.). Since many years, many different image caption recommendation approaches have been developed.

There have been a lot of contributions from the architecture created by the winner of the ILSVRC. Along with the VGG the research made in the field of natural language translation have helped us continuously in bettering the performance in text generation.

Researchers at AI Lab used a Convolution Neural Network for each potential object in the image for producing high-level features of the image. Then a Multiple Instance Learning (MIL) [6] was used for figuring out the best area which matches with each word. This method gave a BLEU score of 22.9% on MS-COCO dataset.

The Vinyals came up with a new model called NIC (Neural Image Caption), Show and Tell model [7], which was nothing but an encoder RNN which was given input through a CNN model for computer vision. After this a group of researchers took the NIC model and modified it. They used a technique that makes use of images datasets and their corresponding captions to study the inter-modal correlations between natural language and image data. The model used by those researchers was based on a new combination of CNN around image fields, the LSTM or bidirectional RNN over textual descriptions, and a planned aim of putting the two modals together via bimodal embedding. Flickr 30K, Flickr 8K and MSCOCO were the datasets used by them to achieve these bests in business results. Jonathan further modified their model in 2015 when he suggested an idea of a model related to dense captioning in which the model detects each of the different areas of the image and then suggests a group of captions. Chen Wang also suggested a model which makes use of multiple LSTM networks and a deep CNN in the year 2016.

Over a period of time there has been enhancements not only in the captioning models but also in score metrics used for evaluating the accuracy of the models. This project has used the BLEU score for evaluation. BLEU - being a standard evaluation metric adopted by many of the groups. Now, new state of the arts metrics has come like CIDEr which are replacing older metrics like BLEU score, etc. CIDEr was proposed by Vedantam [8].

III. APPROACH

Recent developments in the field of technologies related to image captioning has been the main source of motivation for our research work. The model proposed in this paper has an eventual aim as to predict natural language descriptions for various areas of the image.

The research work focuses mainly on obtaining the results for several image captioning models by making use of BLEU score metric and hence comparing the performance of different image captioning models. Various CNN models

such as VGG16, Inception-V3 etc are used for encoding the images and extracting features from the images. Further these encoded images are used with two types of decoders, namely unidirectional LSTM and bidirectional LSTM to obtain the results. We have used greedy search and beam search algorithm to generate the caption from encoded features. The generated caption is then compared with the original caption from dataset on the basis of Bilingual Evaluation Understudy score.

A. Convolution Models : Encoders

This section discusses various convolution models used for the research work. There are two encoders namely, VGG16 and Inception-V3. Each convolution model has been described in brief in the following subsections.

VGG16: VGG16[9] consists of a 16-layer network for the completion of the task of encoding the image. Out of 16 layers present in the VGG16 network, 3 are dense layers and rest 16 are convolution layers. The architecture of VGG16 is shown in Fig. 1. For the feature extraction to be done on the image, the dimension of the image has to be a 224×224 image. We have fixed the length of the stride to be 1 for the CNN layer which have filters of size 3×3 . The next step is Max pooling, it is executed using a window size of 2×2 -pixel with a length of stride taken to be 2.

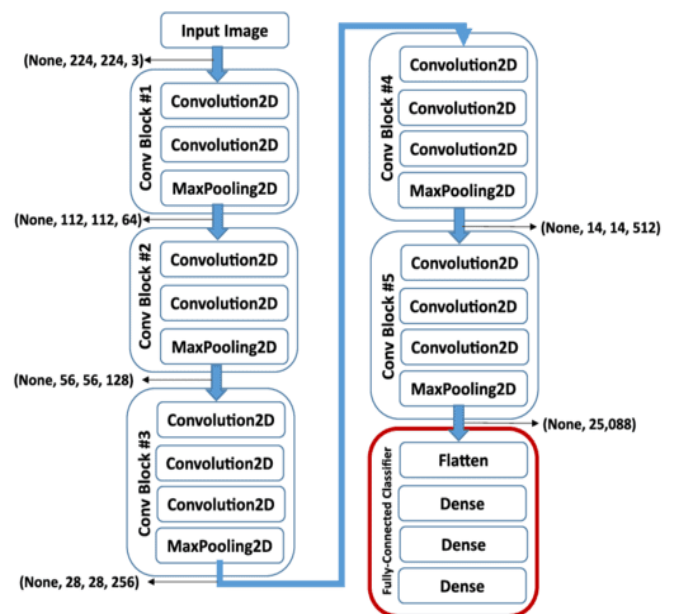


Fig. 1. VGG16 Architecture

Inception-V3: InceptionV3[10] consists of a 48-layer deep convolutional network for performing the task of encoding the image. InceptionV3 stacks together 11 inception modules each of which consists of convolution and max-pooling layers. For the feature extraction to be done on the image, the dimension of the image has to be a 229×229 image. Three fully connected layers of size 512, 1024 and 3 are added to the final concatenation layer. The architecture of Inception-V3 is shown in Fig. 2.

B. Decoders

This section discusses various decoder models used for the generation of captions for images. There are two decoders used in this research work namely, unidirectional

LSTM and Bidirectional LSTM. Each type of LSTM network has been described in brief in the following subsections.

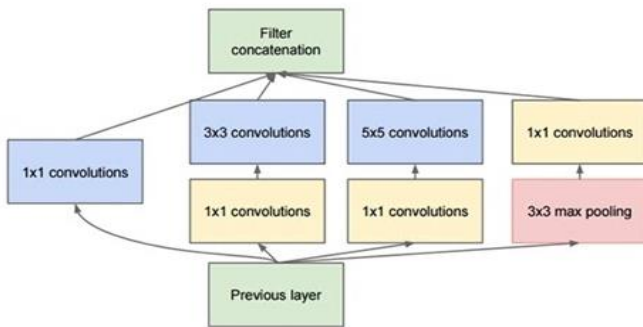


Fig. 2. Inception-V3 Architecture

LSTM (Long Short-Term Memory): LSTM[11] have been widely used by the researchers in the areas of text translation, audio to text conversion etc. As in the traditional RNNs, the straight structures are also present in LSTM, but there is a difference in the building manner of the reiterating modules. The main method by which LSTM preserves the past info is by line running on the top of LSTM network which is called as cell states. All of the modules in the network consist of a cell state. These cell states are fed information with the help of different gates. Fig. 3 shows four contacting layers of our LSTM model.

These gates are composed up of sigmoid function -whose value varies between 0 and 1- so it can be decided how much information is to be passed to the next layer. If the value of the sigmoid function is 1, it means the whole of the information is passed to the next cell else if it is 0 then no information is passed. Hence the cell states help the network to maintain the info in the system.

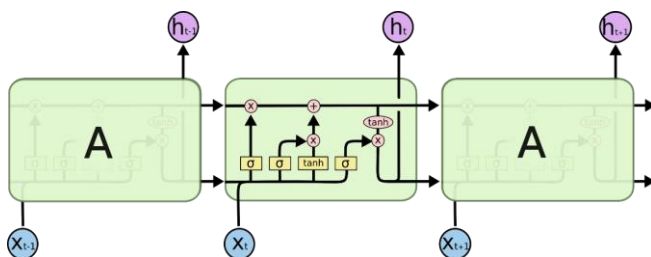


Fig. 3. Four contacting layers of LSTM

Bidirectional LSTM: Bidirectional LSTM[12] are an addendum to the conventional LSTMs and can help in significantly enhancing the performance of the model problems related to sequence classification. A Bidirectional LSTM, or bi-LSTM, is a model for sequence processing that consists of 2 LSTMs: one taking the input in a forward direction, and the other in a backwards direction. Bidirectional LSTMs work upon 2 LSTMs in place of one on the sequence provided as input. Fig. 4 shows the architecture of our bi-LSTM network. The first LSTM trains itself on the input sequence as-is and the second LSTM works upon the reversed copy of the input sequence. By

using the bi-LSTMs the amount of information available to the network is increased effectively, which helps in enhancing the context available to the algorithm and thus result in complete and faster learning of the model.

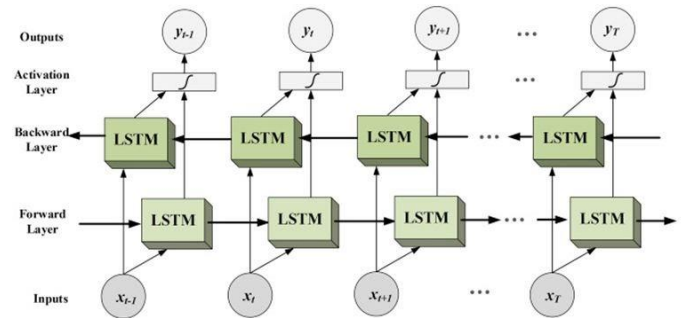


Fig. 4. Bi-Directional LSTM network

C. Datasets Collection

We have used the Flickr 8k[13] dataset for training and validation purposes. This dataset has been provided by University of Illinois at Arban - Champaign. The dataset contains 8000 images and for each image it has corresponding 5 descriptions. By default, the dataset is split into two folders, image folder and text folder. For each image the caption is stored along with the respective id as we have a distinctive image-id for every image in the set. The images in the dataset are divided into three parts: Training set, validation set and Test set. Test and validation set consist of 1000 images each whereas the training set consists of 6000 images.

Apart from this there are other datasets also available like MS-COCO[14] and Flickr30k[15] for captioning images but both these datasets have at least 30,000 images and training the model on these datasets requires a lot of power and is computationally very expensive.



Fig. 5. A random image from dataset along with the following captions.

1. black dog and spotted dog are fighting
2. black dog and tri coloured dog playing on the road
3. two dogs of different breeds looking at each other on the road
4. two dogs on pavement moving toward each other
5. black dog and white dog with brown spots are staring at each other in the street

D. Data Preprocessing

Flickr 8k dataset consists of nearly 6000 train images and for each image we have corresponding 5 descriptions. These text descriptions require some minimal pre-processing before we can use it to train the model.

We first loaded the file containing all the descriptions along with their corresponding image id. We looped through the file and created a dictionary which maps each photo identifier to a list containing textual descriptions for the image. After this we did some cleaning of the textual data in order to reduce vocabulary size. Cleaning of textual descriptions involve: removing punctuations, converting text to lowercase, removing stop words like 'a', 'an' etc. and removing tokens containing digits.

Next step is to create a vocabulary of all the unique words present across all the image descriptions. Finally, for each description which corresponds to an image in training dataset we need to add a '<startseq>' token at the start of each caption and an '<endseq>' token at the end of each caption. The '<startseq>' token signifies the start of a sequence while '<endseq>' token signifies end of a sequence.

E. Feature Extraction

In our research work, image acts as an input to the decoder network. For training the decoder, the image data must be provided in the form of fixed size vectors. Therefore, each image is converted into a fixed size vector which will then be fed as input to RNN.

We use a transfer learning method for extracting features from the images. For this purpose, we used pre-trained models and its weight trained on larger similar data. We computed the image features using these pre-trained models and saved them in a file. Later we loaded these features and fed them into the neural network as the interpretation of the image given in the dataset.

F. Model Training and Evaluation

For training purposes, we used the Google colab notebook. We trained the decoder model on a batch size of 32 and 64 using Adam optimizer and categorical crossentropy as loss function. We used training and validation loss as the metric to evaluate the model after each epoch. We monitored the validation loss of the model during training. When the validation loss of the model improves at the end of an epoch, we saved the model into a file.

At the end of the training period, we used the model with best skill on the training dataset as our final model. The final code for our research work is available at [16].

G. Performace Measures

We have used two algorithms to generate the captions from the features extracted using CNN encoders.

Greedy Search Algorithm using Argmax function: Greedy Search algorithm chooses one best candidate at each step while generating caption. It selects the word with the highest probability by applying the argmax function to the vocabulary of words and selecting the word with the highest probability to generate captions of image. Choosing one best candidate may be optimal in beginning but for complete sentences, it may not be the best choice.

Beam Search algorithm: Beam search [17], [18] algorithm is a greedy tree search algorithm based on heuristics. The advantage over greedy search algorithm is that it selects multiple alternatives at each step instead of one. It selects the top k words with the highest probability from vocabulary of words, where k = beam width. The procedure for beam search is as followed:

1. Select the first k words with the highest probability from the vocabulary of words by applying SoftMax function.
2. For each word selected in the first step, find the conditional probability of the next word given that the previous pair of words occurred.
3. Repeat the process iteratively until the end of sentence. In simple words, at each step we consider the possibility of a pair of words occurring together instead of just focusing on a single word each time while generating a caption.

The number of alternatives selected at each step can be changed with beam width parameter, k. For example, if k=3, three alternatives are selected at each step of beam search.

BLEU Score:

After generating captions from extracted features, the next step is to compare the accuracy of our generated captions with the actual captions given in the dataset. We have used BLEU [19] score metrics as a parameter to measure the accuracy of our generated captions. BLEU score helps in analyzing the text quality which has been generated by the Machine Learning model. BLEU score was among the earliest developed metrics to get such high correlation with actual verdict.

The value of BLEU score always lies between 0-1. If BLEU score is zero, it means machine translation is not relevant to actual description at all. On the other hand, a BLEU score of 1 means that the machine translation is equivalent to actual description.

For calculating BLEU score we followed the following procedure:

- 1) Produced captions by taking all images which belong to the test set.
- 2) After that we used these captions generated by model as our predicted or candidate sentences.
- 3) Next each of the candidate sentences is associated with 5 of the reference sentences which are given by humans.
- 4) The BLEU score of candidate sentences related to each of the references is averaged.

IV. RESULT

The BLEU scores of different models are shown in table I and table III respectively. The y-axis shows the models along with their configurations while the x-axis shows the BLEU scores using the greedy algorithm and beam-search algorithm. From table I, we can infer that given a defined set of configurations (batch size = 64 and optimizer = adam and decoder = unidirectional LSTM), Inception V3 (BLEU-I score = 0.605097) performs better than Vgg16 model (BLEU-I score = 0.578993). From table III, we can infer that given a defined set of configurations, a bidirectional LSTM decoder outperforms a unidirectional LSTM decoder for both Inception V3 and VGG16 encoders. One can also see that

beam search algorithm for generating captions is better than greedy algorithm although the time required for beam search is more. The Inception V3 + Unidirectional LSTM model gives a BLEU-1 score of 0.5695 with batch size =32. But on increasing batch size to 64, the BLEU score improved to 0.605097. This took roughly 8GB of ram. We could not increase batch size to 64 with bidirectional LSTM models as that required more than 12GB of ram which is more than what is available with us. Table II shows an example of caption predicted on an image taken randomly from internet.

TABLE I. BLEU SCORES KEEPING BATCH SIZE = 64

Model and Config	Argmax (Greedy)	BEAM Search
InceptionV3 + Unidirectional LSTM Epochs = 11 Batch Size = 64 Optimizer = Adam	Cross-entropy loss (Lower the better) loss(train_loss): 2.5254 val_loss: 3.1769 BLEU Scores on Validation data (Higher the better) BLEU-1: 0.591272 BLEU-2: 0.340125 BLEU-3: 0.236282 BLEU-4: 0.105637	k = 3 (beam width) BLEU Scores on Validation data (Higher the better) BLEU-1: 0.601173 BLEU-2: 0.349092 BLEU-3: 0.248659 BLEU-4: 0.119507
VGG16 + Unidirectional LSTM Epochs = 7 Batch Size = 64 Optimizer = Adam	Cross-entropy loss (Lower the better) loss(train_loss): 2.6297 val_loss: 3.3486 BLEU Scores on Validation data (Higher the better) BLEU-1: 0.557626 BLEU-2: 0.317652 BLEU-3: 0.216636 BLEU-4: 0.105288	k = 3 (beam width) BLEU Scores on Validation data (Higher the better) BLEU-1: 0.578993 BLEU-2: 0.326569 BLEU-3: 0.226629 BLEU-4: 0.113102

TABLE II. AN EXAMPLE OF AN IMAGE TAKEN RANDOMLY FROM THE INTERNET.

	Predicted Caption: dog is running through the water
---	--

V. CONCLUSION

In this paper, we have used Flickr 8k dataset with various image captioning models to compare the performance of different models. We have used CNN encoders like VGG16, InceptionV3 etc. for converting features into numeric vectors. These features are then passed to unidirectional or a bidirectional LSTM for generating captions. We used the

TABLE III. BLEU SCORES KEEPING BATCH SIZE = 32

Model and Config	Argmax (Greedy)	BEAM Search
InceptionV3 + Unidirectional LSTM Epochs = 11 Batch Size = 32 Optimizer = Adam	Cross-entropy loss (Lower the better) loss(train_loss): 2.5254 val_loss: 3.1769 BLEU Scores on Validation data (Higher the better) BLEU-1: 0.564183 BLEU-2: 0.314968 BLEU-3: 0.210921 BLEU-4: 0.098583	k = 3 (beam width) BLEU Scores on Validation data (Higher the better) BLEU-1: 0.569564 BLEU-2: 0.315819 BLEU-3: 0.219372 BLEU-4: 0.111061
InceptionV3 + Bidirectional LSTM Epochs = 20 Batch Size = 32 Optimizer = Adam	Cross Entropy loss (Lower the better) loss(train_loss): 2.4200 val_loss: 3.0724 BLEU Scores on Validation data (Higher the better) BLEU-1: 0.575166 BLEU-2: 0.332099 BLEU-3: 0.228444 BLEU-4: 0.111307	k = 3 (beam width) BLEU Scores on Validation data (Higher the better) BLEU-1: 0.581609 BLEU-2: 0.339489 BLEU-3: 0.240200 BLEU-4: 0.124673
VGG16 + Unidirectional LSTM Epochs = 7 Batch Size = 32 Optimizer = Adam	Cross-entropy loss (Lower the better) loss(train_loss): 2.6297 val_loss: 3.3486 BLEU Scores on Validation data (Higher the better) BLEU-1: 0.560285 BLEU-2: 0.308491 BLEU-3: 0.210819 BLEU-4: 0.105209	k = 3 (beam width) BLEU Scores on Validation data (Higher the better) BLEU-1: 0.566529 BLEU-2: 0.315291 BLEU-3: 0.212491 BLEU-4: 0.103105
VGG16 + Bidirectional LSTM Epochs = 18 Batch Size = 32 Optimizer = Adam	Cross Entropy loss (Lower the better) loss(train_loss): 2.2342 val_loss: 3.1726 BLEU Scores on Validation data (Higher the better) BLEU-1: 0.568254 BLEU-2: 0.312748 BLEU-3: 0.218816 BLEU-4: 0.112289	k = 3 (beam width) BLEU Scores on Validation data (Higher the better) BLEU-1: 0.579914 BLEU-2: 0.323926 BLEU-3: 0.227842 BLEU-4: 0.113637

BLEU score metric for comparing the accuracy of different image captioning models. To conclude we can say that for all types of Convolutional networks (encoders) the Bidirectional LSTM gave better results than the unidirectional LSTM. Also, for same type of decoder i.e. LSTM or BiLSTM, inceptionV3 encoder model performed better than the VGG16 model. Each type of method has its own merits and limitations like we have seen a BiLSTM performs better than a unidirectional LSTM as a unidirectional LSTM runs an input in only one direction so it preserves context only from the past whereas a BiLSTM runs input in both directions, once in a forward direction and once in a backward direction such that it preserves information from both past and future which helps in understanding the context better. At the same time, hidden layers in BiLSTM are more complex as compared to LSTM

and require huge computational power. Also, BiLSTM cannot be used for purposes like speech translation where you can't wait for whole input before beginning the inference.

VI. LIMITATIONS AND FUTURE WORK

Although experimentation with given models, datasets and hyperparameters show pretty good results but there are certain limitations to the proposed work like we did not have machines with higher processing power. Higher computational powers would have enabled us to further fine-tune the hyperparameters like batch size and learning rates which we believe would have resulted in better performance. Also, the dataset we have used contains only 8000 images. Using larger datasets like Flickr30k, MS-COCO etc. would mean we have more images to train the model on and it will ultimately lead to better accuracy. Also, larger datasets would mean we have larger vocabulary of words to train the model which would lead to better and more grammatically correct captions. But for working with these larger datasets, we would require machines with high computational powers otherwise it will take a lot of time to train the model on these datasets. The work we have done in this paper is just a small part of a large research area, there is lot of research which can be done in this field. For future prospects we suggest following improvements:

1. Using larger datasets: We can make use of larger datasets like MS-COCO, Flickr30k or Stock 3M datasets which will increase the vocabulary size thus enhancing the model accuracy significantly. It will help to generate better and diverse captions for an image.
2. Hyperparameter Tuning: The hyperparameters related to the model can be further fine-tuned to improve the accuracy score of the model.
3. Implementing Attention based Model: Nowadays attention mechanism is becoming quite popular. In future prospects, we can make use of attention-based mechanism which can easily focus on different parts of the image while output sequence is being produced.
4. Apart from this, newer models like inception-v4[20] or inception-resnet[20] can be used to improve the BLEU score. Also, we can make use of other RNNs like Gated Recurrent Unit[21] to have more detailed comparison of different models.

ACKNOWLEDGMENT

We would like to express our gratitude towards our research guide Dr. Akshi Kumar (Assistant professor, DTU) for her constant guidance, supervision and constructive criticism during the successful completion of the research work.

REFERENCES

- [1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Dec. 2016, vol. 2016-December, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [2] X. Chen and C. Zitnick, "Learning a Recurrent Visual Representation for Image Caption Generation," 2014.
- [3] D. V. T. and V. R., "RETRIEVAL OF COMPLEX IMAGES USING VISUAL SALIENCY GUIDED COGNITIVE CLASSIFICATION," J. Innov. Image Process., vol. 2, no. 2, pp. 102–109, Jun. 2020, doi: 10.36548/jiip.2020.2.005.
- [4] Y. Bengio and Y. Lecun, "Convolutional Networks for Images, Speech, and Time-Series," 1997.
- [5] D. E. Rumelhart and J. L. McClelland, "Learning Internal Representations by Error Propagation," in Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations, 1987, pp. 318–362.
- [6] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, vol. 07-12-June-2015, doi: 10.1109/CVPR.2015.7298968.
- [7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, vol. 07-12-June-2015, doi: 10.1109/CVPR.2015.7298935.
- [8] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, vol. 07-12-June-2015, doi: 10.1109/CVPR.2015.7299087.
- [9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.
- [10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, vol. 2016-December, doi: 10.1109/CVPR.2016.308.
- [11] S. Hochreiter and J. J. Uger Schmidhuber, "Long short term memory. Neural computation," Mem. Neural Comput., vol. 9, no. 8, 1997.
- [12] M. Basaldella, E. Antolli, G. Serra, and C. Tasso, "Bidirectional LSTM Recurrent Neural Network for Keyphrase Extraction," 2018, pp. 180–187.
- [13] "Flickr 8k Data | Illinois." <https://forms.illinois.edu/sec/1713398> (accessed Mar. 05, 2021).
- [14] "COCO - Common Objects in Context." <https://cocodataset.org/#download> (accessed Mar. 05, 2021).
- [15] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models," in 2015 IEEE International Conference on Computer Vision (ICCV), Dec. 2015, pp. 2641–2649, doi: 10.1109/ICCV.2015.303.
- [16] "Image Caption - Google Drive." https://drive.google.com/drive/u/2/folders/181xqs33zg5-PIv_VReG8VE13Fi6rbPIi (accessed Mar. 16, 2021).
- [17] "9.8. Beam Search — Dive into Deep Learning 0.16.1 documentation." https://d2l.ai/chapter_recurrent-modern/beam-search.html (accessed Mar. 13, 2021).
- [18] C. Meister, T. Vieira, and R. Cotterell, "Best-First Beam Search," arXiv. 2020, doi: 10.1162/tac1_a_00346.
- [19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," ACL, pp. 311–318, 2001, doi: 10.3115/1073083.1073135.
- [20] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," 2017.
- [21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," 2014.



Comparative study of wind induced mutual interference effects on square and fish-plan shape tall buildings

SUPRIYA PAL, RITU RAJ* and S ANBUKUMAR

Department of Civil Engineering, Delhi Technological University, New Delhi 110 042, India
e-mail: supriya8788@gmail.com; rituraj@dtu.ac.in; sanbukumar@dce.ac.in

MS received 13 September 2020; revised 14 January 2021; accepted 10 February 2021

Abstract. This paper deals with comparative study of wind-induced mutual interference effects on twin Square and Fish- plan shape building models having equal volume. Wind induced pressure and force is measured through experiments under boundary layer wind tunnel for interference and isolated wind incidence condition with building model of scale 1:300. For isolated condition's study two wind incidences i.e. 0° and 180° is considered whereas; for interference conditions five different orientation of twin interfering building models are considered. The distance between the twin building models is fixed at 10% of the height of principal model. The variation of coefficient of wind pressure on different surfaces of the model is shown by contour plots. To examine the mean variation along the faces Face-Average C_p has been calculated, plotted and discussed to a great extent. The interference study is done in order to understand the effects of different conditions that can arise in real life situations and the differences associated with them. The concluding remarks states the dominance of Drag and Lift forces at isolated 0° and 180° wind incidences for Fish- plan shape model and at isolated 0° wind direction at Square shape model. Also, the overall efficiency in terms of Base shear of principal building is enhanced due to interference effect; with maximum efficiency exhibited by Back-to-Back wind interference condition when only Fish- plan shape model is considered. Overall, maximum efficiency in terms of induced wind pressure and base shear is exhibited by Square- plan shape model at Full Blockage condition.

Keywords. Tall building; pressure coefficients; minimum, maximum and average face C_p ; average interference factor; base shear; force interference factor.

1. Introduction

Tall buildings with unconventional configurations have been constructed nowadays to satisfy today's needs for providing aesthetically sound shelters to humans due to population growth. The buildings not only provide shelter but also provide a barrier between the outdoor and indoor environments and the inhabitants residing inside it [1]. Most of the manmade structures are bluff bodies and are subjected to formation of large eddies in its wake [2]. Tall buildings are more susceptible to wind excitations and are more sensitive to dynamic loading [3] for which shape optimization of the cross section is put forward to improve its wind resistance [4]. Wind-induced lateral movement of tall structures may cause annoyance to the occupants (especially in the upper floors) and can also succour structural damage [5]. In addition, it is further seen that, the effects of wind loads are dynamic in nature and can cause collapse of the structure [6]. In designing of tall and slender structures wind loads play fundamental role and thus need for accurate evaluation of such with respect to both collapse

and serviceability conditions [7]. The tall buildings should be designed for the extremes taking into account of past time experiences with dust storms and thunderstorms of these gusts of wind can cause major damage to the structures as compared to local winds [8].

Due to increase in the number of tall and slender structures all the dynamic aspects of wind like flutter, vortex shedding, galloping and unaccountable behavior of wind needs to be studied in detail [9]. Ample amount information is available in different international standards [10–14] for isolated wind incidence conditions only and that too only for regular square, rectangular, cylindrical, etc. plan shape buildings. Due to mutual interference, the wind loads on tall structures may increase or decrease depending upon the parameters such as upstream terrain, shape and size of the buildings, the incident wind directions and the building arrangement as well as spacing. However, because of the complexity of problem with wide range of variables in interference conditions no information can be found in any international codes. The consequences of under specification can be serious and hence, this necessitates more experimental and/or analytical study. Furthermore, the results presented in this study can help structural designers

*For correspondence
Published online: 28 April 2021

to choose among ingenious solutions with an objective to fulfil collapse and serviceability requirements of a structure under extreme wind actions.

It is apparent that wind loads on tall buildings can be reduced by using two approaches; one is use of “Aerodynamic Mitigation” technique and second is use of “Aerodynamic shape optimization” technique [15]. The present study emphasizes on comparative experimental investigation between Square (figure 4(a)) and Fish- plan shape (figure 4(b)) tall building model by considering “Aerodynamic shape optimization” approach. For good comparative study between both the models, same volume is regarded. The purpose of the study is to present pressure distribution on various faces, to depict base forces and also to delineate pressure and force interference factors of Square and Fish- plan shape tall building models for different wind flow at isolated and interference conditions because these results are not incorporated in relevant codes and no analytical formula is available for evaluation of wind effects on such complex plan shaped tall buildings. Square- plan shape model is studied also to validate the experimental method according to various international codes. The functionality of Fish- plan shape structures can be as hotels, museums, institutional buildings, office, hospitals, educational spaces and other public buildings. Isolated Fish- plan shape buildings are adapted triangular- plan shape buildings that are often build at triangular plots on the face of merging roads; the best example is the New York Flatiron (see figure 1).

From the available literature it is evident that most of the previous interference studies are concentrated on finding optimum building configurations and/or distance between the interfering models of square or rectangular plan shape buildings such as, Lam, Zhao and Leung [16] have studied the interference effects on a row of five tall square plan shaped buildings, Amin and Ahuja [17] has calculated the mean interference effects between two rectangular located in close proximity in a configuration of ‘L’ and ‘T’ plan shape buildings, Hui, Tamura and Yoshida [18] have studied peak interference effect of a square plan shape model on a rectangular plan shape model and vice versa and found huge difference in the values for isolated and interference studies. Similarly, Bairagi and Dalui [19] have studied the interference effect between twin rectangular models with varying distance between the two to find out optimum spacing between the twin buildings for 0° and 90° wind incidence conditions. In addition, some researches [20–23] have studied the interference effects among twin square plan shaped models for various distances between the twin interfering buildings. It is evident that no experimental study has been carried out on complex plan shaped tall buildings thus; this accounts for interference study of complex plan shape building with same or different plan shape interfering buildings.



Figure 1. The Flatiron Building, New York city.

Focus of present study is concentrated at experimental investigation for understanding the mutual interference between twin Square and Fish- plan shape Building models. Depending upon various arrangements among twin Fish- plan shape building models four interference conditions (figure 13) i.e., Back-to-Back, Front-to-Front, Front-to-Back and Back-to-Front is taken for the study. As Square- plan shape model is symmetrical about both the axis thus only Full Blockage interference condition (figure 12) is taken into account. Isolated wind incidence conditions at 0° and 180° is also studied as the direction of wind flow in all the interference conditions are either of these wind directions depending upon the orientation principal building model to incident wind.

An approach is made to find out a generalized relation between relative height and C_p variation along vertical centerline at each face for all isolated and interference conditions. Also, for cladding structures surface design determination of position of high pressure and high suction regions is significant as it causes high external compression and high external tension on cladding surface respectively, which further leads to failure of cladding.

2. Experimental programme

2.1 Feature of experimental set-up

The experiments have been conducted in open circuit boundary layer wind tunnel with a section of 2 m x 2 m and 15 m in length. In order to generate uniform flow of wind throughout the wind tunnel, square holed honey-comb is positioned at the entrance of the wind tunnel. Vortex generators and obstructions are placed at the upstream of wind tunnel for developing boundary layer flow conditions. A pictorial representation of wind tunnel facilities is shown in figure 2. The wind is continuously flowing through the tunnel with the help of suction by blower fan, which is producing a constant mean wind velocity of 10 m/sec during experiment. Pressure model is placed at the centre of manually controlled turntable, which rotates the model at various angles. The pressure tapings of 1 mm diameter made up of steel tubes are installed near the edges of each face to study the changes in the variation of pressure due to flow separation. These tapings points and reference pressure points are attached to the pressure transducers for measuring pressure through the Baron instrument attached. Wind pressure on the models was measured using Baratron Pressure Transducer, which was capable of measuring extremely low differential heads. The wind velocity inside the wind tunnel was measured with the help of the instrument “TESTO-480”. A probe was connected to this instrument to measure the wind velocity at different height, which had a length of 1 m. The intensity of the turbulence is defined as the ratio of the standard deviation of fluctuating wind velocity to mean wind velocity.

The variation of mean velocity of wind and turbulence intensity of the same is shown in figure 3. The boundary

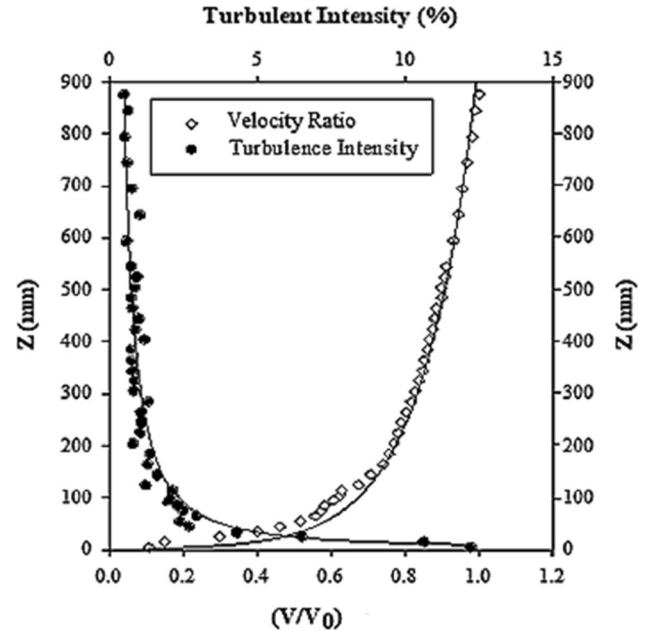


Figure 3. Mean wind velocity and turbulence intensity profile.

layer wind profile is governed by the power law equation (Eq. 1):

$$V = V_0 \left(\frac{z}{z_0} \right)^\alpha \quad (1)$$

Where, V is the mean velocity at height z above the ground, V_0 is the wind velocity at reference height, z_0 is the reference height above the ground, i.e. 900mm for the present experimental work under wind tunnel, and α is the

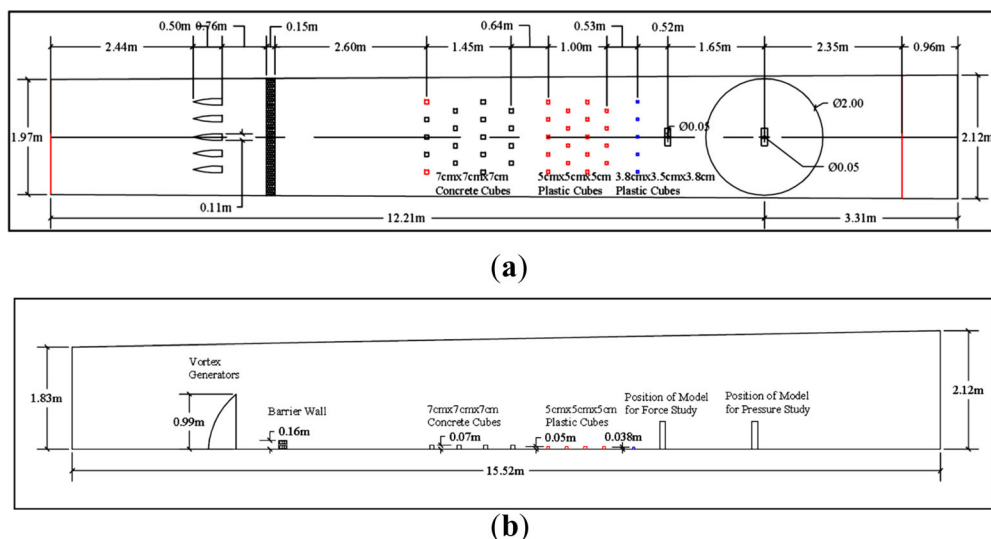


Figure 2. Pictorial representation for arrangement of facilities under wind tunnel at (a) Plan view and (b) Elevation view.

are distributed at seven different height levels at 10 mm, 60 mm, 180 mm, 300 mm, 420 mm, 540 mm and 590 mm from bottom to acquire a wide and clear picture of the distribution of pressure on all faces and sides of the models.

2.3 Validation with international codes

For validation, experimental study was carried out at the wind tunnel for a Square- plan shape isolated building model of 600 mm height and 40000 mm² plan area under present working environment. The experimental study has been validated with different international codes [10–14]. It has been observed that pressure coefficients of windward, leeward and sidewalls of isolating model have appreciable results with the international codes as shown in table 1.

appreciable results with the international codes as shown in table 1.

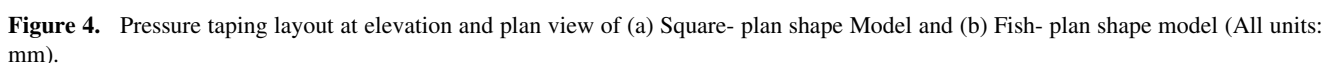


Table 1. Comparison of face pressure coefficient (C_p) on the Square- plan shape tall building.

International Code	Wind angle	Windward side	Leeward side	Side walls
Experimental results	0°	0.71	−0.41	−0.67
	90°	0.73	−0.42	−0.66
AS/NZS: 1170.2:2002	0°	0.80	−0.50	−0.65
	90°	0.80	−0.50	−0.65
ASCE/SEI 7-10	0°	0.80	−0.50	−0.70
	90°	0.80	−0.50	−0.70
EN: 1991-1-4	0°	0.80	−0.55	−0.80
	90°	0.80	−0.55	−0.80
BS: 6399-2	0°	0.76	−0.50	−0.80
	90°	0.76	−0.50	−0.80
IS 875 (part 3)	0°	0.80	−0.25	−0.80
	90°	0.80	−0.25	−0.80

3. Results and discussions

A thorough study of variation of mean pressure and base shear (designated as Drag force coefficient and Lift force coefficient) at various wind incidence conditions is carried out for Square and Fish- plan shape tall building model. This is a very important step to engineer an understanding of the variation of cross section plan of building on associated wind pressure on various faces and Base shear on unit of model before designing the buildings for collapse and serviceability conditions. For each pressure tapping point the pressure coefficient is calculated from the formula Eq. (2) [14]:

$$C_p = \frac{P_a}{0.6V^2} \quad (2)$$

where, P_a is pressure at respective pressure tapping point and V is the mean wind velocity in m/s at the top of building model i.e., 10 m/s for this experiment.

3.1 Distribution of minimum, maximum and average C_p along building periphery at isolated model conditions for

3.1.1 Isolated Square- plan shape building model at 0° wind incidence The wind incidence directions for isolated conditions of a Square- plan shape building model is given in figure 5(a). Detailed experimental study for isolated condition of Square- plan shape model has been carried out at 0° direction of wind incidence only. As the model is symmetrical at both the axis thus similar pressure distribution is also accompanied at 180° wind direction. Pressure coefficients are calculated with the help of Eq. (2) and distribution of minimum, maximum and average C_p at face around building facade for the isolated condition at Square- plan shape model is plotted and shown at figure 6.

Due to suction cladding, materials may oust away during an episode with relatively strong winds. The suction pressure is generated due to flow separation at faces. In suction regions windows panes break and the broken shards end up dispersed outside the building. From structural design, point of view average of C_p at face values may suffice the condition but for cladding surface design maximum and minimum C_p at face magnitudes put huge difference to actual conditions. Larger variation between maximum C_p at a face and minimum C_p at a face shows huge turbulence at the face however, this difference does not indicate fluctuation of wind at the face and thus a thorough study of pressure coefficient distribution is must. The nearness of average face C_p magnitudes i.e. Face values to minimum or maximum C_p at face values indicates larger magnitude of C_p distribution over the face. From figure 6 it is evident that only Face-A is experiencing positive distribution of pressure due to direct exposure of face to incoming wind. Maximum C_p at face of 0.97 is experienced at Face-A which is about 38% higher than average pressure at Face-A whereas; maximum suction of -0.87 is experienced by side faces Face-B and Face-D which is again about 30% higher than average suction at side faces.

3.1.2 Isolated Fish- plan shape building model at 0° and 180° wind incidence The wind incidence directions for isolated conditions of a Fish- plan shape building model is given in figure 5(b). Detailed experimental study for isolated condition of Fish- plan shape model has been carried out at 0° and 180° directions of wind incidences.

Although the results of the other isolated and/or interference studies cannot be compared to the Fish- plan shape model due to the complexity of the model's shape but, from the literature by Sanyal and Dalui [24] it is clear that the distribution of C_p varies with the change on the direction of wind incidence. As the direction of incident wind is changing thus, the magnitude and nature of C_p values at

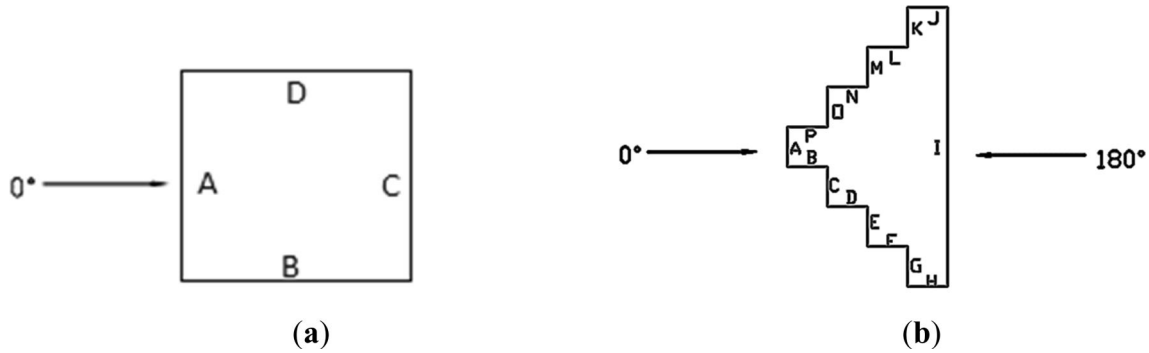


Figure 5. Wind Directions for Isolated building model (a) Square shape model and (b) Fish- plan shape model.

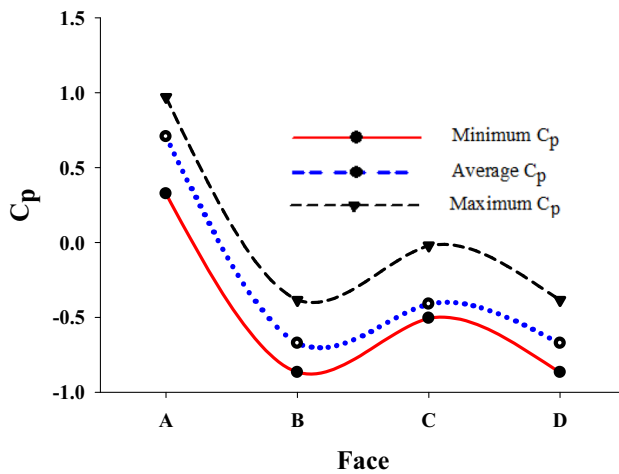


Figure 6. Minimum, maximum and average C_p distribution of Square- plan shape building model along periphery of building for 0° .

inner faces are also changing. Figure 7(a) shows the minimum, average and maximum C_p at face distribution for 0° isolated wind direction along Fish- plan shape building periphery. The overall maximum positive C_p at face of magnitude 0.74 is observed at Face-C which is closely followed by Face-A (0.71) because these faces are in direct exposure to incoming wind. It is observed that at all the faces perpendicular to the wind incident have higher magnitudes of C_p and also associate larger swirls as compared to all the parallel faces; this phenomenon is due to the exposure of perpendicular faces at this wind direction. The maximum suction of -0.58 is observed at Face-H and Face-J which also have very less variation between magnitudes of minimum, maximum and average C_p at face of present study. For pressure distribution at “+” plan shaped model [25] the windward Face-A at 0° wind incidence shows the maximum C_p at of 0.986 which is 39% higher than the maximum C_p at Face-A at present study, this variation is due to difference in the exposed surface area between both

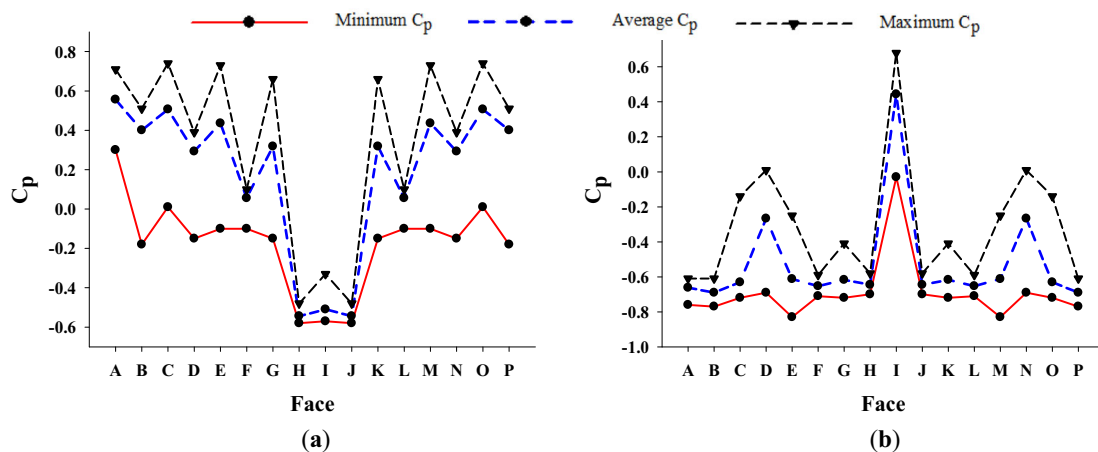


Figure 7. Minimum, maximum and average C_p distribution of Fish- plan shape building model along periphery of building for (a) 0° , (b) 180° isolated wind incidence conditions.

the model's face. The maximum and minimum C_p at face values is 0.56 and -0.56 is observed at Face-A and Face-I respectively for 0° condition. At 0° , overall face values show a decrement of the magnitude with maximum magnitude at Face-A and decreasing towards Face-I thereafter, it is increasing from Face-I to Face-P.

Figure 7(b) shows the minimum, average and maximum C_p at face distribution for 180° isolated wind direction along Fish- plan shape building periphery. Face-I at 180° acts as a shield to the downstream faces due to large elevated area of face thus, only Face-I of 180° condition experiences maximum C_p at face of 0.68 and positive face value of 0.39. As wind is being separated at wide angles from both the edges after being incident on Face-I of principal model thus, less fluctuation of magnitude of face values can be observed at the downstream faces of the principal models for all conditions. At 180° wind condition sudden hike in the magnitude of face value i.e. average value of C_p at face is observed between Face-C and Face-D (-0.63 to -0.27) and then a drop at Face-E (-0.27 to -0.61) shows reattachment of wind at Face-D and hence turbulence at face.

3.2 Distribution of pressure coefficient at faces of building model for

3.2.1 Isolated Square- plan shape building model at 0° wind incidence All the objects existing in nature are bluff bodies. Thus, it is important to study the pressure

distribution on all the surfaces due to incident wind of such bodies over a terrain (figure 8).

The present experiment is carried out at Power law index 0.22 with a mean wind velocity of 10m/s at the top of building model. In the present experiment at square model H/B (where, H = Height and B = width of model) ratio of 3.0 is considered (figure 9), whereas; in pressure distribution at figure 8(a) the study [26] deals model with H/B ratio as 8.0, Power law index of 0.22 and mean wind velocity at the top pf model was of 9.3m/s. Distribution of C_p at front, side and back faces have similar variation along the height with present experimental condition. However, due to variation in the mean wind velocity and mainly H/B ratio, the max C_p at front face is observed as 0.8 which is about 15% lower than that of Face-A at present study. Similarly, back face of [26] experiences higher suction of about 25% to that of Face-C of present study. Figure 8(b) shows distribution of C_p along CAARC (0.23) [27] building model where, H/B ratio is 6.1. The study shows huge difference in magnitude of C_p of back face to that of Face-C at present study mainly due to change in the experimental conditions.

3.2.2 Isolated Fish- plan shape building model at 0° and 180° wind incidence During an episode of wind incidence at such high velocity the edges of all the faces mostly escape the influence of incident wind and thus eddies are formed at such positions. Formation of eddy creates a space devoid of downstream fluid flow and thus, any exit can be provided at this portion of building. However, thorough study needs to be carried out for the effects of openings at pedestrian level due to high

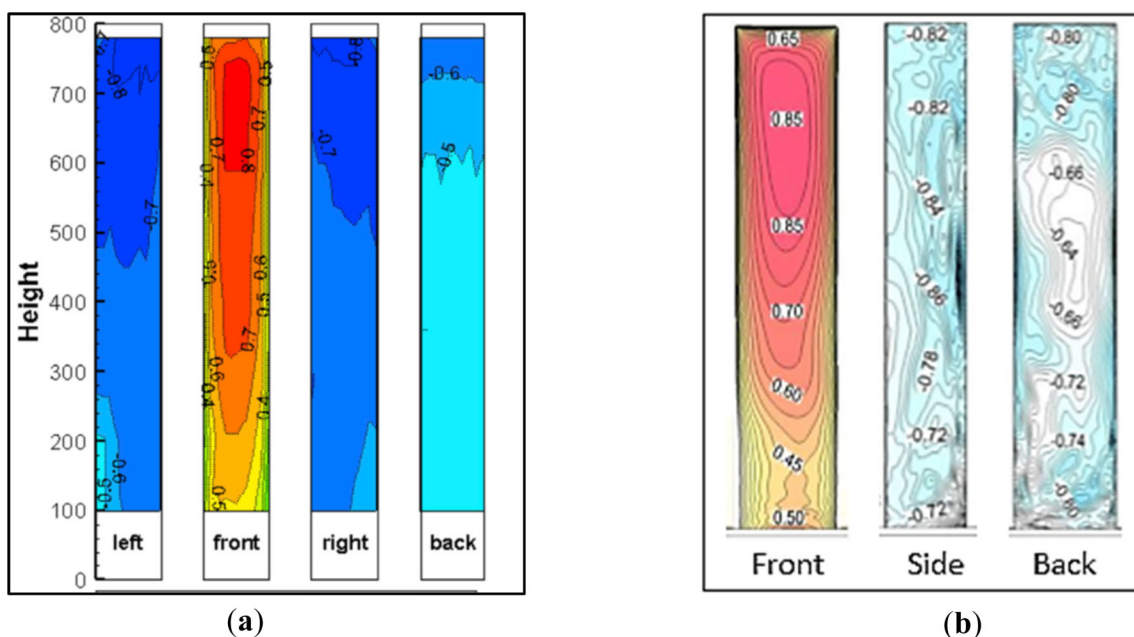


Figure 8. Mean wind pressure coefficient distributions on (a) Square model at wind direction of 0° [26], and (b) rectangular CAARC (0.23) at wind direction of 0° [27].

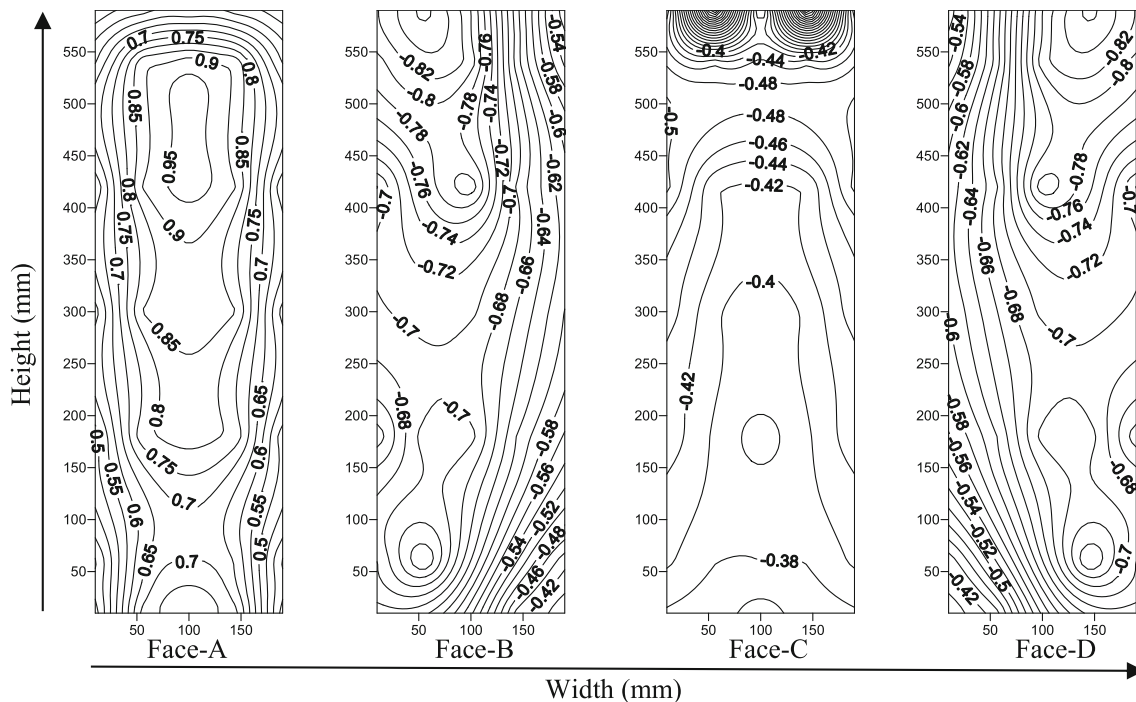


Figure 9. Pressure distribution of Square- plan Shape building model at 0° (isolated) wind incidence condition.

turbulence at ground level. Windward wind load is fundamentally blowing wind force that is pressing the building. The experimental study of pressure distribution on the walls of asymmetric trapped body under turbulent supersonic flow [28] concludes that the cladding pressure varies significantly with angle of attack which is also apparent in the present study from the variation of C_p values.

The opposite placed face pairs such as Face-B and Face-P, Face-C and Face-O etc. would have mirrored C_p distribution over the face in all conditions of wind incidence with or without the presence of interfering model for Fish-plan shape building models. Thus, for figures 10, 11, 17, 18, 19 and 20 distribution of pressure coefficient is shown for Face- A to Face- I only.

The distribution of pressure coefficients at Fish-plan shape model for 0° wind incidence is shown in figure 10. Very minimal variation is evident between 1/3rd to 2/3rd heights at Face-A due to uniform exposure to wind to face at such height. Two large symmetrical vortices are formed in the wake region at Face-I with its centre at 1/3rd height as the wind stream is getting deviated away from Face-I symmetrically from both sides. For 0° wind incidence rear Face-G at “Z” plan building model [29] and rear Face-I of present study shows agreement with the flow pattern due to similar positioning of faces. Also huge agreement is seen between the pressure distribution patterns of “+” plan shaped [25] model and present study model for front and side faces for 0° isolated wind incidence conditions

however, the magnitude differs due to the variation of shape and exposed elevated area between the models of both studies. Distribution of pressure coefficients of face pair Face-B- Face-C, Face-D-Face-E etc of present study is compared to Front and Right face of L- Shape model [30] at 0° wind direction due to similar orientation of face pairs to each other and to incident wind. In both the models formation of pressure region is evident at perpendicular face; wind is then reflected to parallel face. In L-shape model the perpendicular face have higher magnitude of C_p as compared to parallel face whereas; in Fish- plan shape model the perpendicular face have lesser magnitude of C_p as compared to parallel face, this difference in phenomenon is due to the presence of neighbour interfering faces in Fish-plan shape model. All parallel faces to wind direction in 0° wind incidence (Face-B, Face-D, Face-F, etc.) can be used as openings like balconies.

The distribution of pressure coefficients at Fish- plan shape model for 180° wind incidence is shown in figure 11. Face-I is handling the severity of wind due to direct exposure to incident wind. Recently, it has been suggested that the wind pressure coefficient at a point on a surface significantly varies with wind incidence angle and surface curvature [31]. At 180° wind direction, positive pressure occurs at Face-I and negative pressure at all the other faces of the model as the wind stream is getting deviated away from downstream faces after the wind is incident at Face-I. Due to large elevated area of windward Face-K at 0° wind incidence at “E” shaped model [32] and that of Face-I at

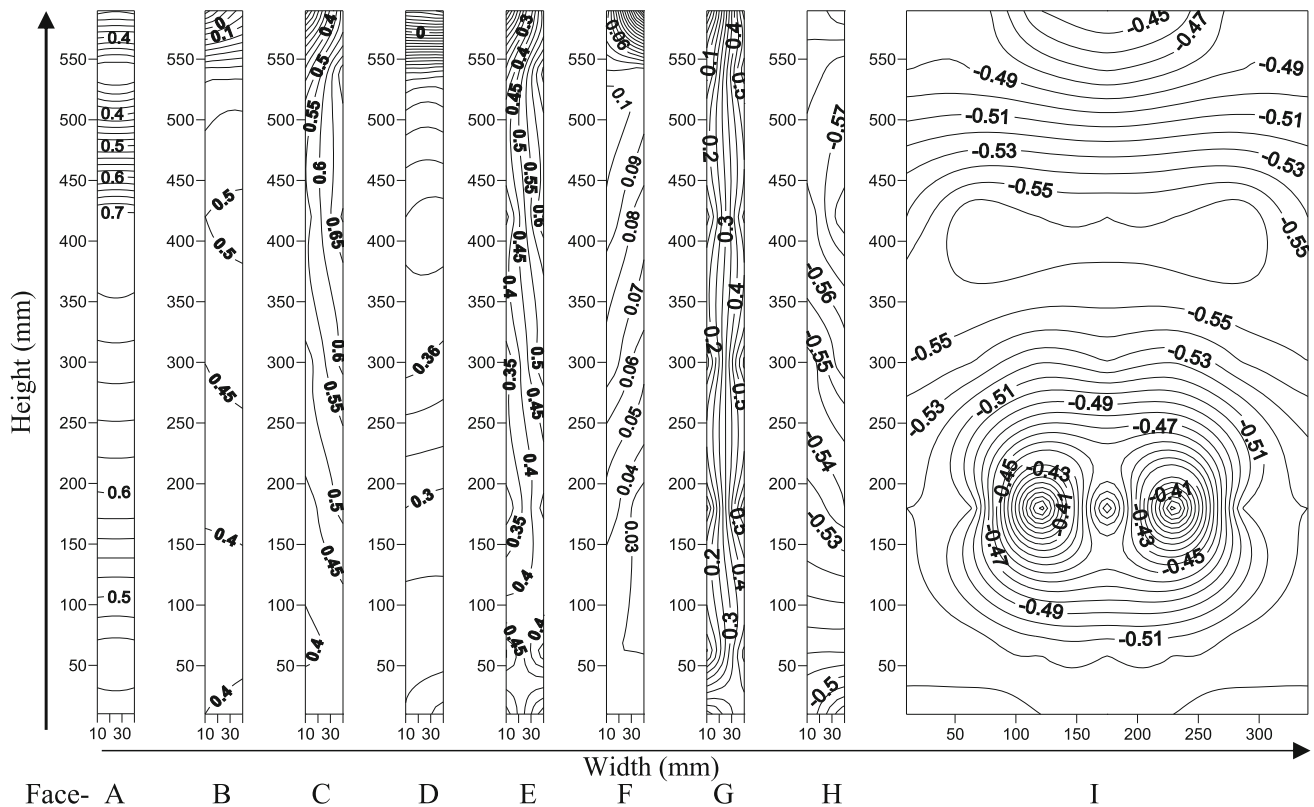


Figure 10. Pressure distribution of Fish- plan shape building model for selective faces at 0° (isolated) wind incidence condition.

180° wind incidence of present experimental study the flow pattern shows similarity. Where, for both the studies maximum C_p is found at $0.70 H$ to $0.75 H$ at vertical centerline of the face which is then further distributed at both the sides symmetrically. Face pair Face-G- Face-F of present study is compared to L-Shape model's [30] right and Front faces at 180° wind direction due to similar orientation of face pairs to each other and to incident wind. In present study, huge swirl of wind is noticeable on Face-F due to high turbulence of wind at Face-G and at neighbour interfering faces whereas; in L-shape model [30] due to variation in size and absence of interfering neighbour faces the phenomenon of swirl is absent. The highest positive C_p of 0.70 is found at Face-I which is about 6% lesser than the overall highest positive C_p of 0.74 (at Face-C and Face-O) as found at 0° wind incidence of present study, this is mainly because of the large difference in the exposed elevated area of face and orientation of model to the incident wind. Overall maximum suction of -0.83 is found at both Face-E and Face-M which is about 43% greater than that found at 0° wind incidence of present study due to higher angle of deflection of wind stream at 180° wind incidence. Face-C and Face-O have little fluctuation of pressure over the face as due to stagnation of wind incident at faces. Variation pattern of wind around the model is different in 180° as compared to

0° wind incidence. Decrease in pressure at windward faces causes less compression at cladding surface whereas, increase in suction leads to high tension at cladding surface.

Interference study

In interference studies blockage in wind tunnel is an important problem associated with wind tunnel tests. Due to interference of nearby buildings the pressure on principal building might increase or decrease depending upon many conditions like terrain category, exterior shape of building, cross-sectional plan of building, aspect ratio, etc.

The Fish- plan shape buildings are symmetrical about one axis thus for the present work only four orientation with 100% blockage conditions are considered with 10% gap between all the twin plan models model i.e. 60 mm as suggested by Cook [33], Houghton and Carruthers [34]. The limitation of the present study lies with the thorough study of tandem and staggered arrangement of twin Fish-plan shape models.

For Square- plan shape model, only Full Blockage interference is considered due to symmetrical shape of model (figure 12). However, for Fish- plan shape model interference study has been carried out for four interfering conditions; figure 13(a) Back to Back; (b) Back to Front; (c) Front-to-Back and (d) Front-to-Front.

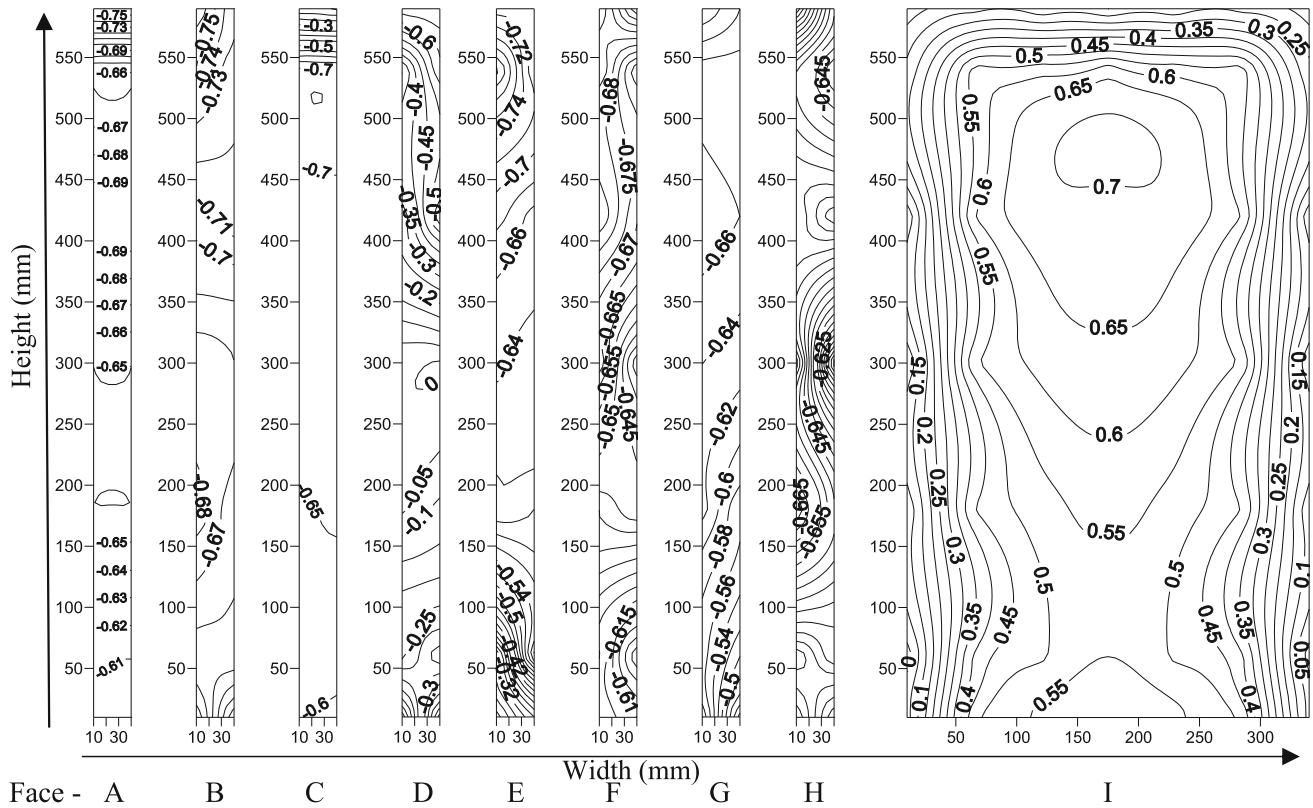


Figure 11. Pressure distribution of Fish- plan shape building model for selective faces at 180° (isolated) wind incidence condition.

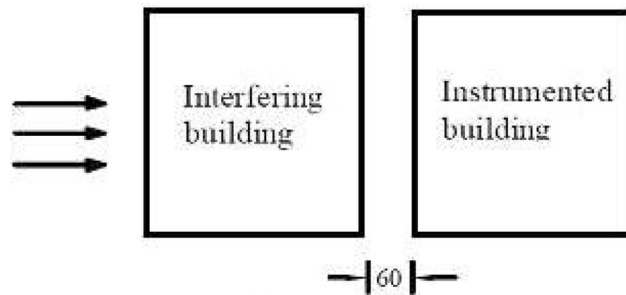


Figure 12. Full Blockage interference condition of Square- plan shape building model. (All units: mm).

3.3 Distribution of minimum, maximum and average C_p along building periphery at interference model conditions for

3.3.1 Full Blockage interference condition of Square model Mutual interference between twin Square- plan shape models results in the decrease of face values as compared to isolated condition of similar shape model. Due to interference between twin Square building models (figure 14) all the faces are experiencing suction as the instrumented model is under wake region of interfering model. About 54% decrease in maximum suction is experienced at faces due to influence of interference.

3.3.2 Interference conditions of Fish- plan shape model Figure 15 shows distribution of minimum, maximum and average C_p at face around building facade for all interference conditions. At Back-to-Back interference condition (figure 15(a)) largest variation between minimum and maximum C_p at face is observed at Face-I with which shows huge turbulence of wind at the face which is evident from figure 10. At Face-E 59% difference between maximum and minimum C_p at face values is explained through huge swirls due to reattachment of wind stream at the face. For Front-to-Back (figure 15(b)) interference condition the distribution of minimum, maximum and average C_p at face values indicates uniform flow at the principal building model due to the orientation of twin interfering models to each other. Irrespective of the orientation of interfering model the principal model at both Back-to-Front (figure 15(c)) and Front-to-Front (figure 15(d)) interference conditions the overall increase in magnitude of face values from Face-A to Face-I then decreasing from Face-I to Face-P is observed. Small rise in face vales for consequent faces shows swirl of wind whereas; huge rise denotes formation of pressure regions at face with higher magnitude of face value. Sudden decrease in face value between consequent faces shows formation of vortices. At figure 15(c) due to orientation of twin models to each other, large suction region is generated

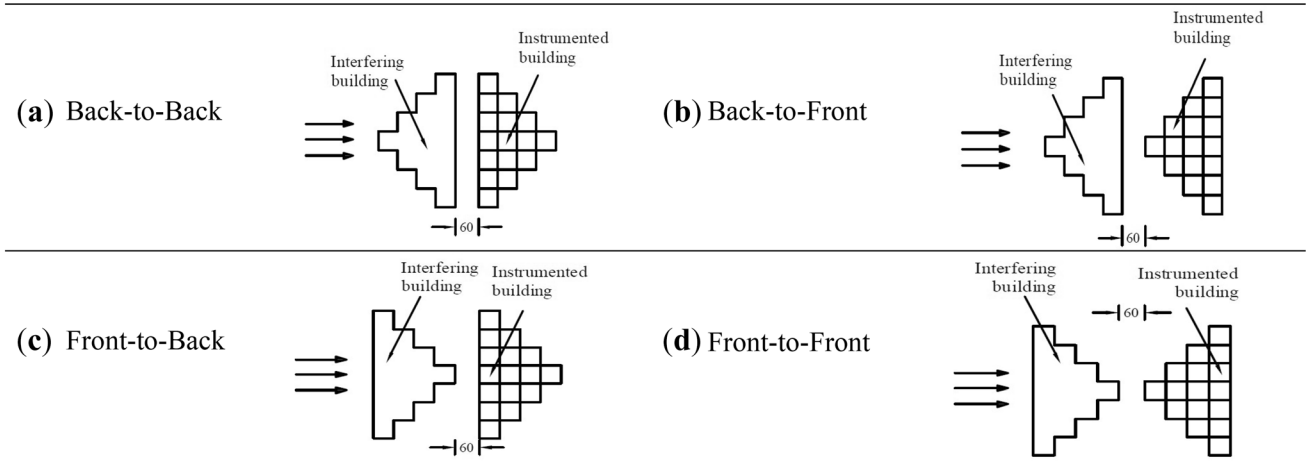


Figure 13. Various Interference Conditions of Fish- plan shape building model. (All units: mm).

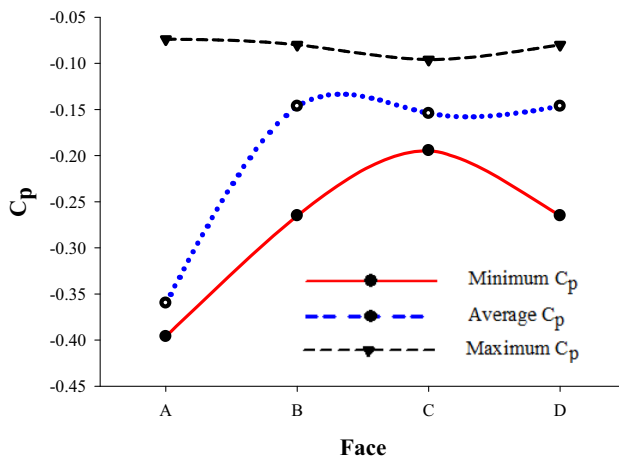


Figure 14. Minimum, maximum and average C_p distribution along periphery of Square- plan shape building model for Full Blockage Interference Condition.

between both the models and thus, largest variation (91%) between minimum and maximum C_p values is observed for Face-C. Face-I at Back-to-Front (figure 15(c)) and Front-to-Front (figure 15(d)) interfering conditions exhibits least turbulence due to least affected orientation of model's face to incoming wind.

3.4 Distribution of pressure coefficient at faces of

3.4.1 Square- plan shape building model at Full Blockage interference condition All the faces of Square- plan shape model at Full Blockage interference condition (figure 16) experiences suction as the principal building is under the wake region created by interfering model. Face-A experiences maximum suction of -0.38 because of the shielding effect of interfering building. The maximum suction at interference condition is about 54%

less than that at isolated condition. Face-B, Face-C and Face-D experiences suction throughout the height but due to unification of wind stream at side faces the suction is reduced to that of maximum suction at isolated condition. In the experimental study for tandem arrangement of models the C_p distribution at $x/b = 1.5$ for side faces show side wash from inner edge [21] which is also evident in the present interference condition however, due the magnitude of suction at present interference condition is very less (about 90%) as compared to buildings in tandem arrangement. Hui, Yoshida and Tamura [35] have studied the interference effect on two rectangular shaped models by keeping the models at parallel and perpendicular arrangement to each other, through study the author have cautioned for detailed study at the edges and corners as these parts are concentrated with huge pressure variations.

The interference study by Kim and Kanda [36], have associated their interference study with the change in the height of the interfering building model to the instrumented model and found that the highest suctions increase with increase in height of interfering building. Due to high complexity of the building plan of model in present study, it is impossible to correlate the results of pressure distributions to previous interference aerodynamic studies. Entrance to building can be provided at Face-A because of the limited cross sectional size of the face. The declaration is partly dependent on architectural requirements to provide safety and security, as the safety and security criteria include areas such as building control. Also, Face-A at ground level experiences lesser turbulence of wind at all wind incidences when compared to other faces at similar conditions. The adjoining edges between the wall pair like Face-B and Face-C and also Face-C and Face-D is least affected as eddy formation is taking place due to flow separation at such areas and thus openings can be provided near the attached edges. Depressed faces like Face-F, Face-L can be provided as emergency exits because at the time of

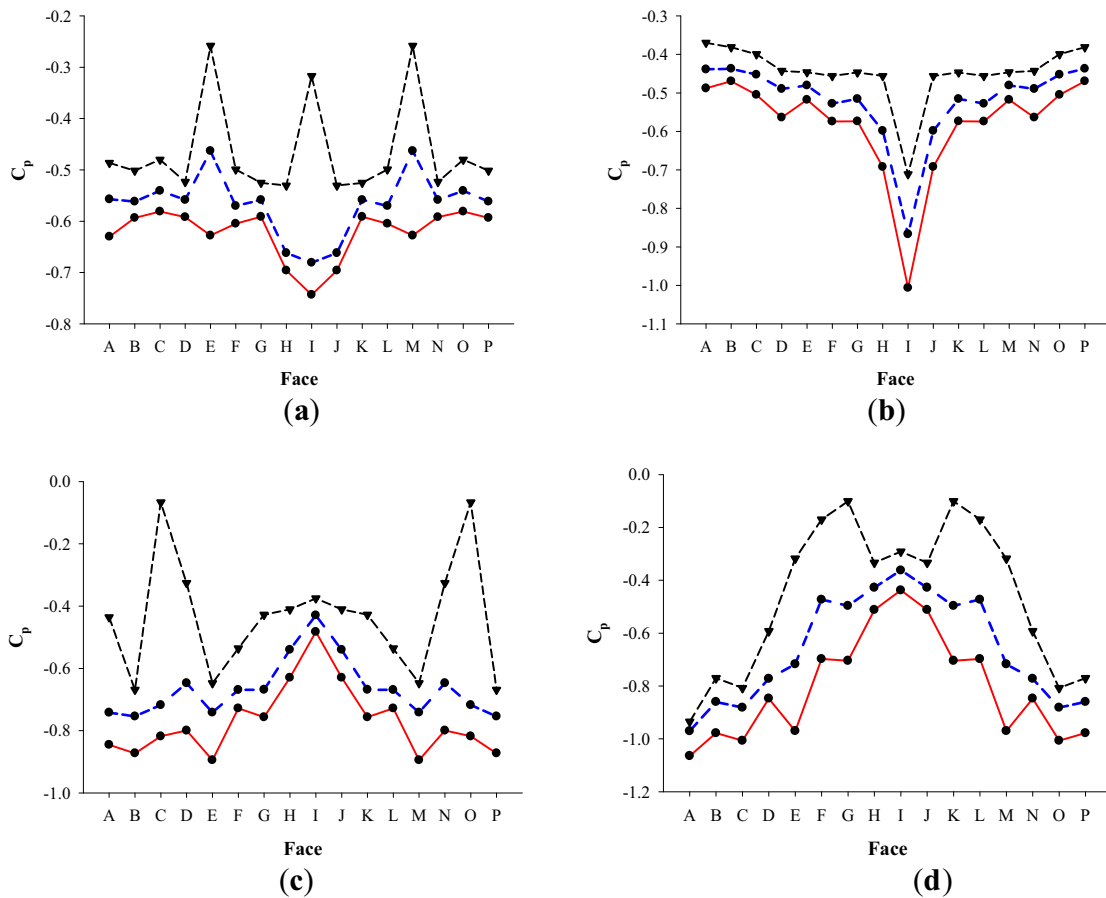


Figure 15. Minimum, maximum and average C_p distribution along periphery of Fish- plan shape building model for (a) Back-to-Back; (b) Front-to-Back; (c) Back-to-Front; (d) Front-to-Front Interference conditions.

emergency they provide fair access to all the interior places. Even, at all the isolated and interference conditions of Fish-plan shape model, less variance of C_p is observed at the faces due to the orientation of both faces to the wind incidence. However, for providing any entrance and exit at any portion of building thorough study for the effects of openings at pedestrian level due to high turbulence at ground level is needed beforehand. Elevated portions like area between Face-G, Face-H and Face-I and between Face-I, Face-J and Face-K can be used as lift areas.

3.4.2 Fish- plan shape building model at all interference conditions In Back-to-Back interference condition (figure 17) the principal and the interfering building models are oriented by placing the back faces i.e. Face-I in front of each other. The principal building is observed to have negative C_p throughout the faces of the model due to the flow separation after the wind is incident on the interfering building and gradually separated in either direction. C_p distribution at Face-I shows little fluctuation from bottom to 5/6th height due to shielding effect of interfering building. Symmetrical flow pattern can be seen

from the figure due to symmetrical plan and orientation of both principal and interfering models. Face-D and Face-N can be seen to have rapid variation in the pressure coefficients from -0.53 to -0.58 throughout the horizontal line due to high swirl of wind caused by reattachment of wind stream at the face. After Face-I all the faces that are parallel to the wind flow have higher magnitude of C_p as compared to all the other non- parallel faces, this characteristic is similar to the isolated 0° condition, however, in this case the nature of pressure coefficients is negative.

In Back-to-Front interference condition (figure 18) the principal and interfering building models are oriented by placing the front face and back face respectively in front of each other. The principal building is observed to have negative C_p throughout the faces of the model due to the shielding effect of the interfering building. Due to upwash large fluctuations of pressure is seen at top 1/6th height at Face-A, Face-B, Face-C, Face-E, Face-P, Face-O and Face-M. Also, suction is observed at 1/3rd height of Face-I due to flow contraction at the rear face of the principal model. At Face-I when the flow is contracted towards the face from

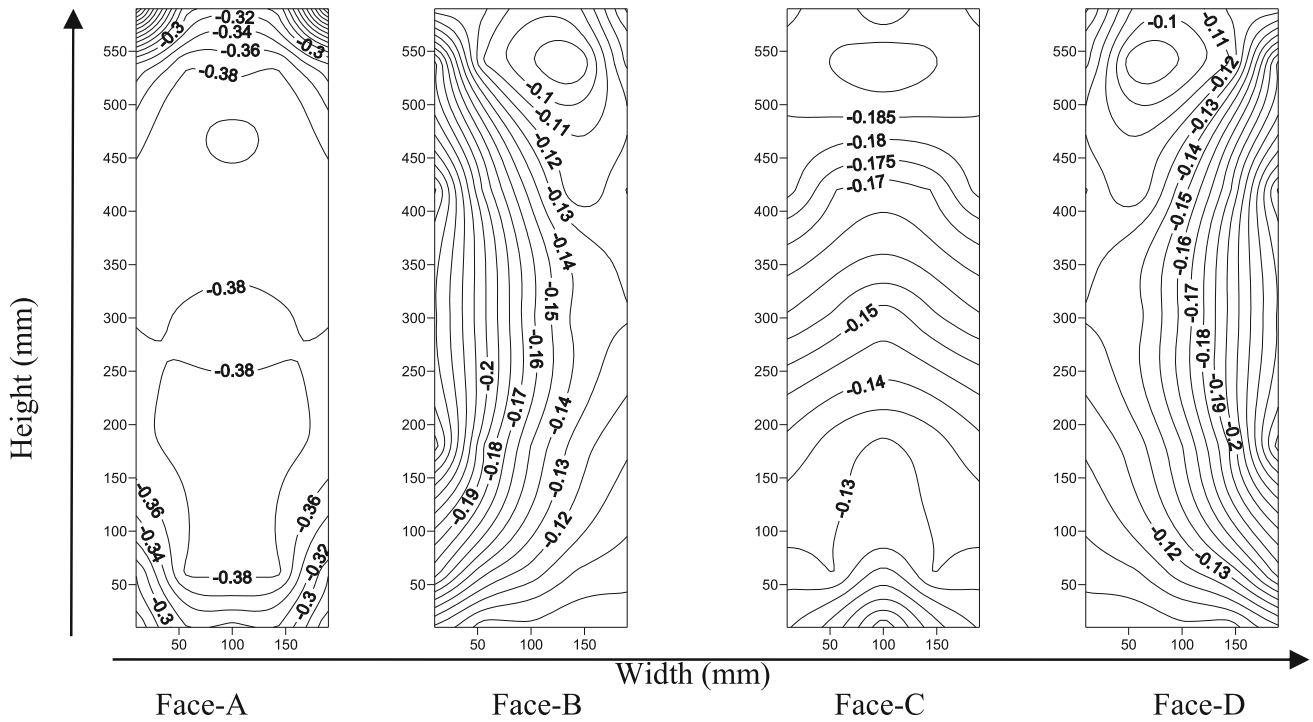


Figure 16. Pressure distribution of Square- plan Shape building model at Full Blockage (interference) wind incidence condition.

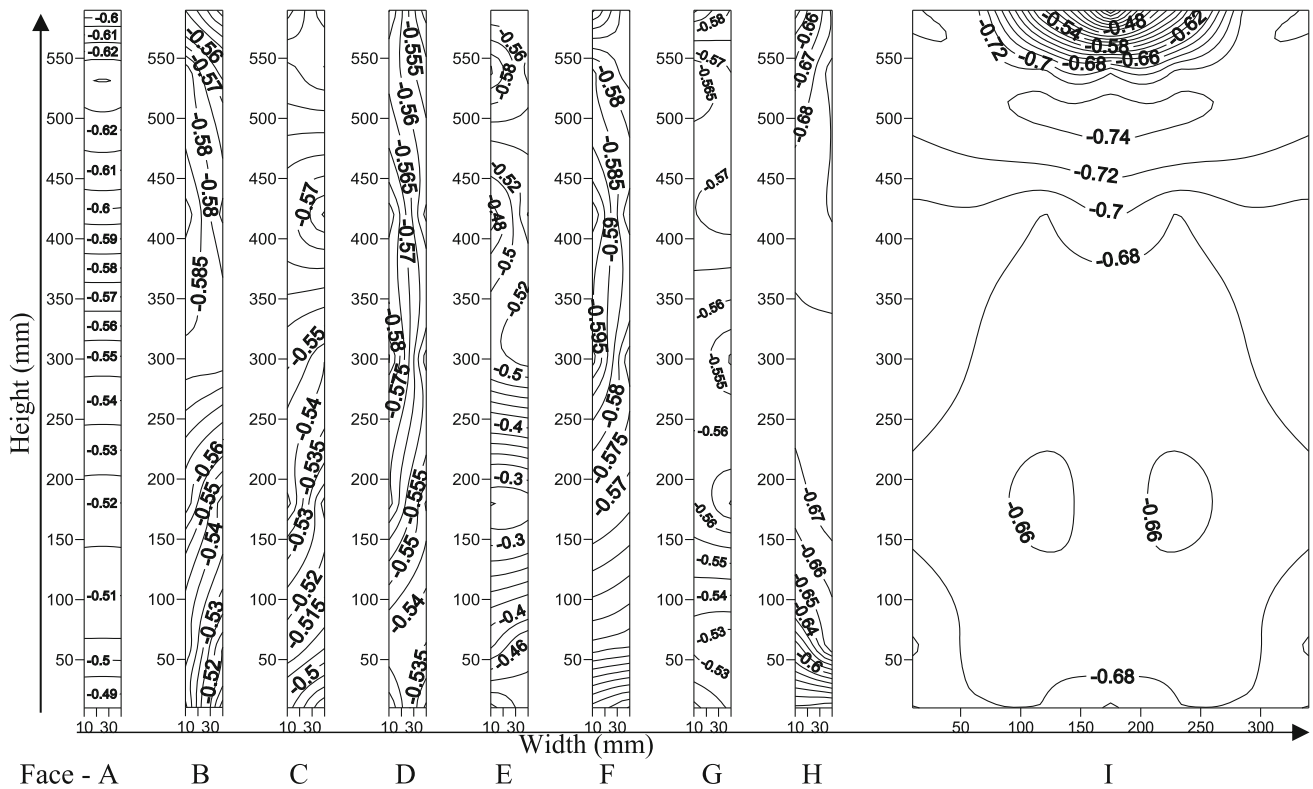


Figure 17. Pressure distribution of selective faces for Fish- plan shape building model at Back-to-Back interference condition.

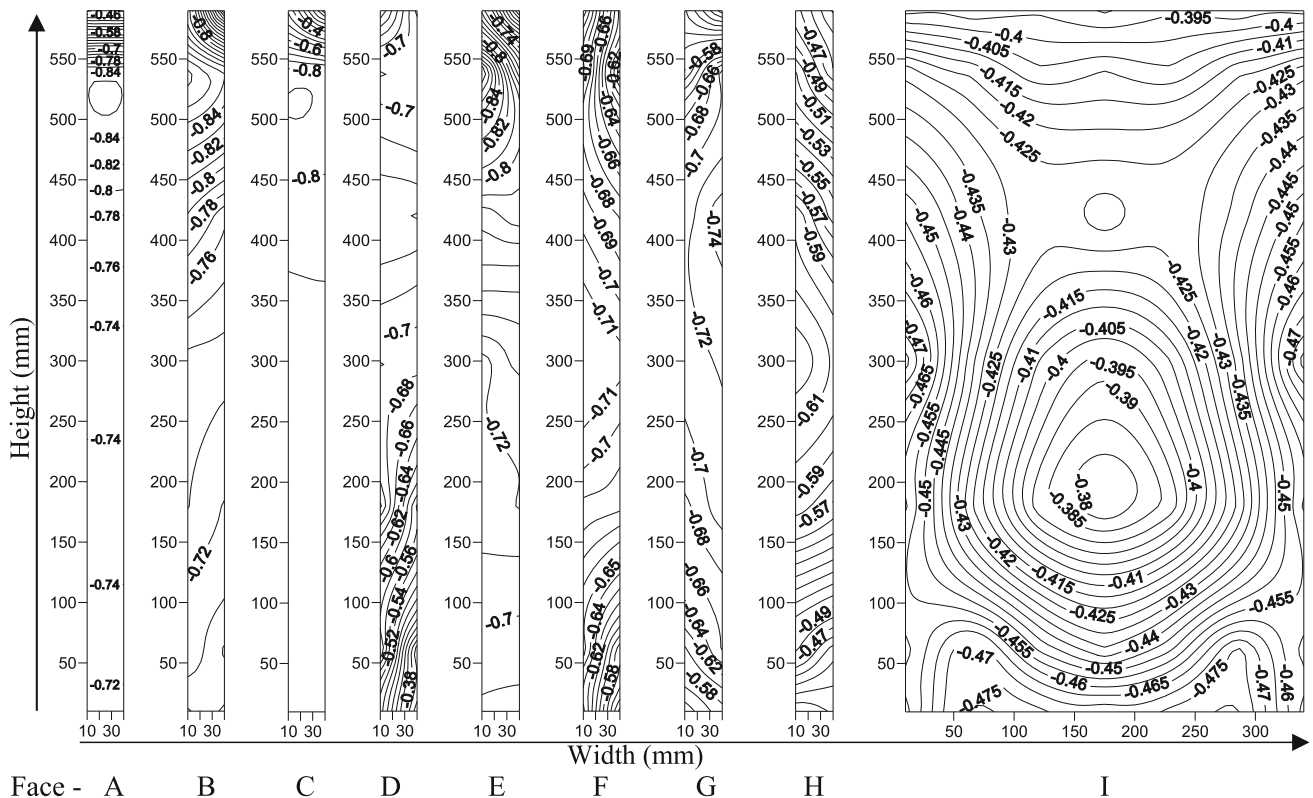


Figure 18. Pressure distribution of selective faces for Fish-plan shape building model at Back-to-Front interference condition.

both sides, rapid changes in the pressure can be observed at both the edges observed as side wash. In this interference condition, due to orientation of both the buildings with respect to each other the downstream wind becomes highly unsettled.

In Front-to-Back interference condition, the principal and interfering building models are oriented by placing the front face and back face respectively in front of each other (figure 19). Two small vortices are observed at 1/3rd height of model at Face-I due to channelling effect arising at the middle of both the models. Further, the wind stream getting deviated at wider angle on both sides thus, generation of wake region can be observed at downstream faces causing decrease in the magnitude of C_p . As a result, highest negative C_p of -0.97 can be observed at Face-I and lowest C_p of -0.38 at Face-A. The wind is seen to be twisting at Face-H and Face-J and also at Face-G and Face-K due to high swirl of stream as a result of interference effect of neighbouring faces.

In Front-to-Front interference condition, the principal and the interfering building models are oriented by placing the Face-A in front of each other (figure 20). The wind after hitting Face-I of the interfering building is deviating on both sides at wider angle, but when it is reaching the principal building model it is further

reattaching to model thus, a channelling effect is created at the middle of both the models. This building model is more or less behaving similar to the model at isolated 0° condition. The difference between the isolated 0° condition and Front-to-Front interference condition is the change in nature of pressure (negative); this is due to the presence of interfering building and thus, formation of wake region at the location of principal building. Face-A is experiencing maximum negative pressure coefficient of -1.06 due to orientation of interfering model and also flow pattern of wind around bluff body. Further downstream side, at Face-F and Face-L the wind is changing its direction, thus twisting of wind stream is observed along the height of model. Large fluctuation (-0.48 to -0.36) in the form of side wash form inner edge is observed at mid height of Face-G and Face-K. Contraction of wind stream towards Face-I is observed, thus experiencing side wash from both the edges symmetrically. Also small vortex is observed at the centreline of Face-I at 1/6th height of model with the centre of -0.295 due to high swirl of wind caused by contraction of wind stream. At 180° isolated wind direction and all interference conditions perpendicular faces (Face-C, Face-E, Face-G, etc) will suffice the wind induced conditions for providing balconies.

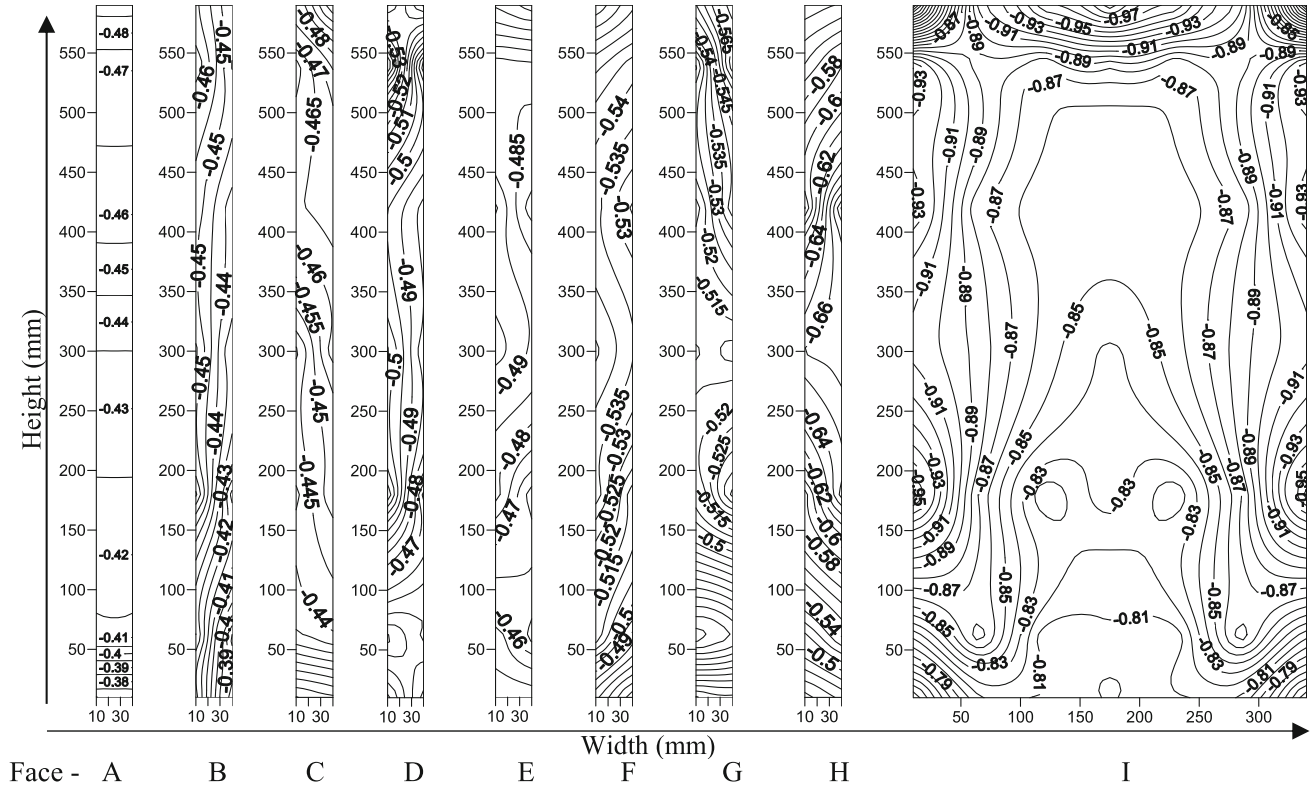


Figure 19. Pressure distribution of selective faces for Fish- plan shape building model at Front-to-Back interference condition.

3.5 Pressure interference factor (average)

The interference effect is an experience caused by presence of one or more objects in the path fluid flow obstructing it to the principal or test object. With the help of this factor the average face C_p variation of the instrumented tall building model is being analysed. Interference factor for Face Average is [22]:

$$\text{Average Interference Factor (IF}_p\text{)} = \frac{\text{Face Average } C_p \text{ with interference}}{\text{Face Average } C_p \text{ without interference}} \quad (3)$$

Interference factors associated with the interference study between square shaped twin building models by changing distance and position between the two in grid form suggests that the study is useful in identification of potential interference issue [20]. In the interference study between twin rectangular building model the overall maximum and minimum IF_p is 1.05 and 0.5 respectively for parallel and perpendicular arrangement among the building models [35]. Whereas, interference factor reaches up to -2.6 for side face in tandem arrangement for a said distance ($x/b = 1.5$) between the twin square plan models [22].

3.5.1 Full Blockage interference condition of Square model All IF_p magnitudes for interference study for Square- plan shape at Full Blockage (figure 21) condition

lies between -0.52 and 0.4. Magnitude of $1 < \text{IF}_p < 1$ represents decrease in C_p of principal building due to shielding effect of interfering building model. Face-A in this interference condition experiences minimum IF_p of -0.52 as due to shielding effect of interfering building model Face-A of principal model is situated in wake region at present interference condition.

3.5.2 Interference conditions of Fish- plan shape model IF_p < 1 represents decrease in C_p due to shielding effect of upstream building resulting with a huge development of turbulence between two interfering twin models. Whereas, IF_p > 1 represents increase in C_p due to upstream building. In the present experimental study for interference between twin Fish- plan shape model (figure 22) only Face-I for Back-to-Back interference condition (figure 22(a)) and Front-to-Back interference condition (figure 22(b)) have IF_p < 1 with magnitudes of -1.55 and -1.98 respectively. Max negative IF_p of the magnitude -11.17 is found at Face-F and Face-L of Back-to-Front interference condition (figure 22(c)) and of magnitude -7.83 in Front-to-Front interference condition (figure 22(d)). Such high negative IF_p values of interference study shows generation of very high suction region at the said faces due to the presence of interfering building. Thus potential interference issue is concentrated at Face-F and Face-L of Back-to-Front and Front-to-Front interference

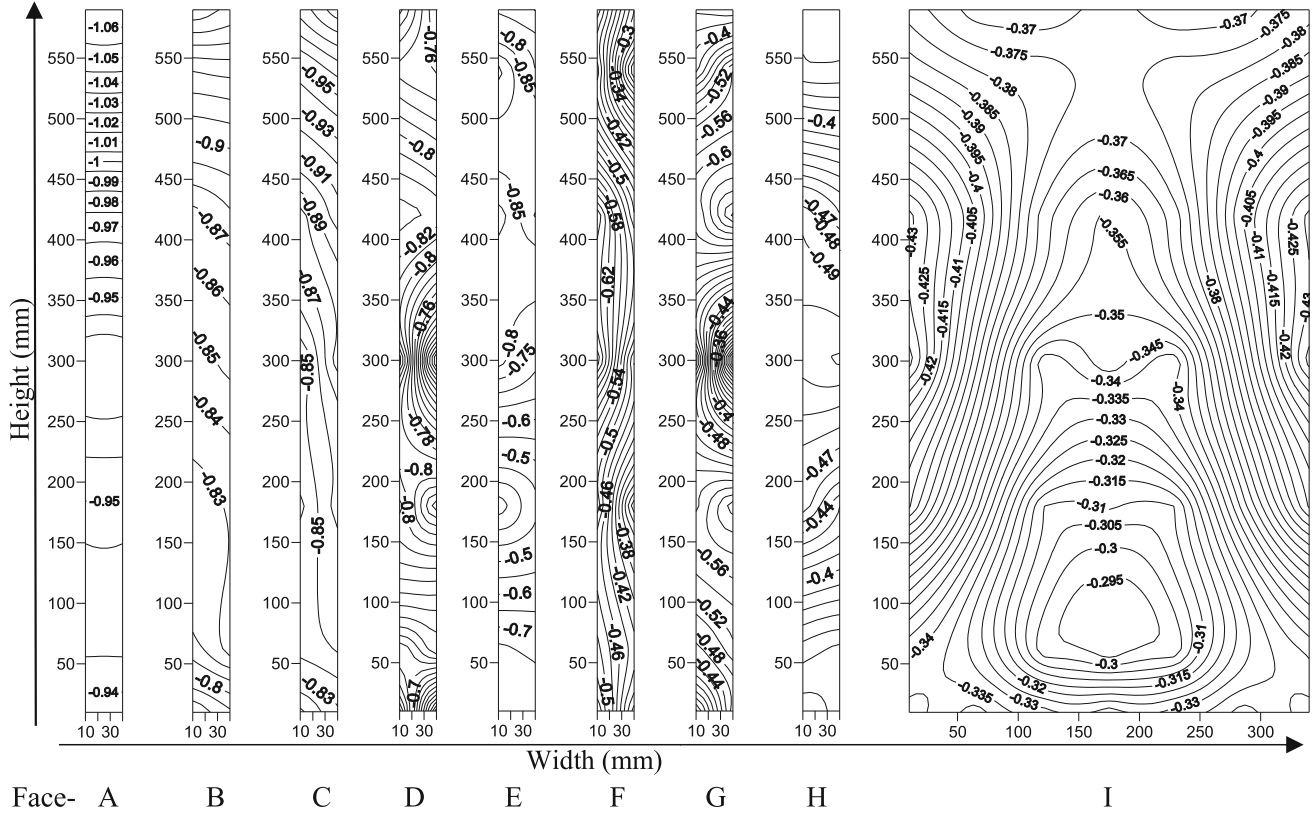


Figure 20. Pressure distribution of selective faces for Fish- plan shape building model at Front-to-Front interference condition.

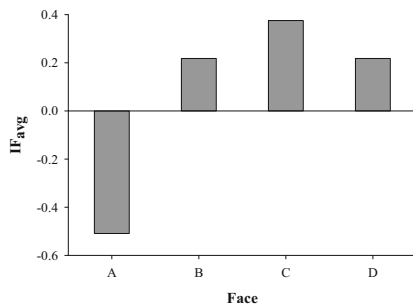


Figure 21. Average Interference Factors (IF_{avg}) of Square- plan shape building model at Full Blockage Interference condition.

conditions. Such high magnitudes of IF_p is not found at any other studies. High magnitudes of IF_p are arising due to the complexity of the plan shape of building and thus accounts for further study of the said cases. Very high IF_p magnitudes show unfavorable positions in buildings from structural design as well as cladding design point of view.

3.6 Base Shear coefficients C_D and C_L

3.6.1 Force Coefficients From the measured pressure at various pressures taping points on all the walls of the principal building model, aerodynamic along-wind force

(Drag force) and across-wind force (Lift force) acting on the model is calculated with the help of formulae as incorporated in the previous studies [22]:

$$F_D = \sum_{i=1}^N (p_i A_i n_{i,along}) \quad (4)$$

$$F_L = \sum_{i=1}^N (p_i A_i n_{i,across}) \quad (5)$$

Where, F_D , and F_L are the along-wind force and across-wind force, respectively. p_i , and A_i are the pressure and tributary area of tap i and $n_{i,along}$ and $n_{i,across}$ are unit direction cosines to the surface of the building.

Force coefficients are representation of forces and calculated with the help of formulae as incorporated in the previous studies [37]:

$$C_D = \frac{F_D}{0.5 \rho B H V^2} \quad (6)$$

$$C_L = \frac{F_L}{0.5 \rho B H V^2} \quad (7)$$

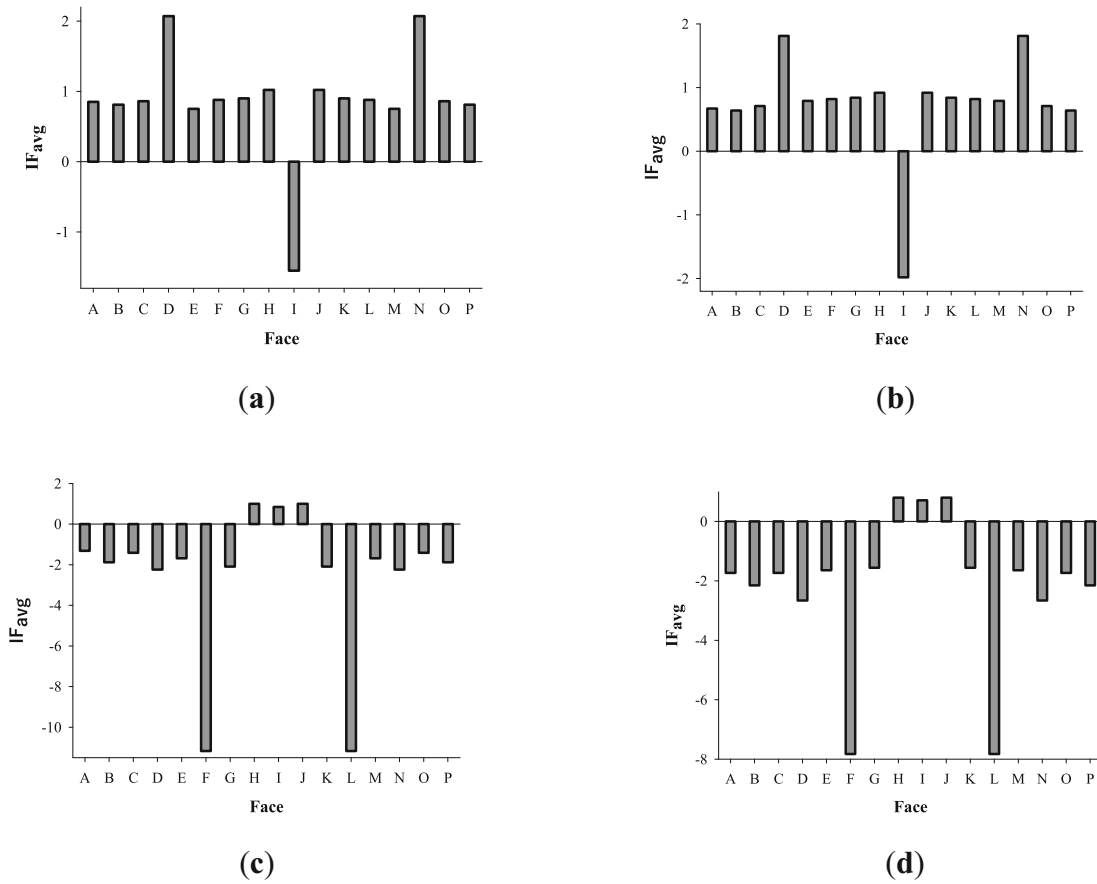


Figure 22. Average Interference Factors (IF_{avg}) (a) Back-to-Back; (b) Front-to-Back; (c) Back-to-Front; (d) Front-to-Front Interference conditions.

Where, C_D and C_L are base force coefficients, B is the dimension of building model perpendicular to incident wind, H is the height of the model.

3.6.2 Force coefficients of Square -plan shape building model The magnitudes of Drag force coefficients when compared to magnitudes of Lift force coefficients portray large variation due to the normal wind directions in the direction of symmetry of model (figure 23). Dominant wind direction of present study is shown by Along isolated wind conditions. Where the C_D at full blockage condition shows decrement, the C_L value shows near about same magnitudes. Due to symmetrical form of principal as well as interference Square-plan shape building models along both the axis thus, resulting in symmetrical flow at full blockage condition around twin building models, non-variation of Lift force coefficient is induced. Shielding effect refers to the situation where, before reaching the principal structure wind has to move through any structure(s) located on the upstream wind side [14]. Due to the shielding effect, it is clear from the contour

plots of all the interference conditions that the principal structure is under suction, i.e. negative pressure. The faces lying at the interface between the twin models experiences maximum suction in all interference conditions. Due to its high magnitude of coefficients, the nature of force and pressure at the interfaces dominates the resulting direction of force and pressure. Thus the resultant net pressure coefficients and net Drag force coefficients is negative in nature is acting against the direction of wind incidence.

3.6.3 Force coefficients of Fish- plan shape building model It has been widely recognized that external shapes of tall buildings play an important role in the generation of wind loads on high-rise structures [38]. In addition, it was relayed by Sakamoto and Haniu [39] and Song *et al* [40] that the fluctuating components of drag force coefficients C_D and lift force coefficients C_L of the principal building model are not significantly affected by the gap distance of the interfering model. Unlike Full blockage condition all the interfering conditions shows variation in C_L because asymmetrical cross-section plan of Fish- plan shape model

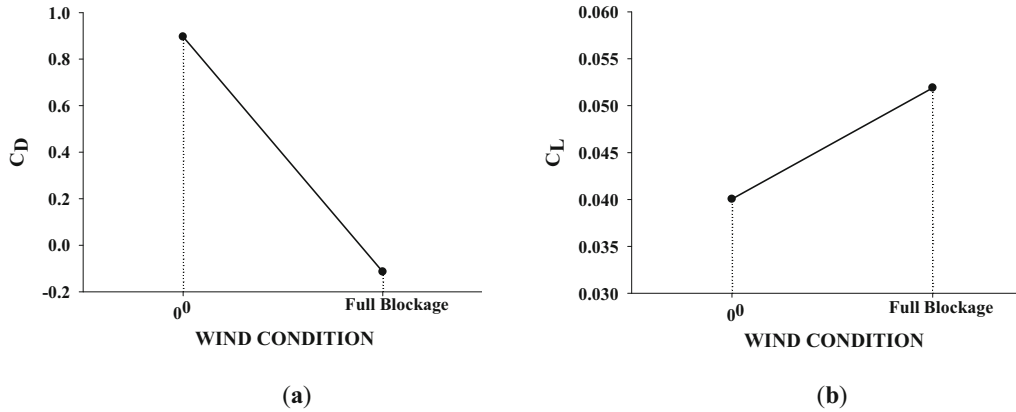


Figure 23. Base force coefficients of Square- plan shape building model at all wind conditions (a) Drag Force Coefficients; (b) Lift Force Coefficients.

along one axis thus, generation of huge turbulence at principal model depending upon the orientation of twin models.

In figure 24 for Fish- plan shape model maximum C_D of 1.1 is found at 180° isolated wind incidence followed by 0° wind incidence with 23% decrement. Maximum C_L is manifested by 0° wind incidence is 0.14 closely followed by 180° wind incidence with 7% decrease. Overall, minimum C_D is exhibited by Back-to-Back inference condition whereas, minimum C_L is exhibited by Front-to-Back and Front-to-Front interference condition. Enormous decrease in magnitude irrespective of nature of C_D and C_L is shown for all interfering conditions when observed against isolated wind conditions mainly due to shielding effect of interfering building model and hence reduced effect of direct exposure to wind. From the results, it is clearly visible that the overall efficiency of principal building is enhanced due to interference effect [22].

3.7 Force Interference Factors

$$\text{Drag Force Interference Factor (IF}_{CD}) = \frac{C_D \text{ at base with interfering building}}{C_D \text{ at base without interfering building}} \quad (8)$$

$$\text{Lift Force Interference Factor (IF}_{CL}) = \frac{C_L \text{ at base with interfering building}}{C_L \text{ at base without interfering building}} \quad (9)$$

The corresponding interference positions of maximum EIF i.e. peak acceleration response interference factors are shown at the top of each bar by x/b, y/b in figure 25 for along-wind and across-wind directions for various breadth ratios between same height of models [41]. The study by Yu *et al* [41] has been performed for twin models for H/B of 6:1, mean wind velocity at top of model is 12.9 m/s and $\alpha = 0.22$. It is found that maximum Along-wind (Drag force) EIF was 1.6 at (x/b, y/b) = (2, 0.9) which shows high

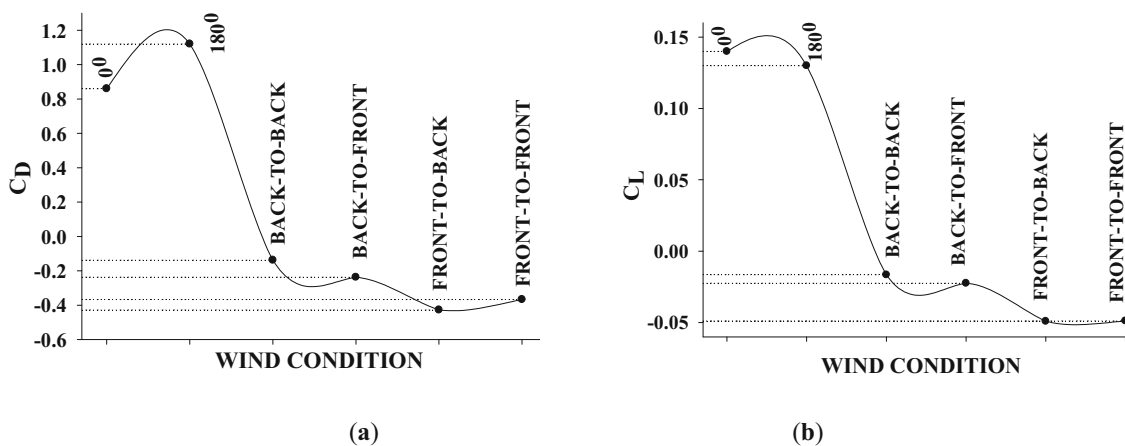


Figure 24. Base force coefficients for Fish- plan shape building model at all wind conditions (a) Drag Force Coefficients; (b) Lift Force Coefficients.

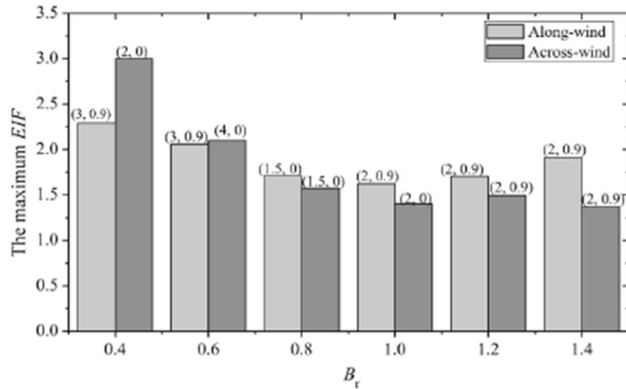


Figure 25. The results of maximum values of EIF (Envelope Interference Factor) [41].

influence of staggered arrangement for $B_r = 1.0$; where, B_r is Breadth ratio among breadth of interfering model and principal building model. Maximum Across-wind (Lift force) EIF is 1.4 for $B_r = 1.0$; at $(x/b, y/b) = (2, 0)$ i.e. 16.67% clear gap between twin models at tandem arrangement; which shows dominance of shielding which is also visible at present experimental work. Similarly, the results from study by Mara [20] deals with models under urban exposure with H/D of 7:1 at a wind velocity of 4.9m/s in tandem arrangement. It is found that at minimum clear gap of 28.6% of height of model, the Along-wind (Drag force) IF_m is 0.6 and Across-wind (Lift force) IF_m is 1.0 for study by Mara [20]. Whereas, in current study for Square-plan shape model at Full Blockage condition for reduced gap between twin models Along-wind (Drag force) IF_{CD} of -0.12 is found to be decreasing with reduced gap between twin models when compared to study conducted out by Mara [20]. For current study where, IF_{CD} is decreasing due to shielding effect but IF_{CL} of 1.25 is found to be increasing at reduced gap between twin models due to formation of turbulent shear layer at side faces of principal building model. For all interference conditions for Fish- plan shape model, magnitudes of IF_{CD} and IF_{CL} (table 2) are very less as compared to said studies; this variation is mainly associated with the external shape of the building. Maximum IF_{CD} (-0.43) is shown by Front-to-Front condition and maximum IF_{CL} (-0.38) is shown by Front-to-Back condition is shown by Front-to-Front condition as orientation of

Table 2. Force Interference Factors for Square and Fish- plan shape building model.

Interference conditions	IF_{CD}	IF_{CL}
Full Blockage	-0.12	1.25
Back-to-Back	-0.12	-0.13
Back-to-Front	-0.28	-0.16
Front-to-Back	-0.38	-0.38
Front-to-Front	-0.43	-0.35

twin models to each other as well as to incident wind increases turbulence between both the models. Under same working environment, the Fish -plan shape principal models at all interference conditions display less influence on lift forces compared to Square-plan shape model at Full blockage condition. Whereas, increased influence on drag force is noticeable at Back-to-Front, Front-to-Back and Front-to-Front interference condition of fish-plan shape model compared to Square-plan shape model at Full blockage interference condition.

3.8 Distribution of pressure coefficient along centre vertical line for each face at all wind incidence conditions

Variation of C_p along vertical centerline of all faces for isolated and interfering wind incidence conditions are shown in figures 26 to 34. All the vertical line plots are simplified images of complex contour plots for faces and gives a broader picture of variation of C_p along height of face. The comparison centerline plots of all faces over varied wind incidences gives a fine picture of the change of flow pattern along faces in a particular wind direction.

An attempt has been made to find out a generalised relation between relative height and C_p along vertical centreline at each face for all isolated and interference conditions. The variations of C_p along height cannot be linearly regressed due to wide variable ranges and hence the curve fitting process of all the isolated and interference parameters is complex. From regression analysis a basic equation (Eq.10) is further developed to consider pressure coefficient distribution along vertical centreline at each face for present experimental conditions.

$$C_p = -a \times \left[\frac{1}{\left(1 + e^{-b \times \left(\frac{H_i}{H}\right)}\right)} \right]^n \quad (10)$$

Where; H_i/H is relative height with H_i varying as 0 mm, 10 mm, 60 mm, 180 mm, 300 mm, 420 mm, 540 mm, 590 mm and 600 mm and H is height of model (600mm), a , b and n are constants and can be seen varying with the faces among various wind incidence conditions.

3.9 Vertical centreline C_p distribution at each face along Square- plan shape model for Isolated and interference condition

3.9.1 Square-plan shape building model for 0° isolated condition The present study exhibits symmetrical flow pattern along vertical centerline due to symmetrical shape of both the models along the direction of wind incidences, this is also noticeable in the study for ‘+’

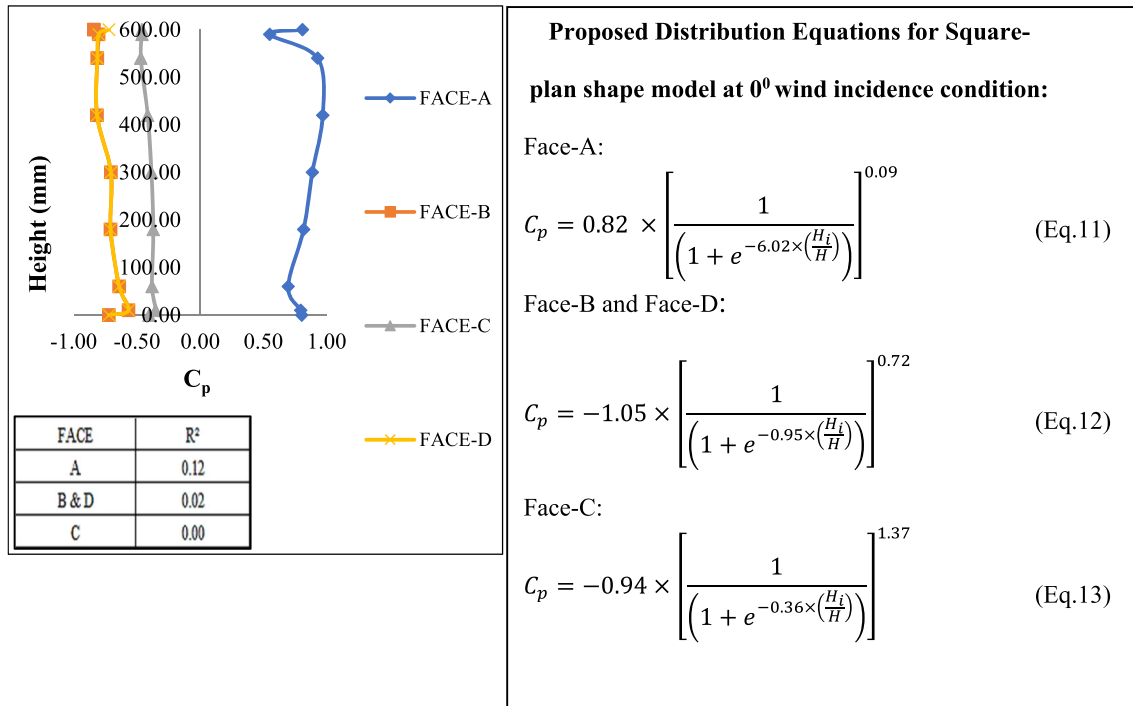


Figure 26. Vertical centreline pressure Distribution of Square- plan shape model at 0° wind incidence condition

plan shape building model at similar working environment. [25].

Eq.(11) to Eq. (13) satisfies the vertical centreline C_p distribution shown in figure 26. Centreline at Face-B and Face-D over line each other due to symmetrical distribution and thus is generalised with the help of Eq.(12).

3.9.2 Full Blockage interference condition of Square model Due to shielding of principal model and vortex generation at Face-A (as validated from figure 16) at Full-blockage condition (figure 27) the vertical centreline shows large variation along height. The centreline distributions of side faces at this interference condition is satisfied by Eq.(15) and front and back faces by Eq.(14) and Eq.(16) respectively.

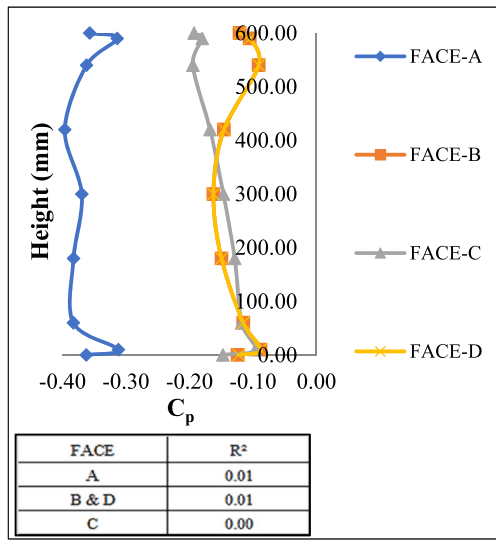
3.10 Vertical centreline C_p distribution at each face along Fish- plan shape model for Isolated and interference condition

3.10.1 Fish- plan shape building model for 0° isolated condition From the studied carried out by Bhattacharyya [32] it is apparent that results of distribution of pressure coefficient over a E-plan shaped building model at windward, leeward and all other faces exhibit huge differences with that of regular cubical model, the similar results are also exhibited by Fish- plan shape model

(figure 28) at isolated 0° wind incidence of the present study, given the experimental environment is same. Although all the vertical centerline C_p distributions satisfy, the distribution along Eq.(10) but due gradual increase in the cross-section dimension of model large variation of C_p distribution among neighboring side faces is seen and thus, generalization of all distributions into one equation is impossible with minimum residual (R^2). Where, R^2 is a statistical measure, which shows the proportion of variance in a regression equation. Eq. (17) to Eq. (20) for satisfies vertical centerline distribution of grouped faces.

The study is taken into consideration because of close resemblance of exterior shape of Triangle- plan shape model [42] to Fish- plan shape model. Figure 29 shows vertical line plot of triangular model at 0° wind direction at Terrain B (Power law index 0.16), China. This power law index is associated with open terrain. Due to fewer obstructions at this terrain category, the distribution of C_p varies when compared to Fish- plan shape model at 180° wind incidence at present study condition. Hence, it is also concluded that Eq.(10) does not satisfy the vertical distribution profile of figure 29.

Only Face-I among all the surfaces for 180° wind incidence condition (figure 30) at shows positive C_p distribution throughout the height, the vertical distribution of which is satisfied by Eq.(23). Due to huge cross-sectional magnitude of Face-I the face acts as shield for upstream faces and thus all other faces are under wake region and



Proposed Distribution Equations for Square- plan shape model at Full Blockage interference condition:

Face-A:

$$C_p = -0.36 \times \left[\frac{1}{\left(1 + e^{-394.19 \times \left(\frac{H_i}{H} \right)} \right)} \right]^{-0.01} \quad (\text{Eq.14})$$

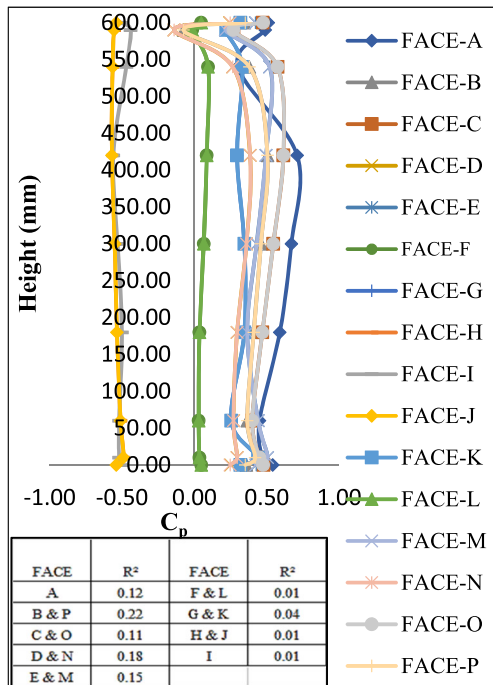
Face-B and Face-D:

$$C_p = -0.23 \times \left[\frac{1}{\left(1 + e^{0.05 \times \left(\frac{H_i}{H} \right)} \right)} \right]^{0.90} \quad (\text{Eq.15})$$

Face-C:

$$C_p = -0.34 \times \left[\frac{1}{\left(1 + e^{-0.77 \times \left(\frac{H_i}{H} \right)} \right)} \right]^{1.55} \quad (\text{Eq.16})$$

Figure 27. Vertical centreline pressure Distribution of Square- plan shape model at Full Blockage interference condition.



Proposed Distribution Equations for Fish- plan shape building model at 0° isolated condition:

building model at 0° isolated condition:

Face-A, Face-C, Face-E, Face-M and Face-O:

$$C_p = 3.43 \times \left[\frac{1}{\left(1 + e^{0.04 \times \left(\frac{H_i}{H} \right)} \right)} \right]^{2.67} \quad (\text{Eq.17})$$

Face-B, Face-D, Face-G, Face-K, Face-N and Face-P:

$$C_p = 0.41 \times \left[\frac{1}{\left(1 + e^{2.51 \times \left(\frac{H_i}{H} \right)} \right)} \right]^{0.22} \quad (\text{Eq.18})$$

Face-F and Face-L:

$$C_p = 0.09 \times \left[\frac{1}{\left(1 + e^{-0.38 \times \left(\frac{H_i}{H} \right)} \right)} \right]^{1.04} \quad (\text{Eq.19})$$

Face-H, Face-I and Face-J:

$$C_p = 0.43 \times \left[\frac{1}{\left(1 + e^{-0.30 \times \left(\frac{H_i}{H} \right)} \right)} \right]^{-0.32} \quad (\text{Eq.20})$$

Figure 28. Vertical centreline pressure Distribution of Fish- plan shape model at 0° wind incidence condition.

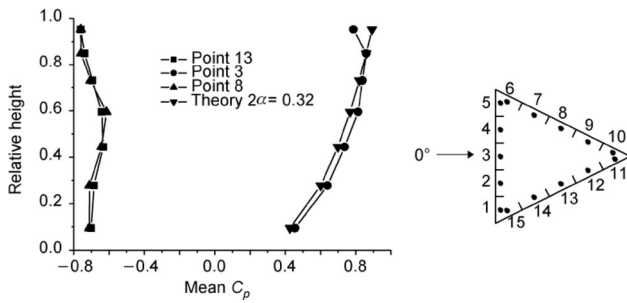


Figure 29. Mean C_p along height of triangular model at 0° wind direction, Terrain B, China [42].

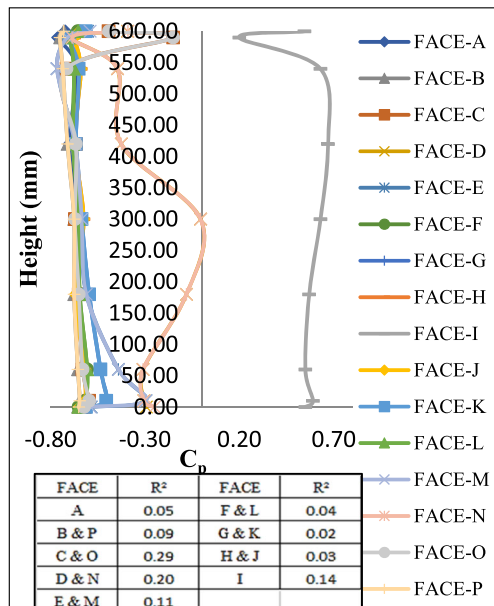
experiences suction. The angle of separation of wind is wide and hence less influence of wind is experienced by upstream faces and thus, Eq.(21) satisfies the vertical distribution of most of the faces. However, Face-D and Face-N undergoes huge variation of C_p throughout its height due to unification of wind stream thus, Eq. (22) meets the condition.

3.10.2 Fish- plan shape building model for 180° isolated condition See figures 30 and 31.

3.10.3 Back-to-Back Blockage interference condition of Fish- plan shape model In Back-to-Back (figure 31)

and Front-to-Back interference (figure 32) condition of Fish- plan shape twin building models due to orientation of principal building model the upstream faces of instrumented building experiences less variation among side faces. In both of these interference conditions Face-I acts as shield to upstream faces. Flow separation is apparent from the edges of Face-I and thus wake region is generated at neighbouring faces. Due to less influence of incident wind, less fluctuation of C_p is noticeable at side faces. Through careful regression analysis Eqs. (24) to (27) is generated which satisfies the vertical centreline profiles of said grouped faces.

3.10.4 Front-to-Back Blockage interference condition of Fish- plan shape model At Back-to-Front (figure 33) and Front-to-Front (figure 34) interference condition due to orientation of principal to downstream interfering model channelling effect is experienced at the interface between twin models. Due to this channelling effect the neighbouring faces are experiencing near about similar distribution and thus are satisfied by distribution Eq.(28) and Eq.(31) respectively. The distribution of upstream faces is governed by point of separation of wind stream as well as cross sectional shape of model which in this condition is gradually increasing. Due to huge fluctuation of C_p at all the surfaces generalising equation for vertical centreline



Proposed Distribution Equations for Fish- plan shape

model at 180° wind incidence condition:

Face-A, Face-B, Face-C, Face-E, Face-F, Face-G, Face-H, Face-J, Face-K, Face-L and Face-M, Face-O and Face-P:

$$C_p = -1.26 \times \left[\frac{1}{\left(1 + e^{-0.26 \times \left(\frac{H_i}{H} \right)} \right)} \right]^{1.20} \quad (\text{Eq.21})$$

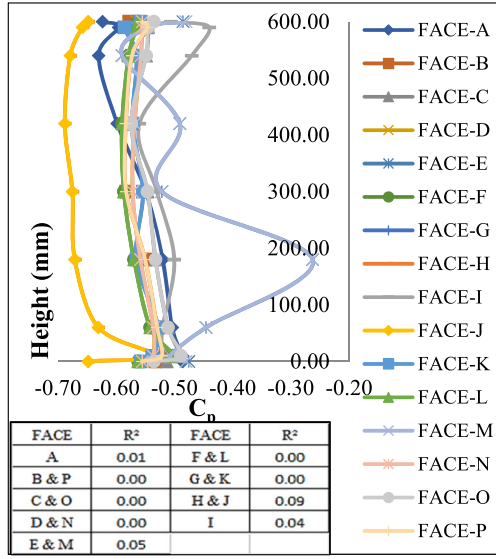
Face-D and Face-N:

$$C_p = -187 \times \left[\frac{1}{\left(1 + e^{-0.22 \times \left(\frac{H_i}{H} \right)} \right)} \right]^{10.14} \quad (\text{Eq.22})$$

Face-I:

$$C_p = 3.34 \times \left[\frac{1}{\left(1 + e^{0.14 \times \left(\frac{H_i}{H} \right)} \right)} \right]^{2.48} \quad (\text{Eq.23})$$

Figure 30. Vertical centreline pressure Distribution of Fish- plan shape model at 180° wind incidence condition.



Proposed Distribution Equations for Fish- plan shape

model at Back-to-Back interference condition:

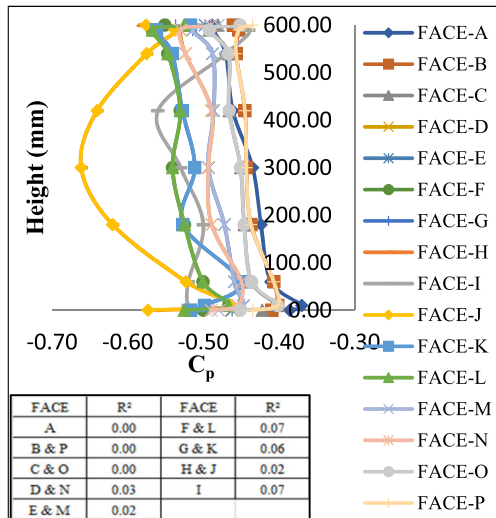
Face-A, Face-B, Face-C, Face-D, Face-F, Face-G, Face-H, Face-I, Face-J, Face-K, Face-L, Face-N, Face-O and Face-P:

$$C_p = -0.57 \times \left[\frac{1}{\left(1 + e^{-5.04 \times \left(\frac{H_i}{H} \right)} \right)} \right]^{0.13} \quad (\text{Eq.24})$$

Face-E and Face-M:

$$C_p = -0.42 \times \left[\frac{1}{\left(1 + e^{4.57 \times \left(\frac{H_i}{H} \right)} \right)} \right]^{-0.05} \quad (\text{Eq.25})$$

Figure 31. Vertical centreline pressure Distribution of Fish- plan shape model at Back-to-Back interference condition.



Proposed Distribution Equations for Fish- plan

shape model at Front-to-Back interference

condition:

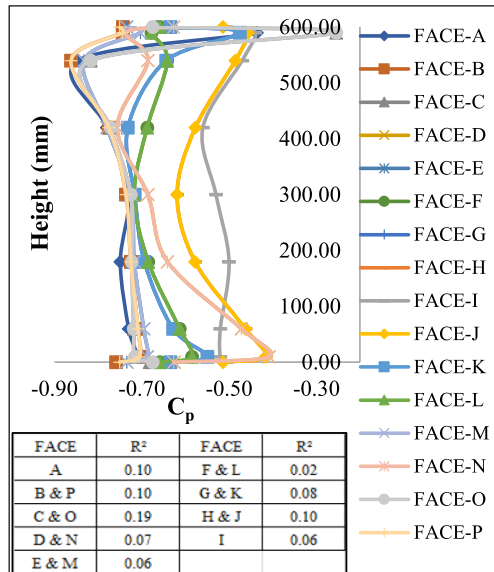
Face-A, Face-B, Face-C, Face-D, Face-E, Face-F, Face-G, Face-K, Face-L, Face-M, Face-N, Face-O and Face-P:

$$C_p = -0.60 \times \left[\frac{1}{\left(1 + e^{-0.94 \times \left(\frac{H_i}{H} \right)} \right)} \right]^{0.64} \quad (\text{Eq.26})$$

Face-H, Face-I and Face-J:

$$C_p = -0.60 \times \left[\frac{1}{\left(1 + e^{-8.48 \times \left(\frac{H_i}{H} \right)} \right)} \right]^{0.25} \quad (\text{Eq.27})$$

Figure 32. Vertical centreline pressure Distribution of Fish- plan shape model at Front-to-Back interference condition.



Proposed Distribution Equations for Fish- plan shape model at Back-to-Front interference condition:

Face-A, Face-B, Face-C, Face-O and Face-P:

$$C_p = -0.37 \times \left[\frac{1}{\left(1 + e^{-0.26 \times \left(\frac{H_i}{H}\right)}\right)} \right]^{-1.0} \quad (\text{Eq.28})$$

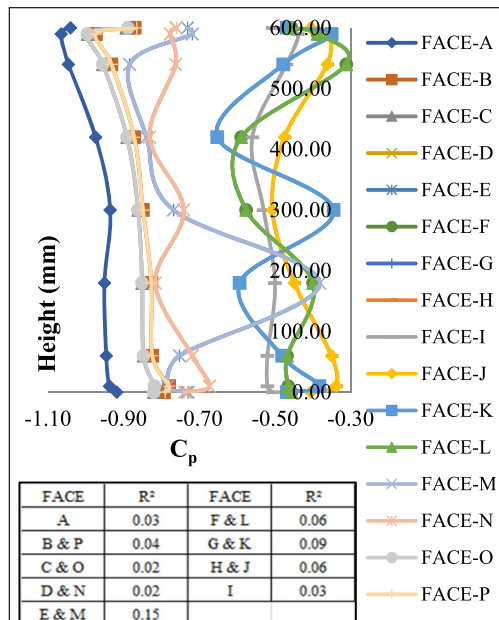
Face-D, Face-E, Face-F, Face-G, Face-K, Face-L, Face-M and Face-N:

$$C_p = -0.74 \times \left[\frac{1}{\left(1 + e^{-2.95 \times \left(\frac{H_i}{H}\right)}\right)} \right]^{0.33} \quad (\text{Eq.29})$$

Face-H, Face-I and Face-J:

$$C_p = -0.40 \times \left[\frac{1}{\left(1 + e^{-0.66 \times \left(\frac{H_i}{H}\right)}\right)} \right]^{-0.66} \quad (\text{Eq.30})$$

Figure 33. Vertical centreline pressure Distribution of Fish- plan shape model at Back-to-Front interference condition.



Proposed Distribution Equations for Fish- plan shape model at Front-to-Front interference condition:

Face-A, Face-B, Face-C, Face-O and Face-P:

$$C_p = -1.34 \times \left[\frac{1}{\left(1 + e^{-0.51 \times \left(\frac{H_i}{H}\right)}\right)} \right]^{0.65} \quad (\text{Eq.31})$$

Face-D, Face-E, Face-M and Face-N:

$$C_p = -2.20 \times \left[\frac{1}{\left(1 + e^{-0.14 \times \left(\frac{H_i}{H}\right)}\right)} \right]^{1.67} \quad (\text{Eq.32})$$

Face-F, Face-G, Face-H, Face-I, Face-J, Face-K and Face-L:

$$C_p = -0.50 \times \left[\frac{1}{\left(1 + e^{2.55 \times \left(\frac{H_i}{H}\right)}\right)} \right]^{0.06} \quad (\text{Eq.33})$$

Figure 34. Vertical centreline pressure Distribution of Fish- plan shape model at Front-to-Front interference condition.

profile is difficult and thus Eq.(29), Eq.(30), Eq.(31) and Eq.(32) and Eq.(33) is generated in order to satisfy the vertical distributions of said faces.

3.10.5 Back-to-Front Blockage interference condition of Fish- plan shape model See figure 33.

3.10.6 Front-to-Front Blockage interference condition of Fish- plan shape model See figure 34.

4. Conclusions

This paper summarizes the findings of an extensive wind-tunnel study on wind induced pressure and base shear on isolated and interference Square and Fish- plan shape tall building models. The experiments are carried out under conditions prevailing power law exponent of 0.22 and under a constant flow of wind with velocity of 10 m/sec. For better comparison the volume of both the models are kept same with plan area of model as 40000 mm² and height as 600 mm. For isolated standing model conditions, 0° and 180° wind directions are considered. For interference studies, five peculiar arrangements of twin building models are taken by considering a constant 60 mm distance i.e. 10% of the height of model between both the twin models. The current study shows that pressure and base shear induced on model building is significantly affected by the model geometry and condition of wind incidence (isolated or interference). The important outcomes of the present study are summarized below.

- Fish- plan shape of building model has differing results of C_p values from that of standard square and rectangular models thus; use of standard codes will not suffice for cladding design.
- Under same working environment the test results for windward face of isolated Fish- plan shape model at 0° i.e. Face-A (Face Value = 0.56) and 180° i.e. Face-I (Face value = 0.44) when compared to square shape model at 0° i.e Face-A (Face Value = 0.71) wind incidences show 21% and 38% decrease in pressure respectively.
- Under same working environment the test results for leeward face of isolated Fish- plan shape model at 0° i.e. Face-I (Face Value = -0.51) and 180° i.e. Face-A (Face value = -0.76) when compared to square shape model at 0° i.e. Face-C (Face Value = -0.41) wind incidences show increase in suction by 24% and 85% respectively.
- At 180° isolated wind incidence condition, Fish- plan shape model experiences about 30% decrease in the overall maximum pressure ($C_p = 0.68$) and also 5% decrease in overall maximum suction ($C_p = -0.83$) due to shielding effect of Face-I when compared to isolated Square shape model with overall maximum pressure of

$C_p = 0.97$ and overall maximum suction of $C_p = -0.87$.

- At Front-to-Front interference condition the Fish- plan shape model experiences huge increase in maximum suction ($C_p = -1.06$) when compared to Full blockage interference condition of Square shape model ($C_p = -0.40$) due to complex cross section plan shape of Fish-plan shape model.
- For the Fish- plan shape model, C_D and C_L (1.12 and 0.13 respectively) are found to be increased by 25% and 225% respectively at 180° isolated condition compared to the isolated Square- plan shape model with C_D and C_L (0.90 and 0.04 respectively) at 0° wind incidence due difference in the cross sectional plan between both the models.
- Maximum increase in C_D is encountered by Fish- plan shape model at Front-to-Back interference condition ($C_D = -0.43$) when compared to than that at Full blockage interference condition of the Square- plan shape model ($C_D = -0.11$).
- At Fish- plan shape model due to complexity of the building model shape, very high IF_p of magnitude - 11.17 is generated at Front-to-Back interference condition and of magnitude -7.83 is generated at Front-to-Front interference condition both thus, attracts potential interference issues at such faces.
- Drag and Lift Force coefficients for interference cases largely depend upon the orientation and cross section shape of interfering model in present study.
- The overall efficiency of principal building is enhanced due to interference effect in both Square and Fish-plan shape model.
- At Fish- plan shape model overall maximum efficiency in shown by Back-to-Back interference condition when Drag force and Lift force is considered.
- Overall maximum efficiency in terms of induced wind pressure and force is exhibited by Square- plan shape model at Full Blockage condition.

5. Recommendations for future research

Based on the present study, it is recommended that future studies should be carried out in the following areas:

- Analytical analysis of the all models at the same working environment.
- Response study of the scaled model.
- Disturbed flow field characteristics and the consequent wind load modification by particle image velocimetry (PIV) for twin building models.
- Study of interference effects on principal building with twin building model in various other configurations like, varying angle of wind incidence on

- 100% blockage in various tandem and staggered arrangements between twin building models.
- v. Study of interference effects on principal building with couple of buildings of different plan shapes in near vicinity.
- vi. Assessment of aerodynamic modifications like openings, corner cut, recessed, chamfered etc. on the wind pressure distribution.
- vii. Dynamic response analysis of the buildings using time varying wind data.

Acknowledgements

We thank Prof. A K Ahuja, Visiting Faculty, Civil Engineering Department, IIT Jammu (India). Without his support and invaluable suggestions this experimental study would not have been feasible.

References

- [1] Hassanli S, Gang H, Kwok K C S and Fletcher D F 2017 Utilizing cavity flow within double skin façade for wind energy harvesting in buildings. *J. Wind Eng. Ind. Aerodyn.* 167: 114–127
- [2] Irwin P A 2008 Bluff body aerodynamics in wind engineering. *J. Wind Eng. Ind. Aerodyn.* 6–7: 701–712
- [3] Xu A, Xie Z N, Fu J Y, Wu J R and Tuan A 2014 Evaluation of wind loads on super-tall buildings from field-measured wind-induced acceleration response. *Struct. Design Tall Special Building.* 23: 641–663. <https://doi.org/10.1002/tal.1065>
- [4] Zheng C, Xie Y, Khan M, Wu Y and Liu J 2018 Wind-induced responses of tall buildings under combined aerodynamic control. *Eng. Struct.* 175: 86–100. <https://doi.org/10.1016/j.engstruct.2018.08.031>
- [5] Aly A M 2013 Pressure integration technique for predicting wind-induced response in high-rise buildings. *Alex. Eng. J.* 52: 717–773
- [6] Gang H, Hassanli S, Kwok K C S and Tse K T 2017 Wind-induced responses of a tall building with a double-skin façade system. *J. Wind Eng. Ind. Aerodyn.* 68: 91–100
- [7] Irwin P A 2009 Wind engineering challenges of the new generation of super-tall buildings. *J. Wind Eng. Ind. Aerodyn.* 97(7–8): 328–334
- [8] Narasimha R and Shrinivasa U 1984 Specification of design wind loads in India. *Sadhana* 7(Part 3): 259–274
- [9] Rao G V N 1988 A survey of wind engineering studies in India. *Sadhana*. 12(Parts 1 & 2)
- [10] AS/ NZS: 1170.2:2011 Structural design actions, Part 2: Wind actions. Australian/ New Zealand Standard. 2011
- [11] ASCE: 7-02 Minimum Design Loads for Buildings and Other Structure. 2002
- [12] BS 6399-2:1997 British standard: loading for buildings part 2. Code of practice for wind loads, British Standard Institution, London. 1997
- [13] EN 1991-1-4:2005/AC: 2010(E) European Standard Eurocode 1: actions on structures-part 1-4: General actions -wind actions, European Committee for Standardization (CEN). Europe. 2010
- [14] IS: 875- Part-3 Code of Practice for Design Loads (other than earthquake loads) for Buildings and Structures- Wind Loads. India. 2015
- [15] Mooneghi M A and Kargarmoakhar R 2016 Aerodynamic mitigation and shape optimization of buildings: review. *Journal of Building Engineering.* 6: 225–235. <https://doi.org/10.1016/j.jobe.2016.01.009>
- [16] Lam K M, Zhao J G and Leung M Y H 2011 Wind-induced loading and dynamic responses of a row of tall buildings under strong interference. *J. Wind Eng. Ind. Aerodyn.* 99: 573–583
- [17] Amin J A and Ahuja A K 2012 Wind-induced mean interference effects between two closed spaced buildings. *KSCE J. Civil Eng.* 16(1): 119–131
- [18] Hui Y, Tamura Y and Yoshida A 2012 Mutual interference effects between two high-rise building models with different shapes on local peak pressure coefficients. *J. Wind Eng. Ind. Aerodyn.* 104–106: 98–108
- [19] Bairagi A K and Dalui S K 2014 Optimization of interference effects on high rise buildings for different wind angles using CFD simulation. *Electronic J. Struct. Eng.* 14
- [20] Mara T G, Terry B K, Ho T C E and Isyumov N 2014 Aerodynamic and peak response interference factors for an upstream square building of identical height. *J. Wind Eng. Ind. Aerodyn.* 133: 200–210
- [21] Yu X F, Xie Z N, Band Zhu J and Gu M 2015 Interference effects on wind pressure distribution between two high-rise buildings. *J. Wind Eng. Ind. Aerodyn.* 142: 188–197
- [22] Zu G B and Lam K M 2018 Across-wind excitation mechanism for interference of twin tall buildings in tandem arrangement. *Wind Struct.* 26(6): 397–441
- [23] Zu G B and Lam K M 2018 Across-wind excitation mechanism for interference of twin tall buildings in staggered arrangement. *J. Wind Eng. Ind. Aerodyn.* 177: 167–185
- [24] Sanyal P and Dalui S K 2018 Effects of courtyard and opening on a rectangular plan shaped tall building under wind load. *Int. J. Adv. Struct. Eng.* 10: 169–188
- [25] Chakraborty S, Dalui S K and Ahuja A K 2014 Wind load on irregular plan shaped tall building—A case study. *Wind Struct.* 19(1): 59–73
- [26] Li Y, Tian X, Tee K F, Li Q S and Li Y G 2018 Aerodynamic treatments for reduction of wind loads on high-rise buildings. *J. Wind Eng. Ind. Aerodyn.* 172: 107–115
- [27] Alminhana W G, Braun L A and Loredou-Souza M A 2018 A numerical-experimental investigation on the aerodynamic performance of CAARC building models with geometric modifications. *J. Wind Eng. Ind. Aerodyn.* 180: 34–48
- [28] Heidari M R, Farahani M, Soltani M R and Taeibi-Rahni M 2009 Investigations of supersonic flow around a long axisymmetric body. *Trans. B Mech. Eng. Scientia Iranica* 16(6): 534–544
- [29] Paul R and Dalui S K 2016 Wind effects on ‘Z’ plan-shaped tall building: A case study. *Int. J. Adv. Struct. Eng.* 8: 319–335

- [30] Gomes M G, Rodrigues A M and Mendes P 2005 Experimental and numerical study of wind pressures on irregular-plan shapes. *J. Wind Eng. Ind. Aerodyn.* 93: 741–756
- [31] Mallick M, Kumar A and Patra K C 2019 Experimental investigation on the wind-induced pressures on C-shaped buildings. *KSCE J. Civil Eng.* 23(8): 3535–3546
- [32] Bhattacharyya B, Dalui S K and Ahuja A K 2014 Wind Induced Pressure on ‘E’ Plan Shaped Tall Buildings. *Jordan J. Civil Eng.* 8(2)
- [33] Cook N J 1985 The designers guide to wind loading of building structures, Part 2: Static structures. Building Research Establishment Report, Butterworths
- [34] Houghton E L and Carruthers N B 1976 Wind forces on buildings and structures: An introduction. Wiley, Hoboken
- [35] Hui Y, Yoshida A and Tamura Y 2013 Interference effects between two rectangular-section high-rise buildings on local peak pressure coefficients. *J. Fluids Struct.* 37: 120–133
- [36] Kim C Y and Kanda J 2013 Wind pressures on tapered and set-back tall buildings. *J. Fluids Struct.* 39: 306–321
- [37] Lam K M, Leung M Y H and Zhao J G 2008 Interference effects on wind loading of a row of closely spaced tall buildings. *J. Wind Eng. Ind. Aerodyn.* 96: 562–583
- [38] Li Y, Li S Q and Chen F 2017 Wind tunnel study of wind-induced torques on L-shaped tall buildings. *J. Wind Eng. Ind. Aerodyn.* 167: 41–50
- [39] Sakamoto H and Haniu H 1988 Aerodynamic forces acting on two square prisms placed vertically in a turbulent boundary layer. *J. Wind Eng. Ind. Aerodyn.* 31: 41–66
- [40] Song J, Tse K T, Tamura Y and Kareem A 2016 Aerodynamics of closely spaced buildings: With application to linked buildings. *J. Wind Eng. Ind. Aerodyn.* 149: 1–16
- [41] Yu X, Xie Z and Gu M 2018 Interference effects between two tall buildings with different section sizes on wind-induced acceleration. *J. Wind Eng. Ind. Aerodyn.* 182: 16–26
- [42] Ming G 2010 Wind-resistant studies on tall buildings and structures. *Sci. China Technol. Sci.* 53(10): 2630–2646



Comparison of Response of Building Against Wind Load as per Wind Codes [IS 875 – (Part 3) – 1987] and [IS 875 – (Part 3) – 2015]

Naveen Suthar and Pradeep Kumar Goyal

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

April 28, 2021

Comparison of response of building against wind load as per wind codes [IS 875 – (Part 3) – 1987] and [IS 875 – (Part 3) – 2015]

Naveen Suthar¹, Pradeep K. Goyal²

¹Post Graduate Student, Department of Civil Engineering, Delhi Technological University (formerly Delhi college of engineering), Delhi, India.

²Associate Professor, Department of Civil Engineering, Delhi Technological University (formerly Delhi college of engineering), Delhi, India.

Abstract

A comparison of wind loads to make a G+11 building in staad and design the building against wind load is presented in this paper. The importance of this study is to calculate the wind load for a structure by the two different code and compare them for better analysis. In present scenario high rise structures have advantages in the populous area and to make more space to live and provide better accommodation in highly populated area around the world. To make the building cost effective and proper design should have to done for more area for living purpose and reduce the cost of structure and safety of structure should be consider in this design. In the recent times, there had been so many catastrophic damages caused by high wind speed in the coastal regions of India which prove that many buildings that are currently in use are not fully wind resistant. In this paper, we have calculated the wind load using static method by the old code [IS: 875 – (Part 3) – 1987] and as per the new code [IS: 875 – (Part 3) – 2015] for zone 4 with terrain category 3 and the building is analyzed using STAAD PRO Software.

they depend on the surface conditions. All the standard wind load codes have their approach to calculate the wind load. they have different formulas and conditions in their map for the calculation of the wind load. For the analysis of wind load, the terrain category 3 has taken for the different wind load and the comparative analysis. STAAD-PRO is very good software for the structure analysis and this software is using by many structural engineers now a days this software can be solved typical problem like static analysis, finite element model, wind analysis and we can also select various load combination in the design by this software to check RCC codes. For the design of beams, columns, lateral bracing and foundations wind loads on the structural frames are required. Wind load is generally taken in to account when the height of building is greater than 150m and low-rise buildings are also affected by wind load. When building is goes increasing, they become flexible and more lateral deflection occur in the building. This paper describes wind analysis of building which is located in zone IV. For the analysis of wind load a twelve-storey building is taken. In this project comparison of result from the IS:875- Part3 (2015) code and IS:875Part3 (1987) are discussed so that we can understand the applicability of wind load analysis using both codes.

I. Introduction

The wind is an important factor in the design of high-rise building. The wind is more important than the earthquake and other important loads. The terrain category is defined according to the roughness and the smoothness of the surface. The wind load is affecting many parameters like construction cost, building strength and another parameter of the building. As per the results which help in the selection of different parameter of the building. Standard codes from different countries use their different terrain categories for the calculation the wind load and

II. Review of literature

high rise structures are currently in demand because of continuously increase in population and technological enhancement as compared to past scenario. In current design practice, the lateral load resisting system of a high-rise building is considered in the design of structure. Structural components such as column, beam, shear wall is considered in the load resisting system of the high-rise building. In the lateral resisting

performance of high-rise building nonstructural component are also consider in the loading. In practice building is system of structural and non-structural but the nonstructural components of the building are considered as non-load bearing component and they are not including in the design of the building.

Kawale and Joshi (2017) analyzed columns, beams, slabs by using IS code [IS: 875 – (Part 3) – 1987] and as per the new code [IS: 875 – (Part 3) – 2015). Thejaswini and Sawjanya (2018) stated the behavior of the junction tower build for the fabric handling purposes in thermal power station subjected to wind load as per IS code [IS: 875 – (Part 3) – 1987] and as per [IS: 875 – (Part 3) – 2015). Sreedharan (2016) comparative study of the seismic and wind analysis for four different structure and three different tracing system are considered for the concentrated load and analyzed. Rajesh et.al., (2016) Found that shear and lateral defection in the building at each story is more at wind load when we compare it to seismic load. higher sections are subjected to high wind so it is good to provide more reinforcement at higher sections to counter the high lateral loads. Mashalkar et al., (2017) studied the effect of wind on different shape as I, C, T and L.

III. METHODOLOGY

RCC framed structure is a combination of beam, column, slab in which beam, column, slab and foundation are inter connected to each other. load transfer of building to soil is through foundation so the foundation must be strong. In frame structure, Load transfers from the slabs to beams, and beams to columns and finally to the foundation.

Bearing walled building is 10 to 12 percent of total framed structure. Monolithic construction is done with R.C.C framed structures. monolithic buildings can easily resist vibrations, wind loading. Load bearing walled can effectively resist earthquake.

Assumptions in Design:

- Using partial factor of safety for loads as 1.5 (as per clause 36.4 of IS-456-2000).
- Partial factor of safety is taken as 1.5 and 1.15 for concrete and steel respectively.

These are the load combinations, which are considered in the design of structures (as per IS 456-2000).

(i) $1.5 \times (\text{Dead load} + \text{Live load})$

(ii) $1.2 \times (\text{Dead load} + \text{Live load} + \text{Wind load})$

When wind load acting in X direction, load combination is considered as $1.2(D+L+W_{in X +ve})$ and wind load acting in Z direction load, the combination of load will be $1.2(D+L+W_{in Z +ve})$.

Therefore, three load combination are considered in this study.

Wind load calculation as per IS 875 part3 (1987):

The design wind speed (V_z) is obtained as per formula given below:

$$V_z = V_b k_1 k_2 k_3 \quad (1)$$

where

V_z = design wind speed at any height z in m/s,

k_1 = probability factor (risk coefficient)

k_2 = terrain, height and structure size factor

k_3 = topography factor

The design wind pressure at height z can be calculated as

$$P_z = 0.6 (V_z)^2 \quad (2)$$

where,

P_z = design wind pressure in N/m² at height z ,

V_z = design wind velocity in m/s at height z .

The total Wind load (F) on particular building or structure is calculated as

$$F = C_f A_e P_z \quad (3)$$

Where,

A_e = effective frontal area

C_f = force coefficient depends upon shape of element plan size & wind dir.

P_z = design wind pressure in N/m² at height z ,

Wind load calculation as per IS 875 part3 (2015)

The design wind speed (V_z) is obtained as per formula given below:

$$V_z = V_b k_1 k_2 k_3 k_4 \quad (4)$$

where

V_z = design wind speed at any height z in m/s,

k_1 = probability factor (risk coefficient)

k_2 = terrain, height and structure size factor

k_3 = topography factor

k_4 = importance factor for the cyclonic region

The basic wind pressure at height z can be calculated as

$$P_z = 0.6 (V_z)^2 \quad (5)$$

where,

P_z = basic wind pressure in N/m² at height z,
 V_z = design wind velocity in m/s at height z.

The basic wind pressure at height z can be calculated as

$$P_d = K_d \cdot K_a \cdot K_c \cdot P_z \quad (6)$$

where,

P_d = design wind pressure in N/m² at height z,
 K_d = wind directionality factor
 K_a = area averaging factor
 K_c = Combination factor

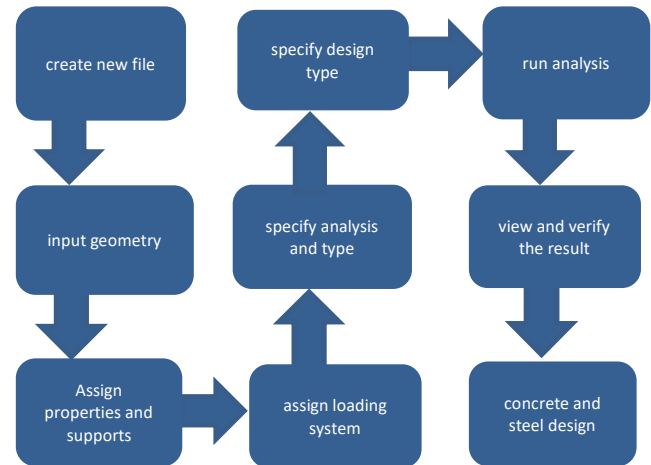
The total Wind load (F) on particular building or structure is calculated as

$$F = C_f \times A_e \times P_d \quad (7)$$

Where,

A_e = effective frontal area
 P_z = design wind pressure in N/m² at height z,
 C_f = force coefficient depends upon shape of element plan size & wind dir.

All the steps are shown in this figure



IV. NUMERICAL STUDY

In this study, a G+11 story building situated in Delhi is considered for comparison of response of building against wind load. The details of building is given in Table 1.

STEPS FOR ANALYSIS OF BUILDING USING STAAD.Pro:

- 1: First, we create nodal point according the dimension, according to the plan we entered the position of plan of building in to the STAAD pro software.
- 2: By using add beam command we add the beam between nodes for beam and column.
- 3: To visualize the 3D view of structure, we simply add transitional repeat command.
- 4: After the completion of structure, we assign support at the bottom as a fixed support. also, we assign material and beam and column dimension.
- 5: Wind loads are calculated as per IS 875 PART 3 and exposure factor is taken as 1. Then wind load is added in load case details in +X, +Z directions.
- 7: Dead loads are calculated as per IS 875 PART 1, including self-weight of structure for external walls, internal walls.
- 8: Live loads are taken as per IS 875 PART 2 and assigned for each floor as 3 KN/m².
- 9: After assigning all the loads, the load combinations with suitable safety factor are taken as per IS 875 PART 5.
- 10: After completed all the steps we have performed the analysis and checked for errors.
- 11: Design of concrete and steel, concrete and steel design are performed as per IS 456: 2000 after the design process, again we performed an analysis for any errors.

Table 1: Building details:

No. of storey	G+11
Size of Column	350 mm × 350 mm
Size of Beam	300mm × 0.500mm
Size of Slabs	150 mm
Live load on slab	3 KN/m ²
Floor finish	3 KN/m ²
Concrete grade in column	M 25
Concrete grade in beam	M 25
Steel grade	Fe 415
Total height of building	36 m
ground storey height	3 m
Height of each floor	3 m
Spacing of frame along length and along width	4m
Thickness of external wall	230 mm

The building, which is considered situated in Delhi. As per IS per code, parameters are given in Table 2.

Table 2: Design Parameter

Basic wind speed	47
zone	IV
city	Delhi
terrain category	3
class	B

Values shown in Table 1 and Table 2 are used for input in the STAAD-Pro software for making the elevation and plan of building and design.

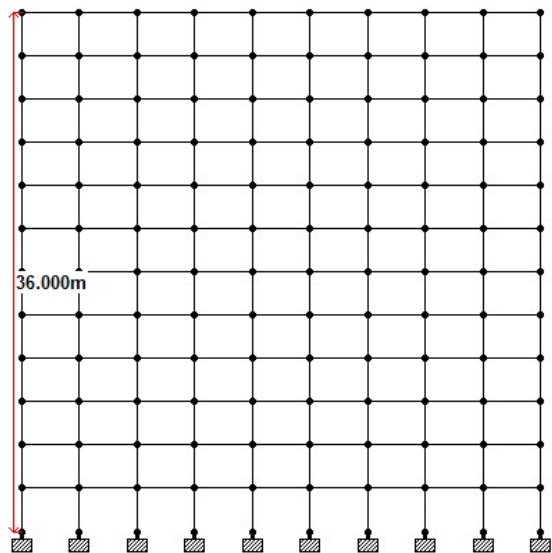


Figure 1: elevation

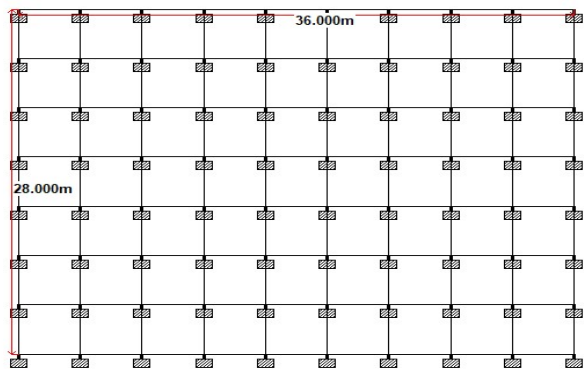


Figure 2: plan

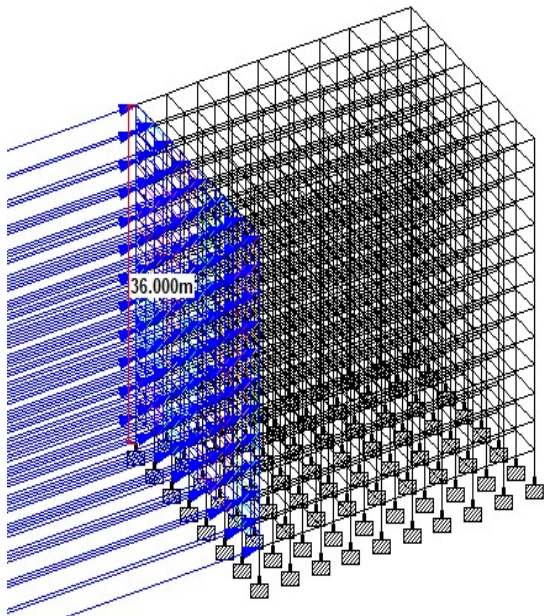


Figure 3: wind load acting in x direction

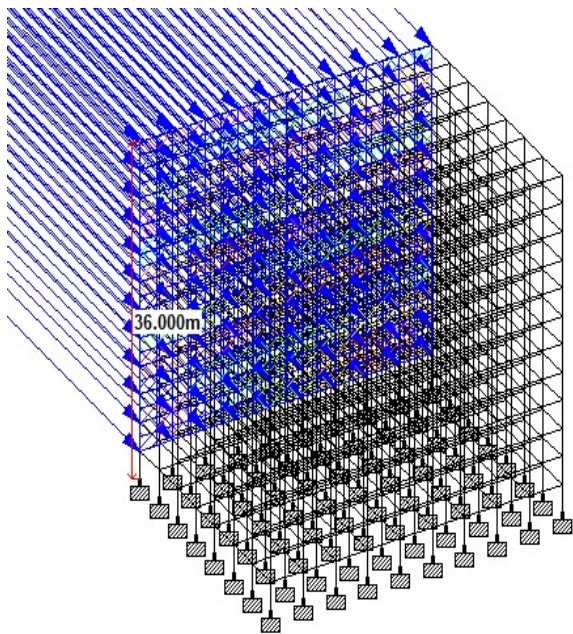


Figure 4: wind load acting in z direction

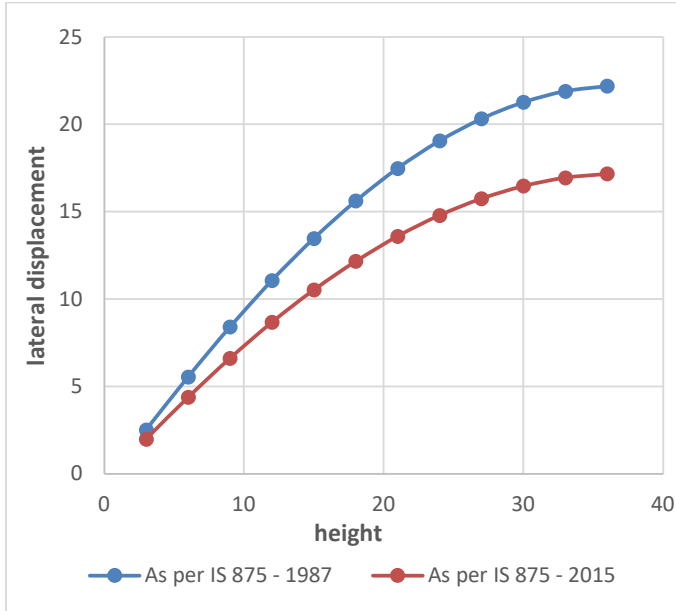


Figure 5: Comparison of Lateral Displacements at different height in x direction

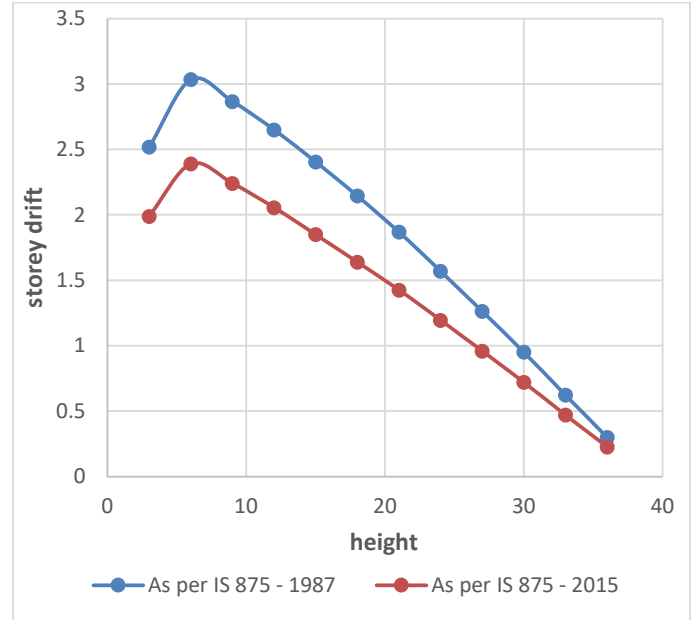


Figure 7: Comparison of storey drift at different height in x direction

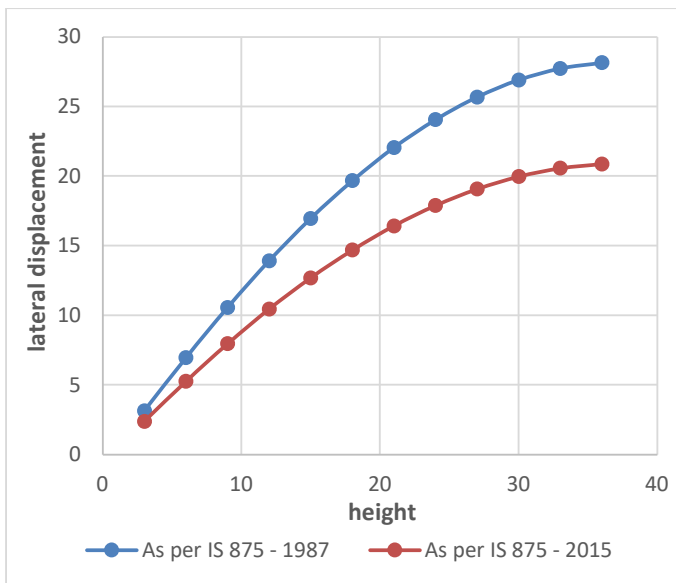


Figure 6: Comparison of Lateral Displacements at different height in z direction

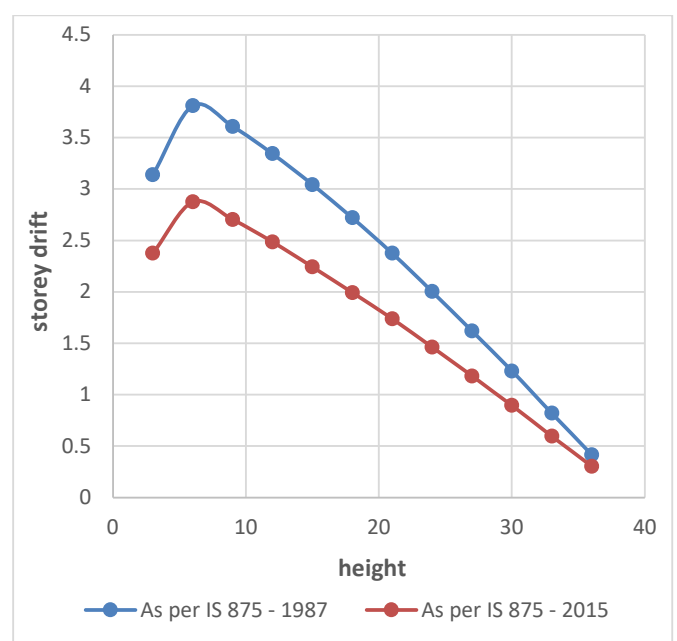


Figure 8: Comparison of storey drift at different height in z direction

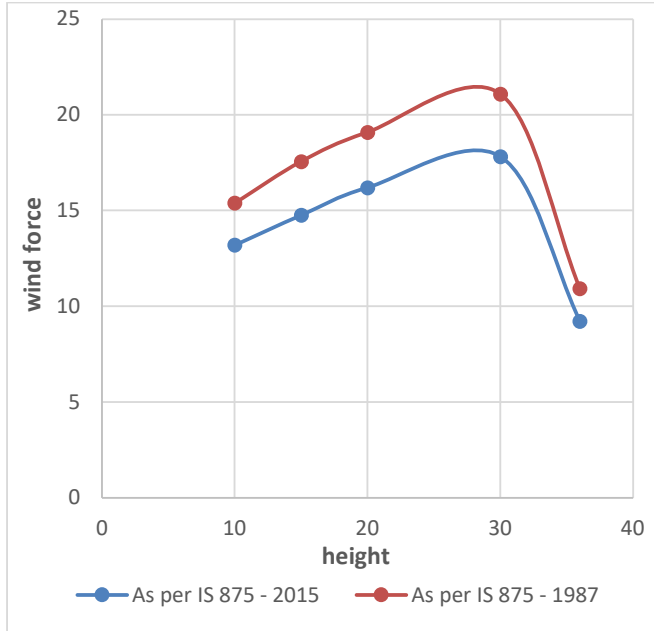


Figure 9: Comparison of wind force at different height in x direction

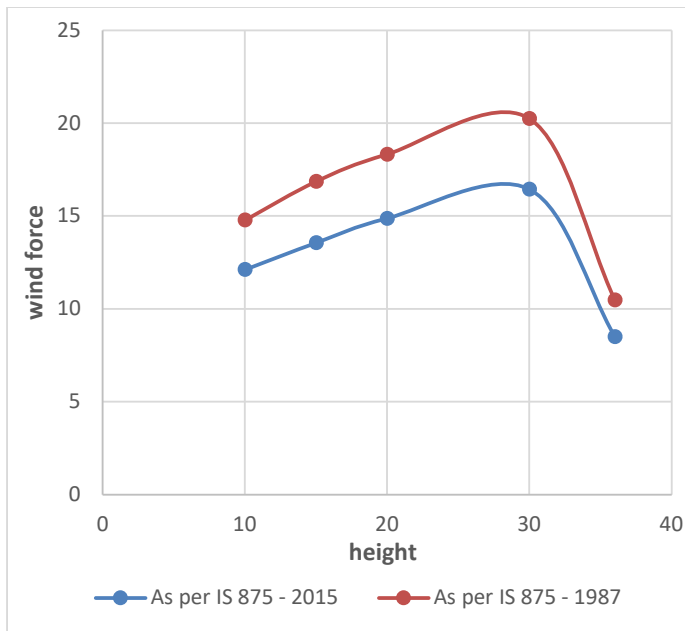


Figure 10: Comparison of wind force at different height in z direction

Table3: Comparison of Lateral Displacements at different height in x direction

Height (m)	Deflection in mm		
	As per IS 875 - 1987	As per IS 875 - 2015	% defection decrease
3	2.518	1.99	20.96902
6	5.553	4.38	21.12372
9	8.418	6.622	21.33523
12	11.068	8.678	21.59378
15	13.473	10.529	21.85111
18	15.618	12.169	22.08349
21	17.487	13.596	22.25081
24	19.057	14.792	22.38023
27	20.32	15.752	22.48031
30	21.271	16.474	22.55183
33	21.895	16.946	22.60333
36	22.196	17.174	22.6257

In table no.3 these are Deflection at each story in x direction, as height increases wind load increases and defection at storey increases. comparison of deflection as per IS 875 – 1987 and as per IS 875 – 2015 mention in above table in x direction.

Table4: Comparison of Lateral Displacements at different height in z direction

Height (m)	Deflection in mm		
	As per IS 875 - 1987	As per IS 875 - 2015	% defection decrease
3	3.14	2.376	24.33121
6	6.952	5.251	24.46778
9	10.562	7.956	24.67336
12	13.907	10.443	24.90832
15	16.951	12.687	25.15486
18	19.672	14.681	25.37109
21	22.049	16.42	25.5295
24	24.054	17.885	25.64646
27	25.674	19.067	25.73421
30	26.905	19.963	25.8019
33	27.725	20.56	25.8431
36	28.143	20.864	25.86434

In table no.4 these are Deflection at each story in z direction, as height increases wind load increases and deflection at storey increases comparison of deflection as per IS 875 – 1987 and as per IS 875 – 2015 mention in above table in z direction.

Conclusion

- The maximum deflection in the top most storey is 22.196 mm for structure which is designed as per Old IS code and 17.174 mm in case of structure which is designed as per new IS Code in x dir.
- The maximum deflection in the top most storey is 28.143 mm for structure which is designed as per Old IS code and 20.864 mm in case of structure which is designed as per new IS Code in z dir.
- Wind force has been decreased as per the new code [IS: 875 (Part 3) 2015]. Percentage decreased is 15.56% along “X” direction and 18.87% along “Y” direction.
- Displacement for the top most storey of G+11 storey building as per new code 22.62% as been decreased along “X” direction and along “Z” direction as per new code 25.86% as been decrease in new code when compared with old code.
- Storey drift for the top most storey of G+11 storey building as per new code 6.37% along “X” direction as been decreased and along “Y” direction as per new code 27.09% as been decreased in new code when compared with old code.
- From the above results it can be concluded that new IS Code [IS: 875 – (Part 3) – 2015] will provide high safety to the structure for static analysis as compared to Old IS Code also structure is economical that designed as per [IS: 875 – (Part 3) – 2015].
- Lateral deflection at each storey shall not exceed 0.002 times the storey height and all the lateral displacement at each story is under permissible limit.

References

- [1]. IS 875: (1987) part3 Indian Standards Code of Practice for Design Loads for Buildings and Structures Part.3 - Wind Loads. Bureau of Indian Standards, India.
- [2]. IS 875: (2015) part3 Indian Standards Code of Practice for Design Loads for Buildings and Structures Part.3 - Wind Loads. Bureau of Indian Standards, India.
- [3] Saurabh Kawale, Dr. S.V. Joshi, Department of Civil Engineering “Analysis of High-Rise Building for Wind Load” International Journal for Scientific Research & Development| Vol. 5, Issue 03, 2017 | ISSN (online): 2321-0613.
- [4] Thejaswini, Dr sawjanya, Department of Civil Engineering “comparative study of old or new code for high Rise Building for Wind Load” International Journal for Scientific Research & Development| Vol. 7, Issue 11, Nov 2018 | ISSN (online): 2278-0181.
- [6] Prof. Sarita Singla, Taranjeet Kaur, Megha Kalra and Sanket Sharma, Civil Engineering Department, Chandigarh, India. “Behaviour of R.C.C. Tall Buildings Having Different Shapes Subjected to Wind Load”. Conf. on Advances in Civil Engineering 2012 DOI: 02. AETACE.2012.3.17.
- [7] B. S. Mashalkar “Effect of Plan Shapes on the Response of Buildings Subjected to Wind Vibrations” IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE) e-ISSN: 2278-1684, p-ISSN:2320-334X PP80-89.
- [8]. Megha Kalra, Purnima Bajpai and Dilpreet Singh. Effect of Wind on MultiStorey Buildings of Different Shapes Indian Journal of Science and Technology, Vol 9(48), DOI:10.17485/ijst/2016/v9i48/10570, December 2016.
- [9]. Higgins, Theodore R. 1979 “Structural Design of Tall Steel Buildings: Council on Tall Buildings and Urban Habitat” Vol. Sb. American Society of Civil Engineers.
- [10]. Vikrant Trivedi “Wind Analysis and Design of G+11 Storied Building Using STAAD-Pro” International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 03 Mar-2018.
- [10]. Md Ahesan, Md Hameed “Comparative Study on Wind Load Analysis Using Different Standards” Vol.7, Special Issue 3, March 2018.
- [11]. Alkesh Bhalerao (2016) “Effect of structural shape on wind analysis of multi storied RCC structures”.
- [12]. Shaikh Muffassir (2016) “Comparative Study on Wind Analysis of Multi-story RCC and Composite Structure for Different Plan Configuration”.

Data Preprocessing based Connecting Suicidal and Help-Seeking Behaviours

Aayush Mittal

Department of Computer Science
and Engineering
Delhi Technological University
aayushmittalhsr@gmail.com

Abhishek Goyal

Department of Computer Science
and Engineering
Delhi Technological University
abhishek.kiitworld@gmail.com

Mohit Mittal

Department of Computer Science
and Engineering
Delhi Technological University
mohitmittal.feb@gmail.com

Abstract—Everyone knows that although there are so many researchers and suicide prevention teams that are active out there to help individuals with mental health problems, many cases of suicide are not able to be detected. Social media acts as a platform for users to express their views online on the internet. How can we help such suicide prevention teams using social media to identify people where there is a possibility that they may develop suicidal ideation thoughts? The main goal of this project is to identify those people from social media who may have suicidal thoughts or may develop suicidal thoughts after some time. We are using the Reddit dataset to find the people who may have had thoughts about suicide before and those people who have other problems like depression or anxiety and express their views by posting in subreddits. We all know that any machine learning algorithm of any classification algorithm cannot work well on a completely raw dataset. So, our next step is doing the preprocessing of the data to extract the data that is relevant for us out of the user's posts for our classification algorithms. Then after that, we will perform the support vector machine (SVM) algorithm to classify the users based on their subreddits that they have posted.

Keywords: Suicidal Ideation; Data Preprocessing; Lemmatization; Stemmatization; Textual Classification; Support Vector Machine (SVM); Multinomial Naïve Bayes; Random Forest.

I. INTRODUCTION

Social media is a great contributor in generating the content that is uploaded by the users on the topics that are related to suicide, depression, etc. Studying and researching this data helps to identify the people that exhibit or can exhibit mental or health-related problems in the future that may awaken the thought of suicide in a person's brain. Approximately 80% of people who have these illnesses start developing these types of thoughts that are of no use, they can't do anything in their lives and their life is totally worthless and they express these thoughts on the internet through social media. If we develop a process or a system that rates a person by giving him/her a certain score by reviewing the content posted and then finding how much chances are there that the user has suicidal thoughts or may develop suicidal thoughts in the future, then such users could be recommended for consulting a doctor or take help from some psychologists.

The organizations that are working in this field like giving counseling to the people who are depressed or are suffering from some other mental illness would also be benefitted. But there is a risk that there may be some privacy or ethical issues in this system. Any data or information related to a user used to detect the transitions can be misused by someone else or

even the system can misinterpret the data of the user because we all know that achieving an accuracy of 100% is impossible..

A shocking fact is that suicide is the 18th leading cause of deaths all across the globe. An approximate 8 lakhs deaths happened in 2017. World Health Organization (WHO) provides a report on attempts for suicide and also has added that if a person commits suicide then it leads to around 20 more deaths. Suicidal ideation basically is used when someone starts talking about depressing things or starts developing suicidal thoughts. Social media is one of the best sources for determining suicidal ideation. Thus, a lot of researchers are motivated to find trends or analyze the data that is related to suicide and present already on different platforms to understand the causes of it and then find a solution for it.

The works that were carried out before by the researchers, focused mainly on finding the words that are related to suicide or suicidal thoughts using three main predictors namely linguistic structure, interpersonal awareness, and interaction. This research helped in finding a lot of mental illnesses like depression, anxiety, stress, etc. The researchers also added a logistic regression model on top of these predictors that helped in improving the results to around 80% accuracy. But the main problem was that although the accuracy was quite good, the results don't tell anything about the main question that is: Are the detected people actually having depression or any suicidal feelings?

In this paper, we are using various classifier algorithms such as support vector machine to solve the problem. The words and phrases in the document that convey the similar meaning are classified into topics like "Individuals having general issues" or "Individuals having suicidal ideation". But still it will provide probabilistic values for each topic that gives document as a mixture of topics rather than being firm on just one particular topic. We perform our analysis on the Reddit dataset. It provides user information such as "user_id", "title", "post_description" and other metadata. It also provides subreddits like "SuicideWatch" and "GeneralIssues" to classify users based on their posts.

We extracted data in 2 time frames: First time frame was from 10-2-2016 to 12-10-2016. We extracted around 22,000

users with approximately 70,000 raw posts. From those 22,000 users, 13,000 users were just subscribers of "GeneralIssues" subsections while remaining 9,000 users subscribed to "SuicideWatch" subsection. From those 9000 users, 869 users were common to GeneralIssues. Second time frame was from 13-10-2016 to 10-02-2017. This was used for the testing dataset. We extracted around 12,000 users with approximately 33,000 raw posts. From those 12,000 users, 7,000 users were just subscribers of "GeneralIssues" subsections while remaining 5,000 users subscribed to "SuicideWatch" subsection. From those 5000 users, 385 users were common to GeneralIssues. The following sections after this are: Section II, in this section we discuss various related works already there. In Section III, we discuss about how we extracted the data and other steps to clean the data. In Section IV, we discuss different research methods that we used. In Section V we have discussed about the results. Then in Section VI we have written the conclusion.

II. RELATED WORK

Here are some research papers that have been written on suicidal ideation.

We all know that internet provides a lot of content on the topics that include different mental health issues related topics like depression, anxiety, trauma etc. Different users on different platforms like facebook, twitter, reddit, instagram etc. post a lot of content which we can use in our research. [1] Choudhury et al. study suicidal expression on Reddit is one such study that focuses on two main topics: first topic is Mental Health (MH) and the second topic is suicide watch (SW). The first topic that is Mental Health focuses on the topics that are related to general mental health like depression, anxiety etc. This is quite general and not as worrying as the second topic that is suicide watch, it mainly focuses more on the topics like suicidal ideation, the main motive is to help those who are having suicidal thoughts by different methods, one such method written in this is psychological therapy. It is written that in the research it was found that those who visited Mental Health section eventually went to Suicide Watch section as well. They created a model which would predict the main reason of the why an individual develops suicidal thoughts but the dilemma was that they were not sure whether an individual is having suicidal thoughts or not. Then from the paper of [6] Chung and Pennebaker, the most important 3 measured were used for analysis namely Linguistic, interpersonal, and interaction. A 4th extra feature was also added that was content which was divided into unigrams and bigrams. Then, logistic regression was used by them to predict whether the point that they were considering, that an individual who visits Mental Health section may also visit Suicide watch section is correct or not. We took our idea of research from this paper to a great extent. [2] Guntuku et al. study focuses on detecting the signs of suicidal thoughts such as depression, mental health issues, anxiety etc from the different online platforms like twitter, instagram, facebook, reddit. They use text presents from this dataset on making n-grams, semantic

request and word check (LIWC), and feeling highlights. These highlights help them to make a directed model utilizing logistic regression and a model utilizing a support vector machine. They grouped the content posts dependent on misery or psychological maladjustment related points. They spoke to results utilizing Receiver Operating Characteristic(ROC) bends with an exactness of 72%. Information extraction and investigation on this paper intently identifies with the examination conveyed in this specific task between help-chasing practices and self-destructive contemplations.[3] Leite et al. is another article that talks about suicidal ideation. Thinking about death or such thoughts that may lead to death as the last option is suicidal ideation. The writers consider suicidal ideation as a very vital component that will tell whether a person having depressive thoughts will eventually commit suicide or not. Then they did a research and analysis on how many percent of deaths that happen are due to suicide. It was found that people in the age group 15 to 25 are most vulnerable and commit more suicides all over the world. So, population age was considered the second component after suicidal ideation. Another feature that the authors used is social isolation. Social isolation basically means that an individual does not want to participate and interact in any social activity, it may be due to lack of expressivity in an individual. Social isolation may lead to a feeling of loneliness or left out in an individual which may lead to the development of suicidal thoughts. The authors said in the article that interacting on social media and expressing thoughts there may help a person to overcome any kind of anxiety or depression. This paper helps in finding out different relations between population age and suicidal thoughts. It helps in distinguishing different age groups that can have suicidal thoughts. [4] Cheng et al. is centered around noticing correspondences identified with self-destructive themes on China explicit online platform named Weibo Suicide Communication (WSC). [5] They overviewed the clients of Weibo and got some information about their inclination towards factors like nervousness, sorrow, stress, and so on that at last prompts performing suicidal ideation on this Chinese Microblog. The Mann-Whitney-Wilcoxon model and the chi-square test model were utilized to separate between WSC people and non-WSC people subsequently giving them a substance of self-destructive correspondence from miniature blogs. In the end, they are attempting to change certain factors and boundaries by making separate models and attempting to discover which one plays out the best to accomplish the last objective of separation. A similar thought of managing web-based media content causes my venture to explore with legitimate utilization of online media content and the highlights that it gives.[7] Pourmand et al. States that adolescents and teens express their thoughts regarding suicide in a much better way on online platforms like facebook, twitter etc. This may be because they feel more comfortable to share their thoughts on social media instead of sharing their thoughts with a doctor or a psychiatrist. Besides, in a cross sectional investigation of 1000 Twitter clients in their 20's, tweeting about self-destructive ideation was fundamentally connected with

evaluations of self-destructive ideation and conduct by means of self-reports, recommending that result measurements assembled through Twitter could address a valuable marker of genuine self-destructive ideation and conduct. Twitter clients who self-distinguish as having schizophrenia are bound to tweet about self destruction content. Besides, Twitter addresses a virtual area where clients are known to make self destruction agreements, looking for different clients to participate in an arranged self destruction attempt. Devotees of major news sources are probably going to retweet content identified with wellbeing and sickness, while content including dysfunctional behavior is related with a higher likelihood of being retweeted. Of all emotional wellness related tweets shared by major news sources, about 30% alluded to suicide. [8] Burnap et al. have utilized Twitter as a mechanism of online media to play out this order examination. They removed some significant catchphrases from Tumblr and other web online journals identified with self-destructive musings and afterward utilized the term frequency/inverse document frequency (TF.IDF) technique to utilize those watchwords and explain them dependent on the most oftentimes utilized words for self-destruction as the connected subject. Given these significant watchwords, they removed those posts from the Twitter dataset utilizing the Twitter API, which is identified with a self-destructive point. After removing the component out of the model, they have utilized order models to be specific support vector machine (SVM), rule-based (decision tree), and Naive Bayes grouping (NB). At long last, they played out a similar examination on the equivalent. This cycle of information extraction and preprocessing is like the Reddit dataset which will be utilized for this task. In this research paper, we will focus on detecting as many suicidal cases as possible by using various classification algorithm so that we can recommend people who are suffering from depression or anxiety to consult a doctor.

III. DATA

1. DATA COLLECTION AND CLEANING: Targeting users with some general concerns and specifically those who express suicidal ideation on massive social media platform is indeed a big challenge. Reddit provides a straightforward way to handle this challenge by providing subreddits, or subforums dedicated to specific topics. First, we will discuss the features and methods of extracting data from Reddit. Following the above, now we will be performing the following preprocessing steps on the selected data set.

2. DATA SELECTION[14]: For our project, we require the posts of the users who reveal some general issues like mental or health-related problems. Also, we would like to extract users who directly express suicidal ideation through their posts. For this purpose, we selected subsections like "r/mentalhealth", "r/depression", "r/trauma", "r/stopsselfharm", "r/survivorsofabuse", "r/rapecounseling", "r/socialanxiety" etc. for targeting users with general issues (henceforth GI). Also, we chose the subsection "r/suicidewatch" for targeting users

who express suicidal thoughts.

3. DATA EXTRACTION[15]: Reddit provides an API to help people extract users' posts and comments for a subreddit. Python Reddit API Wrapper (PRAW) was used for our project, we focused on the year 2016 and got the user information that would help to map users from GI subsections to the SW subsection. A single post from Reddit contains attributes like "user-id", "author name", "subreddit", "post creation time" and finally "post". Python Reddit API Wrapper allows us to extract data according to a given timestamp of the year. For training dataset, we extracted posts within a timestamp of 11-2-2016 to 12-10-2016. We extracted around 22,000 users with approximately 70,000 raw posts. From those 22,000 users, 13,000 users were just subscribers of "GeneralIssues" subsections while remaining 9,000 users subscribed to "SuicideWatch" subsection. From those 9000 users, 869 users were common to GeneralIssues. We extracted the training dataset in two ways. For training *dataset 1* we used a total of 4552 posts out of which 3001 were general issues and 1551 were suicidal. For training *dataset 2* we used a total of 3352 posts out of which 1801 were general issues and 1551 were suicidal. Second time frame was from 13-10-2016 to 10-02-2017. This was used for the testing dataset. We extracted around 12,000 users with approximately 33,000 raw posts. From those 12,000 users, 7,000 users were just subscribers of "GeneralIssues" subsections while remaining 5,000 users subscribed to "SuicideWatch" subsection. From those 5000 users, 385 users were common to GeneralIssues. For final testing dataset we used a total of 4108 posts out of which 3501 were general issues and 607 were suicidal

4. DATA PREPROCESSING: The Raw textual dataset needs some amount of preprocessing before using it as an input for any clustering or classification. Given below are the preprocessing steps:

- A) Split each user's posts into sentences and sentences into words using a technique called tokenization.
- B) Transform all the words to lowercase.
- C) Remove all the punctuations.
- D) Remove all the stop words.

Let's consider an example text that would help us identify why those preprocessing steps are crucial before performing any classification algorithm. One of the posts that we extracted looks something like this: *"I've been feeling depressed on and off for about 2 years, recently there has been more triggers, my anxiety ticks have come back and the depression comes more often (last 2 weeks, every day). The depression gets worse every time, I've read so many suicide stories, ways to do it, etc. but I doubt I would do it but I haven't got that deep yet."* After performing tokenization, removing punctuation and converting into lowercase, we get the given below post: *"feelings depressed recently anxiety come back depression comes often last everyday depression worse every time read many suicide ways doubt would get deep yet"*.

E) Convert all the words in third person format to the first person format and all the verbs in the future and past tense to present tense. This process is called Lemmatization. We have

used wordnet lemmatizer for the same.

F) Reduce all the words to their root form. Convert all the words ending with "ing". This process is called Stemmatization. We have used Porter Stemmer for the same. The main purpose of Porter stemmer is to remove the common grammatical and inflexional endings from English words. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems.

Next, we perform lemmatization and stemming and finally we get the cleaned post-show below: *"feel depressed recent anxiety come back depress come often last everyday depress worse every time read many suicide way doubt will get deep yet"*. As you can see, words in the third person are changed to the first person, and verbs in past or future tense are converted to present. Hence, depressed is converted to just depress. Also, words are converted to their root form. This preprocessing gets rid of redundant words and also combines words that convey the same meaning.

IV METHODS

In our research, we used four types of classifiers so as to detect as many suicidal posts as possible. We divided our training set into two parts: training set1 and training set2. In training set1, the suicidal posts were 1551 and the general posts were 3001, and in training set2 suicidal posts were 1551 only but the general posts were reduced to 1801. We did this in order to see what effect does this have on the accuracy of the model. So, the classifiers that we used are:

Random Forest Classifier: Random Forest is an ensemble classifier made using many decision trees. Ensemble model combines the results from different models. This is a versatile algorithm which can be used both for classification and regression problems. It is one of the most commonly used predictive modeling and machine learning technique. The classifier takes the average of all the prediction so it rules out the chance of over fitting. It uses bagging while creating decision trees so as to reduce the correlations between them. The first step in random forest is that it will divide the dataset into smaller subparts. Every subset need not be distinct, some subsets may be overlapped. The model runs efficiently on large databases and requires almost no input preparation. It is one of the most accurate learning algorithm.

Stochastic Gradient Descent (SGD)[12]: In order to optimize neural networks, an optimization algorithm, or an optimizer is used. There are several optimizers available for this task and Stochastic Gradient Descent, or SGD, is one of the most popular and common optimization algorithm. The gradients of the parameters of neural network are calculated using the Back propagation algorithm and used in the SGD algorithm to update parameters. SDG is an alternative to the standard gradient descent algorithm and is often viewed as a stochastic (or randomized) approximation of the gradient descent. It is

computationally efficient than the standard gradient descent, which makes is the best choice for large scale learning on huge datasets.

Multinomial Naïve Bayes Algorithm: [11]Naïve Bayes is a machine learning algorithm you can use to predict the likelihood that an event will occur given evidence that's present in your data. There are 3 types of Naïve Bayes Model: Multinomial, Bernoulli, Gaussian. We have used Multinomial Naïve Bayes Model in our research work which is preferred when the features used in our dataset describe discrete frequency counts. The model is based on the assumption that past conditions still hold, because if we make predictions from historical values, we will get incorrect results if the present circumstances have changed. The algorithm is based on Baye's Theorem that states:

$$P(M/N) = [P(N/M) * P(M)] / P(N) \text{ where,}$$

$P(M/N)$: Probability that event M will occur given that event N has already taken place.

$P(N/M)$: Probability that event N will occur given that event M has already taken place.

$P(M)$: Probability that M will take place

$P(N)$: Probability that N will take place

Support vector machines[13] popularly known as SVM is a supervised learning algorithm which is mostly used for regression and classification problems as support vector regressor(SVR) and support vector classifier(SVC). Generally SVM is used with datasets of small size as it svm requires large computational powers and a long time to process.The key idea behind SVM is to find a hyperplane that most appropriately separates the features into different classes. Kernel trick is the reason for the popularity of the svm. Kernel trick is method by which we calculate the scalar product of two vectors in some feature space, it is the reason for kernel functions to be called as generalized scalar product. While using the kernel trick we just replace the kernel function in place of scalar product of vectors. Gaussian RBF(Radial Basis Function) is one of the most famous kernel function which is generally applied in SVM models. The value of the RBF kernel depends on the distance from the origin or from some reference point. The format of the RBF kernel is as shown.

$$K(X_1, X_2) = \text{exponent}(-\gamma \|X_1 - X_2\|^2)$$

$$\|X_1 - X_2\| = \text{Euclidean distance between } X_1 \text{ \& } X_2$$

We calculated the dot product i.e. similarity of X1 and X2 using the original space distance.

V. RESULTS

We first trained our classifiers on training dataset1. Then we used them on our testing dataset to get the results. The results of various classifiers for training dataset1 are:

- **Random forest classifier:** The random forest classifier predicted with an accuracy of 85.224%. But the main thing to note here is that the correct suicidal cases detected by random forest are 0. Although the accuracy is quite good, but our motive to identify as many suicidal cases as possible is not accomplished. Here is the confusion matrix obtained from random forest:

	Actual General Issues 3501	Actual Suicidal Ideation 607
Predicted General Issues 4108	3501	607
Predicted suicidal ideation 0	0	0

Fig I: Confusion matrix for random forest classifier for training set1

- **Stochastic Gradient Descent(SGD) classifier:** The SGD classifier predicted with an accuracy of 85.17%. But the numbers of correct suicidal cases detected by it are 0. Here is the confusion matrix obtained:

	Actual General Issues 3501	Actual Suicidal Ideation 607
Predicted General Issues 4106	3499	607
Predicted suicidal ideation 2	2	0

Fig II: Confusion matrix for SGD classifier for training set1

- **Multinomial Naïve Bayes Classifier:** The multinomial Naïve Bayes classifier predicted with an accuracy of 84.98%. But again, although the accuracy is quiet good, the numbers of correct suicidal cases detected by it are only 9. Here is the confusion matrix obtained from the Naïve Bayes Classifier:

	Actual General Issues 3501	Actual Suicidal Ideation 607
Predicted General Issues 4080	3482	598
Predicted suicidal ideation 28	19	9

Fig III: Confusion matrix for Naïve Bayes classifier for training set1

- **Support Vector machine(SVM) classifier:** The SVM classifier predicted with an accuracy of 83.39%.The number of correct suicidal cases detected by it are 32.

This is so far the best classifier for us, although the accuracy of this classifier is least among all the classifiers but still it has detected maximum number of correct suicidal cases. Here is the confusion matrix obtained:

	Actual General Issues 3501	Actual Suicidal Ideation 607
Predicted General Issues 3869	3394	575
Predicted suicidal ideation 139	107	32

Fig IV: Confusion matrix for SVM classifier for training set1

Performance Measure	Random Forest	SGD	Multinomial Naïve Bayes	SVM
Accuracy	0.852	0.851	0.849	0.833
Precision	0.852	0.852	0.853	0.855
Recall	1	0.999	0.994	0.969
F1score	0.92	0.919	0.918	0.908

Fig V: Performance measures for dataset 1

From the above results, we got to know that although the accuracy is quite good, but it is not serving our main purpose that is to detect as many suicidal cases as possible. So we used a training dataset2 which had 1801 general posts and 1551 suicidal posts. We reduced the general posts so that the number of suicidal posts is almost equal to the number of general posts and our training dataset becomes much more balanced. This technique is also known as undersampling. Our dataset is divided into two classes: general posts and suicidal posts. [10] Undersampling technique basically deletes the examples from the class that has more examples than other class so as to balance the dataset. So we have removed the general posts and reduced them from 3501 in training set1 to 1801 in training set2.

After training our classifiers on training dataset2, we used them on our testing dataset to get the results. The results of various classifiers for training dataset2 are:

- **Random forest classifier:** The random forest classifier predicted with an accuracy of 76.266%. Although the accuracy is lesser than the accuracy on training set1, but the number of correct suicidal cases detected by it are 164 as compared to 0 in training set1. So this is a significant increase and serves our purpose better. Here is the confusion matrix obtained:

	Actual General Issues 3501	Actual Suicidal Ideation 607
Predicted General Issues 3412	2969	443
Predicted suicidal ideation 696	532	164

Fig VI: Confusion matrix for random forest classifier for training set2

- Stochastic Gradient Descent(SGD) classifier: The SGD classifier predicted with an accuracy of 67.81%. Also, the numbers of correct suicidal cases detected by it increased to 254. Here is the confusion matrix obtained:

	Actual General Issues 3501	Actual Suicidal Ideation 607
Predicted General Issues 2885	2532	353
Predicted suicidal ideation 1223	969	254

Fig VII: Confusion matrix for SGD classifier for training set2

- Multinomial Naïve Bayes Classifier: The multinomial Naïve Bayes classifier predicted with an accuracy of 73.09%. But again, although the accuracy is quiet not good, but it serves our purpose to detect as many suicidal ideations as possible. The numbers of correct suicidal cases detected by it are 189. Here is the confusion matrix obtained from the Naïve Bayes Classifier:

	Actual General Issues 3501	Actual Suicidal Ideation 607
Predicted General Issues 3231	2813	418
Predicted suicidal ideation 877	688	189

Fig VIII: Confusion matrix for Naïve Bayes classifier for training set2

- Support Vector Machine(SVM) Classifier: The SVM classifier predicted with an accuracy of 63.85%. Although this accuracy is very less as compared to other classifiers but the number of correct suicidal cases detected by it are 293. It is best suited for our purpose. Here is the confusion matrix obtained from SVM classifier:

	Actual General Issues 3501	Actual Suicidal Ideation 607
Predicted General Issues 2644	2330	314
Predicted suicidal ideation 1464	1171	293

Fig IX: Confusion matrix for SVM classifier for training set2

Performance Measure	Random Forest	SGD	Multinomial Naïve Bayes	SVM
Accuracy	0.762	0.678	0.73	0.638
Precision	0.87	0.877	0.87	0.881
Recall	0.848	0.723	0.803	0.665
F1score	0.858	0.792	0.835	0.757

Fig X: Performance measures for dataset 2

VI.CONCLUSION

In this report, we represented a way to identify individuals exhibiting suicidal ideation by targeting their posts and sentiments. We picked Reddit dataset since it provides a better way to organize user posts based on subreddits where we picked subreddits like suicide watch and other general issue subreddits. Users that belong to suicide watch subreddit are the ones expressing suicidal thoughts through social media. Thus, using the posts of those users, we identified other users who express general issues and later on express to show suicidal ideation through their posts. We came up with data preprocessing steps performed on the raw posts of the Reddit dataset. After performing the preprocessing steps we used classification techniques like support vector machine with rbf kernel, multinomial Naive Bayes, SGD classification algorithm. We used two types of training sets, one with more general posts and then in second training set we performed undersampling and reduced our general posts to balance the dataset. Although the accuracy that we got in first training set was much better than the second training set, but the number of correct suicidal cases detected by classifiers in training set2 were much more than the training set1. Random Forest Classifier predicted with an accuracy of 85.224% but it was not able to detect as many suicidal posts. The accuracy predicted by SGD classifier was much less than that of random forest classifier but it was able to detect more suicidal posts. The other two classifiers were also able to detect the suicidal posts. Of all the 4 classification algorithms, SVM classifier was able to detect maximum number of suicidal posts. Eventually, we were able to achieve the goal meant for this project of identifying users exhibiting suicidal ideation, but still this does not implicitly means that the users classified to have suicidal thoughts are actually expressing the same. We can use this model to approach those users for some kind of help that would eventually help them to improve the current status and mind set.

VII. ACKNOWLEDGEMENT

We are very thankful to our mentor Dr. Akshi Kumar for constantly helping, supporting and guiding us in our research and providing us feedback time to time that help us improve the quality of our project.

VIII.REFERENCES

- [1] M. Choudhury, E. Kiciman, M. Dredze *et al.*, "Discovering shifts to suicidal ideation from mental health content in social media," in

Proceedings of the CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, 2016, pp. 2098–2110.

[2] S. Guntuku, D. Yaden, M. Kern *et al.*, “Detecting depression and mental illness on social media: an integrative review,” *Current Opinion in Behavioral Sciences*, vol. 18, pp. 43–49, 2017.

[3] B. Leite, V. Amorim, A. Silva *et al.*, “*The influence of social networks in suicidal behavior*,” *International Archives of Medicine*, vol. 8, 2015.

[4] Q. Cheng, C. Kwok, T. Zhu *et al.*, “*Suicide communication on social media and its psychological mechanisms: an examination of chinese microblog users*,” *International Journal of Environmental Research and Public Health*, vol. 12, pp. 11 506–11 527, 2015.

[5] Q. Cheng, S.-S. Chang, and P. Yip, “*Opportunities and challenges of online data collection for suicide prevention*,” *The Lancet*, vol. 379, pp. 53–54, 2012.

[6] C. Chung and J. Pennebaker, “The psychological functions of function words,” *Social Communication*, vol. 1, pp. 343–359, 2007.

[7] Pourmand, A. *et al.* Social media and suicide: a review of technology-based epidemiology and risk assessment. *Telemed. e-Health* **25**, 880–888

[8] P. Burnap, W. Colombo, and J. Scourfield, “Machine classification and analysis of suicide-related communication on twitter,” in *Proceedings of the*

26th ACM conference on hypertext & social media. Association for Computing Machinery, 2015, pp. 75–84.

[9] <https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589>

[10] Lusa *et al.* Joint use of over-and under-sampling techniques and cross-validation for the development and assessment of prediction models.

[11] Eyheramendy, S., Lewis, D.D., Madigan, D.: On the naive Bayes model for text categorization. In: Ninth International Workshop on Artificial Intelligence and Statistics, pp. 3–6 (2003)

[12] Zadrozny and Elkan, “Transforming classifier scores into multiclass probability estimates”, SIGKDD’02, <http://www.research.ibm.com/people/z/zadrozny/kdd2002-Transf.pdf>

[13] B. Schölkopf A. Smola and K.-R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation*, Vol. 10, 1998, pp. 1299-1319.

[14] Kumar, T. Senthil. "Data Mining Based Marketing Decision Support System Using Hybrid Machine Learning Algorithm." *Journal of Artificial Intelligence* 2, no. 03 (2020): 185-193.

[15] Suma, V., and Shavige Malleshwara Hills. "Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics." *Journal of Soft Computing Paradigm (JSCP)* 2, no. 02 (2020): 101-110.

Deep and Shallow Covariance Feature Quantization for 3D Facial Expression Recognition

Walid Hariri, Nadir Farah, Dinesh Kumar Vishwakarma

Abstract—Facial expressions recognition (FER) of 3D face scans has received a significant amount of attention in recent years. Most of the facial expression recognition methods have been proposed using mainly 2D images. These methods suffer from several issues like illumination changes and pose variations. Moreover, 2D mapping from 3D images may lack some geometric and topological characteristics of the face. Hence, to overcome this problem, a multi-modal 2D + 3D feature-based method is proposed. We extract shallow features from the 3D images, and deep features using Convolutional Neural Networks (CNN) from the transformed 2D images. Combining these features into a compact representation uses covariance matrices as descriptors for both features instead of single-handedly descriptors. A covariance matrix learning is used as a manifold layer to reduce the deep covariance matrices size and enhance their discrimination power while preserving their manifold structure. We then use the Bag-of-Features (BoF) paradigm to quantize the covariance matrices after flattening. Accordingly, we obtained two codebooks using shallow and deep features. The global codebook is then used to feed an SVM classifier. High classification performances have been achieved on the BU-3DFE and Bosphorus datasets compared to the state-of-the-art methods.

Index Terms—Facial expression, CNN, Codebook, BoF paradigm, Covariance matrices.



1 INTRODUCTION

Facial expressions reported up to 55% of face to face communication while only 7% and 38% of the emotional expression are allocated to oral language and vocal tone respectively [1]. Therefore, the recognition of human emotion from facial expression images has become a very interesting research field in computer vision and pattern recognition. The majority of this research focuses on recognizing six basic expressions [2], [3], [4] namely: happy (HA), sad (SA), disgust (DI), surprise (SU), fear (FE), and angry (AN), defined by Ekman and Friesen [5] and accepted as universal emotions. Due to the diverse sources of variability in 2D and 3D facial images, FER has proven to be a very difficult task. Such variations can be environment-related (light conditions, occlusions caused by certain objects), subject-related (location variation), and acquisition-related (image size, distortion, vibration, and other imperfections).

Among previous facial expression recognition (FER) works, we distinguish 2D data-based methods [6], [7]. Despite the good performance which has been achieved in 2D FER, it is still a challenging task as it has to deal with two main issues: illumination and pose. Moreover, the texture, image resolution, and color are not necessarily

the same when the images are acquired in unconstrained conditions. In such a situation, 2D FER struggles to achieve high performance.

3D data, on the other hand, rely on facial structure and provide more geometrical characteristics and are less sensitive to light conditions [8], [9] and pose variations [10]. They have the potential to maintain both geometric and topological facial structural details with the depth information. Accordingly, in 3D FER, 3D data can efficiently capture all the facial parts' movement of the face also in unconstrained conditions. This particularity explains the robustness of the 3D modality for FER due to the recent progress of 3D acquisition techniques and low-cost 3D sensors (e.g., Microsoft Kinect, Intel RealSense) [11]. Thereby, various databases for 3D facial expression analysis have appeared and have been used by the research community to evaluate their algorithms. Among these databases, we find BU-3DFE [11], BU-4DFE, and Bosphorus [12] that contained the six basic emotions. Differently, FRGC v2.0 and GAVAB, present a set of expression variations, but not with a regular distribution thereby, they are not recommended for the FER.

3D FER methods in the literature can be carried out using purely 3D data or of its combination with texture and time variation information. 3D face scan can be also mapped into 2D representations such as three normal component maps, curvedness map, Gaussian curvature map, Mean curvature map, geometry map and texture map, etc [13]. 2D-3D multi-modality could give diversity to the extracted features and outperforms single-modality-based methods. Achieving a good performance, however, requires efficient and discriminative features. In the following, we present and discuss the most important existing FER methods in the literature, and their use in multi-modal representation.

- M. Hariri was with the Department of Computer Science, Badji Mokhtar Annaba University, Algeria, BP12, 23000.
E-mail: hariri@labged.net
- M. Farah was with the Department of Computer Science, Badji Mokhtar Annaba University, Algeria.
E-mail: farah@labged.net
- M. Vishwakarma was with the Department of Information Technology, Delhi Technological University, New Delhi-110042, India.
E-mail: dinesh@dtu.ac.in

Manuscript received.

2 LITERATURE REVIEW

In recent years, Convolutional Neural Networks have been widely deployed in a large range of tasks including image classification systems, particularly facial expression classification [4], [14], [15]. The three main advantages of using CNNs for deep learning are the elimination of handcrafted extraction of features, the cutting-edge recognition results, and the ability to use pre-trained networks for other recognition tasks.

Deep CNNs are generally applied to FER using 2D images to learn deeper feature representations of facial expression. Despite their potential, they can't achieve high performances when dealing with considerable illumination and pose variations [16]. To overcome this problem, various approaches have used 3D data for learning CNN, such as Volumetric CNN [17], Field probing neural networks [18], 3D Graph-CNN [19], multi-view CNN [20] and Vote3D [21]. The high cost of these operations and even more the very large point clouds is a bigger challenge. To obviate this drawback, recent methods proposed to normalize the 3D face image to lower dimensions (e.g. 2D depth image, principle curvatures map), all of which are jointly fed into CNN for feature learning and classification. For example, Jan et al. 2018 [22] learned deep CNN features from a 2D texture map and a depth map extracted from 3D face scans, then an SVM is applied for the classification. They thus take advantage of CNN as a deep feature extractor for FER applications. However, normalizing the 3D face scans to lower dimensions may divest the geometrical and topological information. These limitations make the 3D FER a very challenging task. Therefore, the multi-modal 2D+3D has become a frequent approach for FER, commonly used in literature. One of the most efficient approaches that successfully utilized a 2D+3D multi-modal-based system is proposed in [23]. From 3D face scans, the authors extracted six 2D map representations involve texture map, and combined them through feature learning and fusion learning into a single end-to-end training framework. In [24], RGB images are combined with depth maps to a deep CNN from scratch, in addition to transfer learning using two pre-trained models (ResNet50 and VGG-19) as a hierarchical feature representation. From the state-of-the-art review, one can notice that learning a deep model from scratch is not suitable for 3D FER due to the lack of a large amount of data. Alternatively, pre-trained models show very high interest to overcome this problem and it can be applied in two different strategies. In the first strategy, the facial expression images are fed to the pre-trained model, and the fully connected (FC) layers were typically replaced by one or more additional layers then the networks re-trained to adapt the weights of the added layers for FER. This strategy is called fine-tuning because the pre-trained models are adapted to the FER problem. As an example, a supervised fine-tuning on a small dataset is applied in [25] for FER. The second strategy aims to apply the pre-trained models as feature generators, then uses the obtained features extracted from a FC layer or convolution layer. Traditional ML algorithms (e.g. SVM classifier) can be added to the system, trained with the generated deep features to improve the performance and to avoid the over-fitting problem. However, the FC layers of the pre-trained

models are more dataset-specific features (generally pre-trained on ImageNet dataset) which is a very different dataset, thus, this strategy is not suitable for FER in such a case.

On the other hand, multiple pre-trained CNN models of different architecture that represent different feature abstraction levels can be combined using ensemble model. The global decision takes into account the decision of each model by applying a vote or weighted sum operation. For example, weighted prediction scores of the pre-trained VGG-16 and AlexNet models are computed in [26] to classify the expressions. Despite the high performances achieved by these models, the problem of over-fitting is still a grand challenge because the pre-trained models are generally trained with ImageNet dataset which has a lot of data belonging to very different classes, and the performance is highly depend on the ensemble strategy that cannot be meaningful for FER.

3 CONTRIBUTION OF THE PAPER

We propose in this paper a 2D+3D multi-modal approach that performs two deep pre-trained models (AlexNet, VGG-16) applied to generate deep features from 2D (depth and curvature) maps, and handcrafted features extracted from the 3D images. The obtained features are firstly co-varied and normalized using a feature quantization to feed an SVM classifier. Our objectives are threefold: **I)** we avoid the over-fitting that can be occurred after fine-tuning the pre-trained models to small FER datasets. **II)** local features extracted from the 3D images may provide more geometrical and spatial characteristics that could be lacked after mapping the 3D images onto a 2D (depth or curvature) map. **III)** Ensure a high discriminative power from the pre-trained models instead of applying scratch training, which is time-consuming, and is not suitable for small 3D FER datasets.

Accordingly, instead of using the generated deep features directly to feed a ML classifier, we take full advantage of using covariance matrices as local descriptors which have many benefits: **I)** they provide a natural way for fusing multi-modal features that can be of different dimensions. **II)** compact representation since covariance matrices can be computed from different sized regions, the obtained covariance descriptors are of the same size whatever the size of their features. **III)** ability to compare any regions of different sizes.

It is also important to note that the classical use of deep learning approaches mainly focused on handling data in Euclidean space. However, it is not always the case when dealing with other structures such as Symmetric Positive Definite (SPD) matrices (the space is indicated also by Sym_d^+) that deal with non-Euclidean space. Therefore, various methods have been proposed to be able to apply deep learning approaches on non-Euclidean space, i.e. Lie groups [27], SPD manifolds [28] or Grassmann manifolds [29].

In order to use SPD matrices as an input of classical classifiers that usually assume a Euclidean geometry, Harandi et al. 2014 [30] and Jayasumana et al. 2013 [31] have applied a non-linear mapping into a high dimensional space using kernel-based methods via the use of a projection matrix. This solution preserves the SPD structure, however, it is still

incapable to deal with non-linear learning. SPD learning, on the other hand, solves this problem and offers the possibility to introduce a non-linearity during learning SPD matrices in the network. This motivated us to use this strategy to get a high discriminative representation, and to accurately classify the expressions. To have the same dimensionality of the resulting features, we employed a feature quantization to feed an SVM classifier.

The main contributions of this paper are summarized as follows:

- Deep features generated from the last convolutional layer of two modified pre-trained deep CNNs (VGG-16 and AlexNet) using 2D depth and curvature map images.
- Geometrical features are captured from the 3D images.
- The extracted features are embedded into a covariance pooling for the dimensionality reduction.
- To enhance the discrimination power, deep covariance learning is used through two additional layers (BiMap layer + Eigenvalue Rectification).
- Feature quantization of flattened deep and shallow covariance matrices is carried out which takes full advantage of both geometrical features and deep features of the whole face.
- An experiment is performed on public datasets such as BU-3DFE and Bosphorus. Further, the performance is compared to similar state-of-the-art methods and shows the superiority of the proposed methodology.

The rest of the paper is structured as follows, an overview of the proposed method is given in Section 4 including four steps: pre-processing, feature extraction, the deep learning of SPD matrices as well as the shallow and deep Bag-of-Features paradigm. Experimental results and a detailed comparative study are presented in Section 5. The obtained results are analyzed and comprehensively discussed in Section 6. Conclusions end the paper.

4 METHOD

The proposed methodology consists of a deep and shallow features combination based approach as given in Fig. 1. In order to exploit the discriminative power of the deeply learned features using CNN, and the efficiency of covariance descriptors as a compact representation, we propose in this paper a deep and shallow feature combination method using covariance descriptors to handle the problem of FER on the two challenging BU-3DFE and Bosphorus datasets. Shallow features extract spatial and geometrical characteristics from the 3D face images. In the deep stage, we feed the pre-trained CNNs with the normalized face images. To pool the feature maps spatially from the CNN, we propose to use covariance pooling, and then employ the manifold network to deeply learn the second-order statistics. BoF paradigm is then applied to quantize the deep and shallow covariance matrices after flattening. SVM is applied to classify the expressive faces. In the following, we explain each stage of the proposed method from the pre-processing to the classification.

4.1 Pre-processing and data transformation

Data pre-processing aims to eliminate deficiencies like holes, and unnecessary regions such as hair, neck, and clothes from the face surface. We thereby employ some corrections and filters; involving a *smoothing process* that relieves spikes, a *cropping filter* to keep only the desired portion of the face, a *filling holes*, and a *median filter* to withdraw spikes. In the case of deep covariance, each 3D face model is firstly transformed to 2D map images namely: 2D depth and principal curvatures. *The 2D depth map* shows the gray value of each image pixel which represents the depth of the associated point on the 3D facial scan. *The Principal curvatures map* on the other hand represents the principal curvature values over the 3D mesh. It is approximated by the local cubic fitting method [33]. All 2D faces are then normalized to 224×224 pixels.

4.2 Feature extraction

Once the 3D face scans have been pre-processed, we extract covariance descriptors from deep and shallow features as follows:

4.2.1 Deep covariance features :

We extract deep features using VGG-16 and AlexNet models (see Fig. 2 and Fig. 3).

VGG-16 face model [32] is a deep CNN model pre-trained on the ImageNet database [34]. It contains 16 layers, trained on the ImageNet dataset which has over 14 million images and 1000 classes. This model has been successfully used for face recognition [35] and facial expression recognition [36].

The second deep CNN model used to extract deep features is AlexNet [14] pre-trained on the ImageNet database. It consists of 25 layers including convolution, fully connected, pooling, Rectified Linear Units (ReLU), normalization. AlexNet significantly outperformed the runner-up with a top-5 error rate of 15.3% in the 2012 ImageNet challenge [37].

We only consider the feature maps (FMs) at the last convolutional layer of VGG-16 and AlexNet models, also called channels. Instead of using the softmax function to classify faces, we propose to extract covariance matrices as facial descriptors using the obtained FMs to get a higher discriminative power compared to the extracted deep features. Accordingly, we obtain a more efficient and compact representation that encodes the correlation between the extracted non-linear features within different spatial levels.

Let $\mathcal{P} = \{M_i, i = 1 \dots m\}$ be the set of feature maps extracted from the 2D map images (i.e. 2D depth and principal curvatures) using VGG-face and AlexNet models separately. Each patch M_i encodes the local geometric and spatial properties of the face image.

The extracted features $\phi(f)$ are arranged in a $(c \times w \times h)$ tensor, where w and h denote the width and height of the FMs, respectively, and c is the number of *FMs*. Each feature map M_i is vectorized into a n -dimensional vector with $n = w \times h$, and the input tensor is transformed to a set of n observations stored in the matrix $[v_1; v_2; \dots; v_n] \in R^{c \times n}$. Each observation $v_i \in R^c$ encodes the values of the pixel

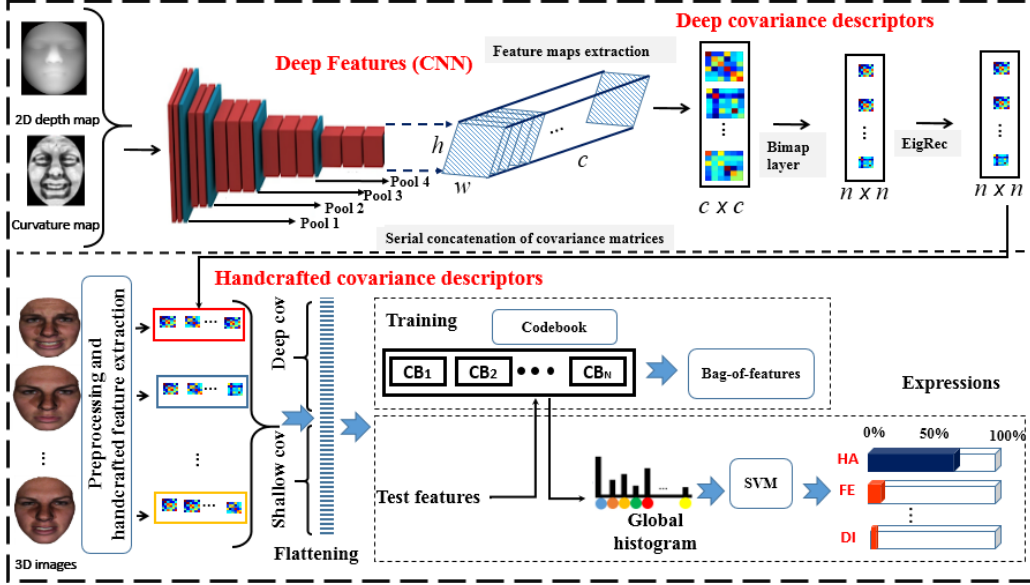


Fig. 1: The Proposal Overview.

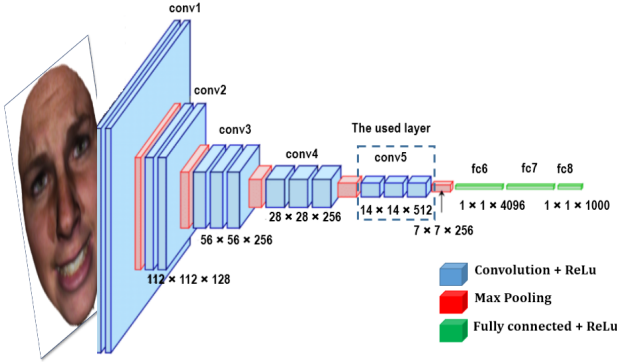


Fig. 2: VGG-16 network architecture proposed in [32].

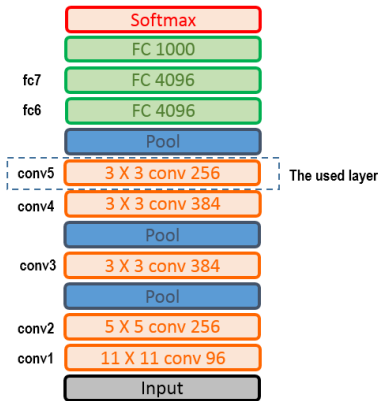


Fig. 3: AlexNet network architecture proposed in [14].

i across all the m feature maps. Finally, we compute the corresponding $(c \times c)$ covariance matrix X_i , defined by Eq.1.

$$X_i = \frac{1}{n} \sum_{j=1}^n (f_j - \mu)(f_j - \mu)^T \quad (1)$$

4.2.2 Shallow covariance features:

We extract 40 patches of the same size from the face surface. Each patch has a reference point that is positioned in its center. We refer to these reference-points by m and to each patch by ρ_i . The covariance matrices are then computed from each patch using their geometric features as proposed in [38].

Let $\mathcal{P} = \{\rho_i, i = 1 \dots m\}$ be the set of patches extracted from a 3D face. Each patch \mathcal{P}_i defines a region around a feature point $p_i = (x_i, y_i, z_i)^t$. For each point f_j in ρ_i , we extract a feature vector F_j , of dimension d , which encodes the local geometrical information and spatial characteristics of the point. In our experiments, we use the following feature vector defined as Eq.2:

$$F_j = [x_j, y_j, z_j, C, M, D_j] \quad (2)$$

where x_j , y_j and z_j are the three-dimensional coordinates of the point p_j . M and C are Mean curvature and Curvedness respectively. D_j is the distance of p_j from the origin. Each patch is defined by a covariance matrix, which is defined by Eq.3:

$$P_i = \frac{1}{n} \sum_{j=1}^n (F_j - \mu)(F_j - \mu)^T \quad (3)$$

where μ is the mean of the feature vectors $\{F_j\}_{j=1 \dots n}$ computed in the patch ρ_i , and n is the number of points in ρ_i .

4.3 Deep learning on SPD matrices

This section introduces the structure of SPD matrices and explain how to learn them on the Sym_d^+ space.

Geometry of SPD matrices: the space of covariance matrices is presented as follows:

The $m \times m$ SPD matrix X is has the particularity of $y^T X y > 0$ while $y \in \mathbf{R}^m$. The space of $m \times m$ SPD matrices,

indicated by Sym_d^+ is not an Euclidean space but a non-linear Riemannian manifold of size $m \times (m + 1)/2$ that supports a Riemannian metric (i.e. geodesic distance) called the Affine Invariant. It is given by:

$\delta_R(\mathbf{A}, \mathbf{B}) = \left\| \log \left(\mathbf{A}^{-1/2} \mathbf{B} \mathbf{A}^{-1/2} \right) \right\|_F$, where \mathbf{A} and \mathbf{B} are covariance matrices.

SPD matrix learning: consists of reducing the SPD matrices size and enhance their discrimination power while preserving their manifold structure. Thus, the output of the SPD matrices learning must still be SPD matrices too. For this reason, we can distinguish linear or non-linear transformations of SPD matrices. Among linear transformation, the Bimap layer has been used in [39]. Non-linear transformation is similar to ReLU layers in CNN, it can be found in [40]. Finally, to flatten the obtained covariance matrices, we apply eigenvalue decomposition (EIG) algorithm to feed a SVM classifier. In the following, we present these steps.

4.3.1 Linear transformation

We apply a bilinear mapping on the covariance matrices using the BiMap layer in order to reduce their size and to be able to concatenate them serially with handcrafted based covariance matrices. The BiMap is considered as a fully connected layer since it preserves the geometrical structure of covariance matrices while reducing dimension as follows: if \mathbf{X}_{k-1} be input SPD matrix, $\mathbf{W}_k \in R^{d_k \times d_{k-1}}$ be weight matrix in the space of full rank matrices and $\mathbf{X}_k \in R^{d_k \times d_k}$ be output matrix, then k^{th} the bilinear mapping f_k^b is defined by Eq.4:

$$\mathbf{X}_k = f_k^b(\mathbf{X}_{k-1}; \mathbf{W}_k) = \mathbf{W}_k \mathbf{X}_{k-1} \mathbf{W}_k^T. \quad (4)$$

4.3.2 Non-linear transformation

Tuning up the covariance matrices is necessary to exploit the ReLU-like layers to introduce a non-linearity to the context of the deep SPD matrices. To do so, we apply the Eigenvalue Rectification presented in [28].

We note \mathbf{X}_{k-1} the input covariance matrix and \mathbf{X}_k the output one. ϵ be the eigenvalue rectification threshold, accordingly, the k^{th} layer f_r^k is defined by Eq.5:

$$\mathbf{X}_k = f_r^k(\mathbf{X}_{k-1}) = \mathbf{U}_{k-1} \max(\epsilon \mathbf{I}, \Sigma_{k-1}) \mathbf{U}_{k-1}^T \quad (5)$$

where \mathbf{X}_{k-1} and Σ_{k-1} are defined by eigenvalue decomposition $\mathbf{X}_{k-1} = \mathbf{U}_{k-1} \Sigma_{k-1} \mathbf{U}_{k-1}^T$.

4.3.3 SPD matrices flattening

Log Eigenvalue Layer is applied to flatten the SPD matrices while preserving the Manifold structure. It consists of applying eigenvalue decomposition and log as matrix operation.

We note \mathbf{X}_{k-1} the input covariance matrix and \mathbf{X}_k the output one. The LogEig layer applied in the k^{th} layer f_l^k is defined by Eq.6:

$$\mathbf{X}_k = f_l^k(\mathbf{X}_{k-1}) = \log(\mathbf{X}_{k-1}) = \mathbf{U}_{k-1} \log(\Sigma_{k-1}) \mathbf{U}_{k-1}^T \quad (6)$$

Where $\mathbf{X}_k = \mathbf{U}_{k-1} \Sigma_{k-1} \mathbf{U}_{k-1}^T$ is an eigenvalue decomposition and log is an element-wise matrix operation [41].

4.4 Shallow and deep Bag-of-Features paradigm and classification

The two sets of the flattened covariance matrices are of different sizes since they are computed from two different types of feature vectors. To overcome this problem, we apply the BoF paradigm proposed in [42] on the two sets of the extracted covariance descriptors (deep and shallow ones after flattening), separately. If we have N covariance matrices of size $d \times d$, the obtained feature vector after flattening is of size $N \times \frac{(d \times d - 1)}{2}$.

Note that the quantization of the obtained feature vectors for each image leads to two predefined number of histogram bins/codewords (deep and shallow histograms). This makes each histogram independent of the number of the obtained feature vectors and the image size. In the following, we refer to the codebooks obtained from the co-varied deep features by *deep codebook*, and to that obtained from the co-varied shallow features by *shallow codebook*. Once the two histograms are computed, we pass to the classification stage to assign each test image to its expression class. To do so, we apply SVM classifier where each face is represented by the global histogram which is the concatenation of the deep and the shallow codebooks.

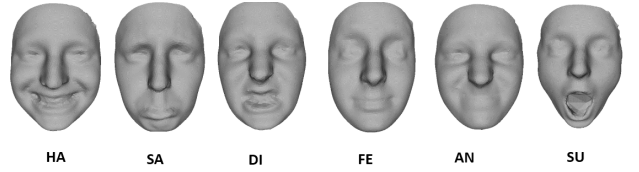


Fig. 4: 3D face models from Bosphorus dataset.

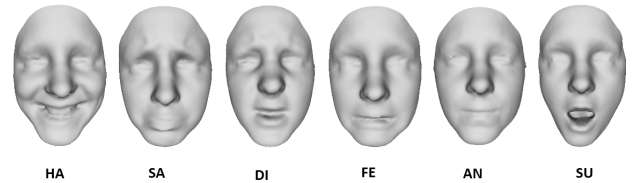


Fig. 5: 3D face models from BU-3DFE dataset.

5 EXPERIMENTS

This section presents the datasets used to evaluate the proposed method and the experimental results compared to other methods in the literature using the same datasets.

5.1 Dataset description

The Bosphorus dataset [12] which was made for testing the algorithms using 3D and 2D facial images for facial analysis and recognition tasks. This dataset includes 105 subjects with many variations (total of 4666 images). It was obtained using structured-light technology to get 3D scans with the six expressions and neutral scan for each subject. Dimensions are defined by 0.3 mm, 0.3 mm, and 0.4 mm respectively. Fig. 4 presents some examples from the Bosphorus dataset for the same subject.

BU-3DFE dataset (Binghamton University 3D Facial Expression) [11] is a multi-view facial expression database

of 2500 images captured in lab-controlled environment. It contains 3D expressive face scans with texture scans of 100 subjects. The basic facial expressions are induced by various ways and head poses with four intensity degrees. we can also find a neutral face associated to each subject. Fig. 5 presents an example of six expressive faces from BU-3DFE dataset. To make a fair comparison with the state-of-the-art methods, we apply the same protocol by excluding the neutral face and only use the six expressive faces.

5.2 Method evaluation

In this section, we present the different evaluation protocols and the experimental results to demonstrate the efficiency of the proposed method. After the pre-processing step, we extract deep and shallow based covariance descriptors as follows:

Deep feature-based covariance descriptors: from each 2D map image (2D depth and principal curvatures), we employed the last convolutional layer of CNN models to extract deeper features from each face image. Using VGG-16, this layer contains 512 feature maps having the size of 14×14 . Thus, we obtain 512×512 covariance descriptors as presented in Fig. 2. When dealing with AlexNet model, we obtain covariance descriptors of size 256×256 . Next, we apply dimension reduction of the obtained covariance matrices with BiMap layer. The dimensionalities of the VGG-16 SPD transformation matrices are set to 512×250 , 250×100 , 100×50 . AlexNet SPD matrices are transformed to 256×150 , 150×100 , 100×50 respectively (see Section 4.3.1). Finally, the obtained SPD matrices are tuned up to introduce a non-linearity to the context of the deep SPD matrices as presented in Section 4.3.2.

Shallow feature-based covariance descriptors: we extract 40 patches from the face surface. Each patch has a reference point that is positioned in its center. We refer to these reference-points by m and to each patch by ρ_i . The radius of each patch is simply given by $r = 15 \times r/100$, where r is the radius of the whole facial shape defined as a bounding sphere.

To describe each patch region, we compute the covariance matrices of size 6×6 using the corresponding feature vector: $[x, y, z, C, M, D]$ as presented in Section 4.2.2.

Features quantization : we finally apply the BoF paradigm on the two sets of covariance matrices described above after flattening as described in Section 4.4. We thus obtain two global histograms for quantization. SVM classifier is then applied on the serial concatenation of the two histograms to classify the six facial expressions.

5.2.1 Performances on BU-3DFE dataset

To assess the proposed system, a 10 fold-cross validation protocol is employed. 90 subjects of the BU-3DFE dataset are then used for training, where 10 subjects are used for the test. Results are averaged across the ten-folds and displayed in the following sections. Fig. 6 (A) shows the classification rate of each expression. From Fig. 6 (A), it is clear that the proposed method recognizes better the expressions of happiness and surprise. Anger and Sad expressions have the lowest performance by 95.20% and 91.95% respectively. These performances are explained by the fact that distinguishing between these two expressions is a difficult task,

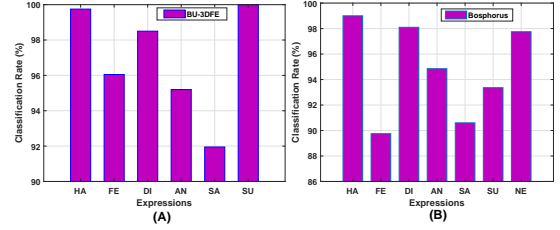


Fig. 6: Facial expression classification performance on (a) BU-3DFE and (b) Bosphorus datasets.

which explains their confusion as presented in Fig. 7. It should be noted that these results are obtained using the best system setting according to the codebook and deep covariance matrices size. More details about the system setting and the effect of the codebook sizes can be found in the following sections.



Fig. 7: Confusion matrix of BU-3DFE dataset.

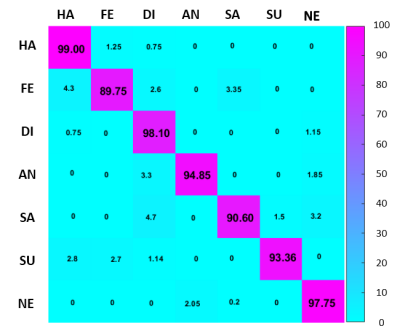


Fig. 8: Confusion matrix of Bosphorus dataset.

5.2.2 Performances on Bosphorus dataset

To evaluate the proposed system on the Bosphorus dataset, we follow the standard protocol (10 fold-cross validation) as in prior methods [43]. Classification performances are shown in Fig. 6 (B). With this dataset, it is clear that happiness and disgust are the best recognized expressions with classification rates of 99.00% and 98.10% respectively. Fear and sad expressions are more challenging and give 89.75% and 90.60% respectively. thus, distinguishing the subjects of Bosphorus dataset with fear and surprise expressions is a very hard task, which explains their confusion as presented in Fig. 8. It should also be noted that these results are obtained using the best system setting.

TABLE 1: Comparison between the performance of the proposed method and state-of-the-art methods on the BU-3DFE dataset.

Method	Images	Landmark	Classifier	Rate
Zhen et al. 2016 [44]	3D mesh	automatic	SVM+HMM	83.20%
Jan et al. 2018 [22]	3D depth+texture	automatic	SVM	88.54%
Li et al. 2017 [23]	2D+3D	automatic	SVM	86.86%
Vo et al. 2019 [16]	2D+3D	automatic	softmax	84.30%
Huynh et al 2016 [45]	2D+3D	-	CNN	92.73%
Derkach et al. 2017 [46]	3D	86 manual	SVM	81.03%
Jiao et al. 2020 [47]	2D+3D	automatic	-	89.72%
Wei et al. 2018 [48]	2D+3D	30 automatic	SVM	88.03%
Ly et al. 2019 [49]	2D+3D	automatic	SVM	87.66%
Shao et al. 2019 [50]	2D+3D	automatic	Softmax	86.50%
Fu et al. 2019 [51]	2D+3D	automatic	SVM	82.89%
<i>3D mesh / 2D depth map</i>				
Our (Shallow COV)	3D mesh	40 automatic	SVM	90.50%
Our (Deep VGG-16 COV)	2D depth	automatic	SVM	94.50%
Our (Deep AlexNet COV)	2D depth	automatic	SVM	94.20%
Our (Deep VGG-16 COV + Deep AlexNet COV)	2D depth	automatic	SVM	94.99%
Our(Deep VGG-16 + Shallow COV)	3D mesh+2D depth	automatic	SVM	96.73%
Our(Deep AlexNet + Shallow COV)	3D mesh+2D depth	automatic	SVM	96.15%
Our(Deep AlexNet VGG-16 + Shallow COV)	3D mesh+2D depth	automatic	SVM	96.90%
<i>3D mesh / Principal curvatures map</i>				
Our (Deep VGG-16 COV)	Principal curvature	automatic	SVM	92.14%
Our (Deep AlexNet COV)	Principal curvature	automatic	SVM	92.00%
Our (Deep VGG-16 COV + Deep AlexNet COV)	Principal curvature	automatic	SVM	93.00%
Our(Deep VGG-16 + Shallow COV)	3D mesh+Principal curvature	automatic	SVM	94.71%
Our(Deep AlexNet + Shallow COV)	3D mesh+Principal curvature	automatic	SVM	94.32%
Our(Deep VGG-16 + AlexNet + Shallow COV)	3D mesh+Principal curvature	automatic	SVM	95.15%

TABLE 2: Comparison between the performance of the proposed method and state-of-the-art methods on the Bosphorus dataset.

Method	Images	Landmark	Classifier	Rate
Ramya et al.2020 [52]	3D	automatic	SVM	87.69%
Wang et al. 2013 [53]	3D	automatic	SVM	76.56%
Jiao et al. 2020 [47]	2D+3D	automatic	-	83.63%
Vo et al. 2019 [16]	2D+3D	automatic	softmax	82.40%
Fu et al. 2019 [51]	2D+3D	automatic	SVM	75.93%
Wei et al. 2018 [48]	2D+3D	30 automatic	SVM	82.50%
<i>3D mesh / 2D depth map</i>				
Our (Shallow COV)	3D mesh	40 automatic	SVM	86.17%
Our (Deep VGG-16 COV)	2D depth	automatic	SVM	92.20%
Our (Deep AlexNet COV)	2D depth	automatic	SVM	93.30%
Our (Deep VGG-16 COV + Deep AlexNet COV)	2D depth	automatic	SVM	93.95%
Our(Deep VGG-16 + Shallow COV)	3D mesh+2D depth	automatic	SVM	94.26%
Our(Deep AlexNet + Shallow COV)	3D mesh+2D depth	automatic	SVM	94.55%
Our(Deep VGG-16 + AlexNet + Shallow COV)	3D mesh+2D depth	automatic	SVM	94.77%
<i>3D mesh / Principal curvatures map</i>				
Our (Deep VGG-16 COV)	Principal curvature	automatic	SVM	91.16%
Our (Deep AlexNet COV)	Principal curvature	automatic	SVM	91.50%
Our (Deep VGG-16 COV + Deep AlexNet COV)	Principal curvature	automatic	SVM	92.50%
Our(Deep VGG-16 + Shallow COV)	3D mesh+Principal curvature	automatic	SVM	93.88%
Our(Deep AlexNet + Shallow COV)	3D mesh+Principal curvature	automatic	SVM	93.95%
Our(Deep VGG-16 + AlexNet + Shallow COV)	3D mesh+Principal curvature	automatic	SVM	94.25%

5.3 Results comparison and analysis

5.3.1 BU-3DFE

Table 1 presents the performance comparison of the proposed method with those of the literature on the BU-3DFE dataset. The reported results are obtained using the same setting of evaluation. When dealing with the whole dataset, Jan et al. 2018 [22] achieved 88.54% using a deep CNN model with hand-crafted features. Derkach et al. 2017 [46] used graph laplacian features and obtained an accuracy of 81.03%. In [48], the authors applied Wasserstein distance with transformed training strategy and obtained 88.03%. Ly et al. 2019 [49] applied deep 2D and 3D multi-modal approach and SVM classifier. They achieved 87.66%. Finally, Shao et al. 2019 [50] have conducted three pre-trained CNNs (i.e. shallow network, dual-branch CNN and CNN with transfer learning technique) only on 5 expressions and obtained 86.50%.

The proposed method overcomes the previous methods and achieved 96.90% using the quantization of the combination VGG-16, AlexNet, and Shallow covariance features.

Table 3 shows a comparison between the obtained classification rates and state-of-the-art ones that aim to recognize the six prototypical expressions. It is clear that the proposed method achieved the best classification rate with FE, DI, AN, SA and SU expressions using the combination of shallow and deep based features by (96.05%, 98.50%, 95.20%, 91.95% and 100%) respectively. With HA expression, Huynh et al. 2016 [45] obtained 100% as highest rate.

5.3.2 Bosphorus

Table 2 presents the performance comparison between the proposed method with those of the literature on the Bosphorus dataset. It is clear that the proposed method gives the best accuracy (94.77%) using the quantization of the combination VGG-16, AlexNet, and Shallow covariance features. The use of deep covariance features (VGG-16 and Alexnet) separately with shallow ones achieves a slightly lower recognition performance by 94.26% and 94.55% respectively. This combination has improved the previous one obtained by each deep model separately by 92.20% and 92.30%.

The proposed method outperforms prior methods. For example, Ramya et al. 2020 [52] applied a transfer learning based-technique to fine-tune the pre-trained model AlexNet after computing local binary pattern (LBP) and local bidirectional pattern features. They achieved 87.69%. Also, Vo et al. 2019 [16] has achieved 82.40 using multi-view CNN. In [56] the authors used 2D transformed images and the texture information, they obtained 76.98%. Finally, Jiao et al. 2020 [47] applied VGG-16 and a trained network from scratch and obtained 82.50%.

Table 4 shows a comparison between the obtained classification rates and state-of-the-art ones in order to recognize the six prototypical expressions and also the neutral face. The reported results show that Ramya et al. 2020 [52] outperformed the other methods with FE expression by 96.92%. When dealing with SU expression, Wang et al. 2013 [53] has achieved the highest classification rate by 95.60%. Our proposed method on the other hand outperformed the state-of-the-art methods when dealing with the expressions: HA, DI, AN, SA and NE by 99.00%, 98.10%, 94.85%, 90.60% and 95.75% respectively.

5.4 Effect of the codebook size

To further evaluate the performance of the proposed method, we study in this section the effect of the codebook size on the classification rate. We evaluated 7 sizes (i.e. 16, 32, 64, 128, 256, 512 and 1024). By intuition, if the codebook size is too small, the histogram feature loses discriminant power to classify the expressions, whereas the performance increases when the codebook size grows.

Fig. 9a presents the improvement of the classification rate according to the codebook size using the shallow and VGG-16 co-varied features on BU-3DFE dataset. The best classification rate is achieved by the combination of (Deep depth + shallow) codebooks, followed by (Deep depth + shallow), (Deep depth), and (Deep curvature) codebooks respectively. As expected, the shallow codebook gives the lowest classification rate since quantized covariance descriptors don't capture deeper features compared to VGG-16 ones. Nevertheless, integrating a shallow codebook with a deep codebook improves the performance of the proposed method. Overall, the classification rate grows when the codebook size gets bigger and almost stabilizes from the size 512.

Fig. 9b presents the variation of the classification rate according to the codebook size using the shallow and AlexNet co-varied features. The same discipline of the previous figure is maintaining here, where the combination (Deep depth + shallow) gives the highest performance.

The same study has been conducted on the Bosphorus dataset. From Fig. 10a and Fig. 10b, we can notice that the shallow codebook gives the lowest classification rate as shown before on BU-3DFE. When dealing with deep-based codebook, curvature-based codebook initially outperforms depth-based codebook with small codebook size (i.e. from 16 to 256). In contrast, using grand codebook size (i.e. 512 and 1024), the depth-based codebook becomes better and slightly outperforms the curvature-based one. This discipline can be explained by the fact that depth information needs a grand quantized feature to be sufficiently encoded.

The performance of the combination between VGG-16 and AlexNet co-varied features are presented in the Fig. 11 according to the codebook size. Using the two datasets, the highest performance is realized by the combination *Depth (VGG-16 + AlexNet) + Shallow* and *Curvature (VGG-16 + AlexNet) + Shallow* codebooks respectively. This achievement can be explicated by the fact that VGG-16-based and AlexNet codebooks are complimentary. Moreover, the shallow-based codebook enhances the performance of the proposed method through the face's encoded geometry and topological information. The lowest performance, in contrast, is obtained using the combination *Curvature (VGG-16 + AlexNet)*. Thus, deeper features extracted from curvature map images may lack some geometrical information that can be provided by the shallow features. This further demonstrates that the combination of the two codebooks (shallow and deep ones) provides high discriminative power and thus is suitable for multi-class SVM classification to overcome the huge challenges in the 3D FER task.

TABLE 3: Comparison of classification rates (%) per expression of our proposed method (shallow features only + best combination) with state-of-the-art methods on the BU-3DFE dataset.

Method	HA	FE	SA	AN	DI	SU	Average
Zhen et al. 2016 [44]	94.6	63.3	79.2	79.5	85.7	96.1	83.2
Berretti et al. 2010 [54]	86.9	63.6	64.6	81.7	73.6	94.8	77.53
Li et al. 2017 [23]	96.26	79.24	81.18	82.08	84.94	97.43	86.86
Huynh et al. 2016 [45]	100	86.7	87.5	91.3	95.2	95.7	92.73
Derkach et al. 2017 [46]	89.50	65.12	77.20	85.58	75.31	93.50	81.03
Lemaire et al. 2013 [55]	89.8	64.6	74.5	74.1	74.9	90.9	78.13
Vo et al. 2019 [16]	87.50	66.88	81.25	80.00	79.06	91.25	84.30
Fu et al. 2019 [51]	92.25	70.75	78.91	80.92	78.67	95.83	82.89
Our method (Shallow only)	95.5	89.67	83.33	86.00	92.37	96.33	90.50
Our method (Best)	99.75	96.05	91.95	95.20	98.50	100	96.90

TABLE 4: Comparison of classification rates (%) per expression of our proposed method (shallow features only + best combination) with state-of-the-art methods on the Bosphorus dataset.

Method	HA	FE	SA	AN	DI	SU	NE	Average
Wang et al. 2013 [53]	92.50	62.80	74.50	63.50	70.60	95.60	-	76.56
Fu et al. 2019 [51]	92.97	63.83	65.97	77.37	67.03	88.40	-	75.95
Ramya et al. 2020 [52]	83.08	96.92	78.46	87.69	87.69	93.85	86.15	87.69
Azazi et al. 2015 [43]	97.50	86.25	67.50	82.50	90.00	83.75	81.25	84.10
Our method (Shallow only)	93.00	81.00	79.75	86.25	85.25	90.50	87.50	86.17
Our method (Best)	99.00	89.75	90.60	94.85	98.10	93.36	97.75	94.77

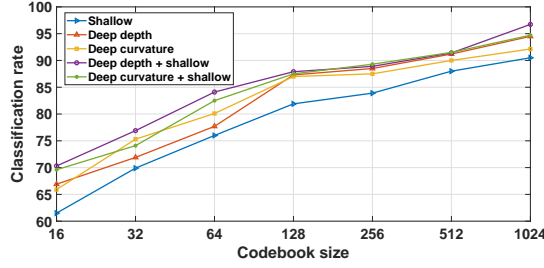
6 DISCUSSION

The reported results above show that our proposed method achieved state-of-the-art performance without using additional training data or facial registration. The reliable accuracy is obtained thanks to the high discrimination power of the SVM classifier. This high-performance power cannot be achieved without using proper discriminative face descriptors. Basically, deep and shallow features-based covariance descriptors capture all the information of the facial expressions and their correlation. Moreover, we show that deep covariance-based descriptors give higher classification rates comparing to shallow ones. This is a predictable discipline due to the high efficiency of deep features reinforced by the covariance matrices learning through the BiMap layer and the non-linear transformation. Since covariance matrices inherit their performance from the used features, CNN models are designed to extract the best features from the 2D transformed face images (i.e. depth and curvature maps). Shallow features on the other hand extract more details about geometrical and spatial information. More particularly, the reported results show that the pre-trained deep features on depth map images outperformed the curvature-based one when dealing with BU-3DFE and Bosphorus datasets. This difference is obtained due to the higher description of the face structure insured by the 2D depth map images.

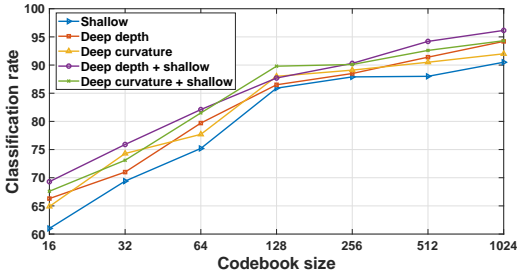
We can also notice that incorporating VGG-16 and AlexNet models boosts the performance of the proposed framework compared to the use of each model separately. This is because the two models have different architectures. Thereby, the convolution layer used to generate our deep features are relatively different; each of them encodes different feature abstraction levels that could capture complementary characteristics from the facial expression images.

Embedding additional pre-trained models are supposed to improve the performance of the proposed framework, however, the choice of the model should be thoroughly studied according to the number of layers and the degree of similarity between the trained database used to initialize these models to be used for the FER task.

The obtained experimental results further demonstrate the capacity to notably improve the FER performance using pre-trained deep network structures combined with a shallow structure. This combination through the two codebooks could overcome the shortage of training data and overfitting problem. We can conclude that these two covariance-based methods are complementary and their quantization using the BoF paradigm is suitable for the 3D FER task. It's worth noting that this complementarity is achieved due to the 2D+3D multi-modality, where each modality can capture different characteristics of the facial expression. This is an advantage compared to state-of-the-art methods that generally extracted shallow and deep features from the same 2D image modality, for example, Yang et al. [57] extracted deep features using pre-trained models, and LBP features from the same 2D images. When dealing with the same 3D datasets, the fine-tuning applied by Ramya et al. 2019 of the AlexNet model with additional shallow CNN (called multi-channel framework) doesn't outperform our proposed method using the AlexNet model separately (87.69% and 93.30%; respectively). The reported results are the overall accuracies obtained using Bosphorus dataset. Thereby, Our proposed method outperforms Ramya et al.'s method with six expressions from seven. This finding further demonstrates the efficiency of the proposed system through the quantization of the co-varied generated deep features.



(a) VGG-16

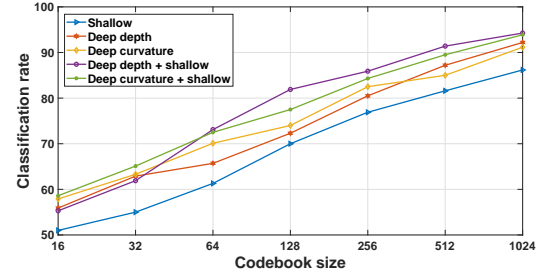


(b) AlexNet

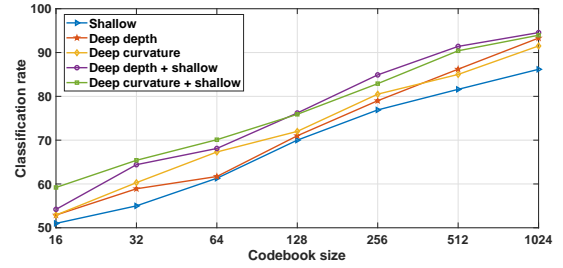
Fig. 9: Effect of the codebook size on the classification performance. The reported results are obtained on the BU-3DFE dataset using the deep codebooks of VGG-16 and AlexNet separately, each deep codebook is combined with the shallow codebook. The displayed sizes are the same for each codebook.

7 CONCLUSION

A new methodology for 3D FER is presented. To deal with the significant variations from facial expression images, we have employed covariance matrices that ensure an efficient combination of different features extracted from distinct image modalities. To do so, we have adopted two different types of features to extract the covariance matrices (i.e. deep and shallow features). The purpose is to capture not only the best features extracted from the last convolutional layer of VGG-16 and AlexNet models but also to capture more geometrical and spatial details from the 3D expressive faces. Note that our method is generic so that other pre-trained deep learning models can be added (e.g. ResNet50), it can also be computed from other 2D transformed images such as Shape Index or Curvedness maps. Covariance matrices, however, belong to the Riemannian structure of Sym_d^+ space. Therefore, classifying facial expressions using their covariance-based descriptors needs a global and optimal description to be able to employ traditional classification algorithms. Therefore, we first apply linear and non-linear transformations using deep covariance matrices learning in order to reduce their size and enhance their discrimination power while preserving their manifold structure. Second, we flatten the obtained covariance matrices (deep and shallow ones). The BoF paradigm is then applied to the two sets of the obtained features of the flattened covariance matrices. A multi-class SVM is finally employed to classify the expressions after being trained using the global quantized feature vector (i.e. deep and shallow codebooks).



(a) VGG-16



(b) AlexNet

Fig. 10: Effect of the codebook size on the FER performance. The displayed results are obtained on the Bosphorus dataset using the deep codebooks of VGG-16 and AlexNet separately, each deep codebook is combined with the shallow codebook. The displayed sizes are the same for each codebook.

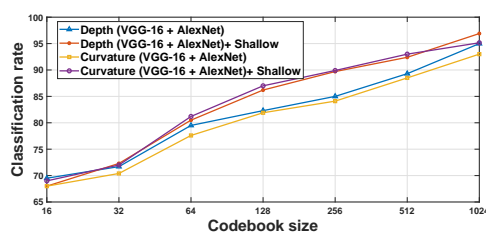
The displayed performances obtained on the BU-3DFE and Bosphorus datasets demonstrate that the deep codebook gives higher classification rates comparing to the shallow codebook. Furthermore, their combination outperformed the deep codebook single-handed. This discipline proves that the two codebook-based methods are complementary and efficiently classify 3D facial expressions. Moreover, the reported results also confirm the advantage of the application of the BoF paradigm to classify facial expressions using 3D data in comparison to state-of-the-art methods. In future work, dynamic features could be added to the proposed system to boost the classification rate of the 3D FER (BU-4DFE dataset). Also, we will investigate the use of point cloud as input in CNN by applying the PointNet architecture proposed in [58]. We also look at the application of generative models to provide additional training data to deal with the problem of small 3D FER datasets and to further enhance the efficiency of the proposed framework.

ACKNOWLEDGMENTS

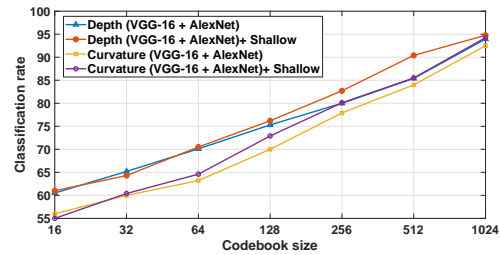
The authors would like to thank the DGRSDT, Algeria.

REFERENCES

- [1] A. P. Ismail, "The usage of combined components of verbal l, vocal and visual (3-v components) of children in daily conversation: Psycholinguistic observation," *ELS Journal on Interdisciplinary Studies in Humanities*, vol. 2, no. 2, pp. 290–301, 2019.



(a) BU-3DFE



(b) Bosphorus

Fig. 11: Effect of the codebook size on the FER performance on the Bosphorus and BU-3DFE datasets. The deep codebooks of VGG-16 and AlexNet are combined with each other, and with the shallow codebook as well.

- [2] T. Fang, X. Zhao, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris, "3d facial expression recognition: A perspective on promises and challenges," in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*. IEEE, 2011, pp. 603–610.
- [3] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.
- [4] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, 2020.
- [5] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.
- [6] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.
- [7] H. Mahersia and K. Hamrouni, "Using multiple steerable filters and bayesian regularization for facial expression recognition," *Engineering Applications of Artificial Intelligence*, vol. 38, pp. 190–202, 2015.
- [8] K. Dutta, D. Bhattacharjee, M. Nasipuri, and O. Krejcar, "Complement component face space for 3d face recognition from range images," *Applied Intelligence*, vol. 51, no. 4, pp. 2500–2517, 2021.
- [9] H. Patil, A. Kothari, and K. Bhurchandi, "3-d face recognition: features, databases, algorithms and challenges," *Artificial Intelligence Review*, vol. 44, no. 3, pp. 393–441, 2015.
- [10] O. Ocegueda, T. Fang, S. K. Shah, and I. A. Kakadiaris, "3d face discriminant analysis using gauss-markov posterior marginals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 728–739, 2013.
- [11] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*. IEEE, 2006, pp. 211–216.
- [12] A. Savran, N. Alyüz, H. Dibeklioglu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3d face analysis," in *European Workshop on Biometrics and Identity Management*. Springer, 2008, pp. 47–56.
- [13] G. R. Alexandre, J. M. Soares, and G. A. P. Thé, "Systematic review of 3d facial expression recognition methods," *Pattern Recognition*, vol. 100, p. 107108, 2020.
- [14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [15] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610–628, 2017.
- [16] Q. N. Vo, K. Tran, and G. Zhao, "3d facial expression recognition based on multi-view and prior knowledge fusion," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSp)*. IEEE, 2019, pp. 1–6.
- [17] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2015, pp. 1912–1920.
- [18] Y. Li, S. Pirk, H. Su, C. R. Qi, and L. J. Guibas, "Fpnn: Field probing neural networks for 3d data," in *Advances in Neural Information Processing Systems*, 2016, pp. 307–315.
- [19] K. Papadopoulos, A. Kacem, A. Shabayek, and D. Aouada, "Face-gcn: A graph convolutional network for 3d dynamic face identification/recognition," *arXiv preprint arXiv:2104.09145*, 2021.
- [20] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 945–953.
- [21] D. Z. Wang and I. Posner, "Voting for voting in online point cloud object detection," in *Robotics: Science and Systems*, vol. 1, no. 3, 2015, pp. 10–15 607.
- [22] A. Jan, H. Ding, H. Meng, L. Chen, and H. Li, "Accurate facial parts localization and deep learning for 3d facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 466–472.
- [23] H. Li, J. Sun, Z. Xu, and L. Chen, "Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2816–2831, 2017.
- [24] O. K. Oyedotun, G. Demisse, A. El Rahman Shabayek, D. Aouada, and B. Ottersten, "Facial expression recognition via joint deep learning of rgb-depth map latent representations," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 3161–3168.
- [25] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 443–449.
- [26] Y. Fan, J. C. Lam, and V. O. Li, "Multi-region ensemble convolutional neural network for facial expression recognition," in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 84–94.
- [27] Z. Huang, C. Wan, T. Probst, and L. Van Gool, "Deep learning on lie groups for skeleton-based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6099–6108.
- [28] Z. Huang and L. Van Gool, "A riemannian network for spd matrix learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [29] Z. Huang, J. Wu, and L. Van Gool, "Building deep networks on grassmann manifolds," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [30] M. T. Harandi, M. Salzmann, S. Jayasumana, R. Hartley, and H. Li, "Expanding the family of grassmannian kernels: An embedding perspective," in *European Conference on Computer Vision*. Springer, 2014, pp. 408–423.
- [31] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi, "Kernel methods on the riemannian manifold of symmetric positive definite matrices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 73–80.
- [32] K. Simonyan and A. Zisserman, "Very deep convolutional

- networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [33] J. Goldfeather and V. Interrante, "A novel cubic-order algorithm for approximating principal direction vectors," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 1, pp. 45–63, 2004.
- [34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [35] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition," in *BMVC*, vol. 1, no. 3, 2015, p. 6.
- [36] Y. Ma, X. Wang, and L. Wei, "Multi-level spatial and semantic enhancement network for expression recognition," *Applied Intelligence*, pp. 1–14, 2021.
- [37] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.
- [38] W. Hariri, H. Tabia, N. Farah, A. Benouareth, and D. Declercq, "3d facial expression recognition using kernel methods on riemannian manifold," *Engineering Applications of Artificial Intelligence*, vol. 64, pp. 25–32, 2017.
- [39] T. Zhang, W. Zheng, Z. Cui, and C. Li, "Deep manifold-to-manifold transforming network," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4098–4102.
- [40] M. Engin, L. Wang, L. Zhou, and X. Liu, "Deepkspd: Learning kernel-matrix-based spd representation for fine-grained image recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 612–627.
- [41] D. Acharya, Z. Huang, D. Pani Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 367–374.
- [42] S. O'Hara and B. A. Draper, "Introduction to the bag of features paradigm for image classification and retrieval," *arXiv preprint arXiv:1101.3354*, 2011.
- [43] A. Azazi, S. L. Lutfi, I. Venkat, and F. Fernández-Martínez, "Towards a robust affect recognition: Automatic facial expression recognition in 3d faces," *Expert Systems with Applications*, vol. 42, no. 6, pp. 3056–3066, 2015.
- [44] Q. Zhen, D. Huang, Y. Wang, and L. Chen, "Muscular movement model-based automatic 3d/4d facial expression recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1438–1450, 2016.
- [45] X.-P. Huynh, T.-D. Tran, and Y.-G. Kim, "Convolutional neural network models for facial expression recognition using bu-3dfe database," in *Information Science and Applications (ICISA) 2016*. Springer, 2016, pp. 441–450.
- [46] D. Derkach and F. M. Sukno, "Local shape spectrum analysis for 3d facial expression recognition," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 41–47.
- [47] Y. Jiao, Y. Niu, T. D. Tran, and G. Shi, "2d+ 3d facial expression recognition via discriminative dynamic range enhancement and multi-scale learning," *arXiv preprint arXiv:2011.08333*, 2020.
- [48] X. Wei, H. Li, J. Sun, and L. Chen, "Unsupervised domain adaptation with regularized optimal transport for multimodal 2d+ 3d facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 31–37.
- [49] T. S. Ly, N.-T. Do, S.-H. Kim, H.-J. Yang, and G.-S. Lee, "A novel 2d and 3d multimodal approach for in-the-wild facial expression recognition," *Image and Vision Computing*, vol. 92, p. 103817, 2019.
- [50] J. Shao and Y. Qian, "Three convolutional neural network models for facial expression recognition in the wild," *Neurocomputing*, vol. 355, pp. 82–92, 2019.
- [51] Y. Fu, Q. Ruan, Z. Luo, Y. Jin, G. An, and J. Wan, "Ferlrtc: 2d+ 3d facial expression recognition via low-rank tensor completion," *Signal Processing*, 2019.
- [52] R. Ramya, K. Mala, and S. S. Nidhyananthan, "3d facial expression recognition using multi-channel deep learning framework," *Circuits, Systems, and Signal Processing*, vol. 39, no. 2, pp. 789–804, 2020.
- [53] Y. Wang, M. Meng, and Q. Zhen, "Learning encoded facial curvature information for 3d facial emotion recognition," in *Image and Graphics (ICIG), 2013 Seventh International Conference on*. IEEE, 2013, pp. 529–532.
- [54] S. Berretti, A. Del Bimbo, P. Pala, B. B. Amor, and M. Daoudi, "A set of selected sift features for 3d facial expression recognition," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 4125–4128.
- [55] P. Lemaire, M. Ardabilian, L. Chen, and M. Daoudi, "Fully automatic 3d facial expression recognition using differential mean curvature maps and histograms of oriented gradients," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013. IEEE, 2013, pp. 1–7.
- [56] S.-Y. Chun, C.-S. Lee, and S.-H. Lee, "Facial expression recognition using extended local binary patterns of 3d curvature," in *Multimedia and Ubiquitous Engineering*. Springer, 2013, pp. 1005–1012.
- [57] B. Yang, J. Cao, R. Ni, and Y. Zhang, "Facial expression recognition using weighted mixture deep neural network based on double-channel facial images," *IEEE Access*, vol. 6, pp. 4630–4640, 2017.
- [58] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 652–660.

Degraded Document Image Binarization using Novel Background Estimation Technique

Harshit Jindal*, Manoj Kumar[†], Akhil Tomar[‡] and Ayush Malik[§]
Department of Computer Science Engineering, Delhi Technological University
New Delhi, India

Email: *harshitjindal2000@gmail.com, [†]mkumarg@dce.ac.in, [‡]akhiltomar098@gmail.com, [§]ayushmalik03@gmail.com

Abstract—Over the past few decades, the use of scanned historical document images has increased dramatically, especially with the emergence of online libraries and standard benchmark datasets like DIBCO. The historical documents are usually in very-poor conditions containing noises like large ink stains, bleed-through, liquid spills, uneven-background, spots, faded-ink, weak/thin text that makes the task of binarization very difficult. In this paper, we propose an effective degraded document image binarization algorithm that performs accurate text segmentation. Our method first estimates the background utilizing information from neighboring pixels and filter smoothening. The next step is background subtraction that helps in the compensation of background distortions. The document is segmented using Otsu thresholding, and then we process the image to remove the remaining noise and maximize text content using labelled connected components. Our method outperforms several existing and widely used binarization algorithms on F-measure, PSNR, DRD, and pseudo F-measure when evaluated on H-DIBCO 2016 and H-DIBCO 2018 datasets and can very effectively detect faint characters from a document image.

Index Terms—Document Image Processing, Degraded Document Image Binarization, Thresholding, Background estimation, Noise Removal, Otsu Thresholding, Bilateral Filtering

I. INTRODUCTION

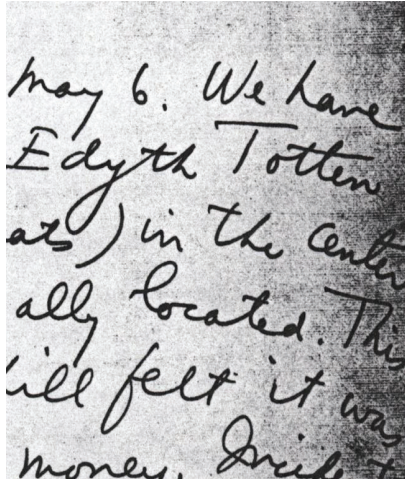
Image binarization is the pre-processing step required for digital image processing [1] and analysis tasks like moving object detection or finding the region of interest in an image like the text in a degraded document [2], [3]. Binary images take less storage memory, document layout analysis, document skew detection, and faster computations for the specific application. The most common way of binarizing an image is to use a thresholding algorithm that segments the image into background and foreground based on a globally or locally computed threshold value for pixel intensities. Ancient Historical Documents, Manuscripts are stored over the years in archives, libraries. Over time, their quality is affected due to several environmental factors like ageing of the paper, humidity, mishandling by humans, and dust. The presence of degradations makes the task of binarization of documents and preserving them digitally a tough job. With increasing interest in historical document analysis, there has been rapid development in document image binarization. Historical documents have various degradations like faded ink or faint-characters, bleed-through, uneven illumination, contrast variation, smear. New techniques keep coming out regularly because there is no single algorithm that can handle all these degradations

and there is always room for improvement in the quality of binarization.

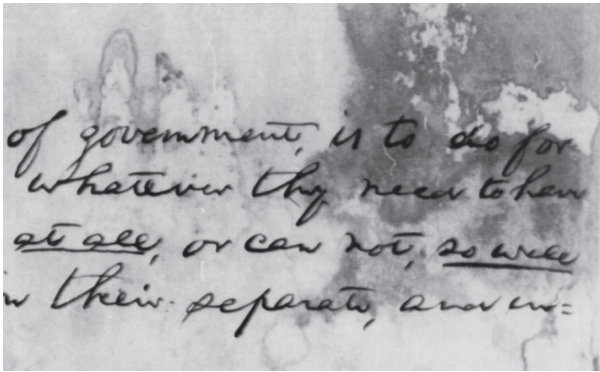
According to Sezgin and Sankur in [4], there are six kinds of thresholding methods, histogram-shape based methods, spatial methods, clustering-based methods, entropy-based methods, object-attribute based methods, local methods. We can also describe the thresholding algorithms as Global or Local depending on the method used to compute the threshold value. In global methods [5], a single threshold value is computed for the image based on entropy, histogram, or a clustering algorithm that segments the image into two classes: foreground(text) and background. Whereas local thresholding algorithms [6] divide the image into subparts and use the information from neighboring pixels in a small fixed-sized window of the image to determine the threshold locally for one subpart at a time. Global thresholding methods efficiently extract text from high-quality documents. But their performance starts to lag when degradations are present and adaptive thresholding algorithms that make a local estimate of the threshold for each pixel produce better results.

Over time several algorithms have been proposed to convert ancient images into their binary form, Global thresholding algorithms like Otsu [7], Kittler [8], etc., and Local thresholding algorithms like Niblack [9], Sauvola [10], etc. Otsu's method performs well with images having a bimodal histogram but can't handle degradations like uneven illumination, bleed-through, or fainting text. Kittler's algorithm assumes a gaussian distribution of pixel values for each pixel level to find the segmentation cut-off and works well for high-quality documents. Niblack's method calculates the mean and standard deviation in a fixed-sized window for each pixel and hence computes a local threshold. It effectively recognizes text but introduces tons of noise. Sauvola tried to modify Niblack's method to reduce noise which gave better results in some cases. But, it also resulted in a lower text detection rate.

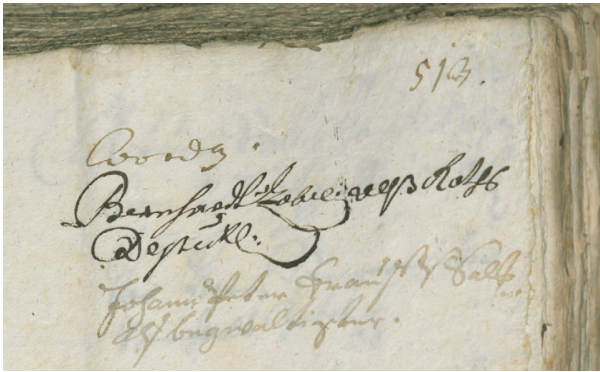
Many methods have incorporated these algorithms and other techniques like contrast normalization, background estimation, stroke-width detection to form hybrid algorithms that perform much better than individual algorithms. Before, moving to our approach we discuss the development of such binarization algorithms. Kim et al. [14] consider the image as a 3d terrain on which water is poured to fill valleys representing text. The water-filled image is subtracted from the original image followed by Otsu's method to form the final binarization



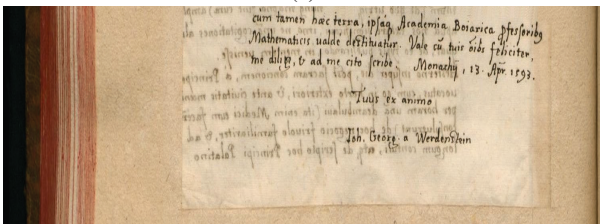
(a)



(b)



(c)



(d)

Fig. 1: Degraded Handwritten Document images from DIBCO dataset [11], [12], [13] illustrating various degradations like contrast variation in (a), smear in (b), faint ink in (c), and bleed-through in (d).

result and we get a nature-inspired unique hybrid binarization algorithm. Adaptive local thresholding algorithm also utilizes the document-specific domain knowledge. Gatos et al. [15] make use of the Weiner filter to pre-process the input image and then the Sauvola's algorithm to estimate the background. The difference between the pre-processed image and the estimated background gives an intermediate result which is finally passed through post-processing to produce the final result. Although this algorithm is better than Sauvola, it is still unable to process bleed-through or recover faint characters from the image. After this came, Lu's [16] Iterative polynomial smoothing-based background-estimation algorithm. It then makes use of a normalized image thresholded by Otsu to detect the text-stroke width. The final result is produced by local thresholding information based on the detected stroke edges and the mean intensity. This method won the first DIBCO competition (DIBCO 2009) [11] but still, it's not perfect. This method is based on the local contrast and hence sometimes high contrast background is difficult to handle. Su et al. [17] used minimum and maximum intensity in the local window to form a contrast image and then binarized it to detect text edges and were able to produce better results than Lu. This method unlike Lu performed well on documents with bleed-through but this also suffers from the faint text.

Nina's binarization algorithm [18] was a six staged background estimation based algorithm that used median filtering for background estimation, bilateral filter, recursive Otsu, contrast compensation, and despeckling. This algorithm can effectively capture all the text edges. Singh et al. [19] proposed an adaptive four-step method that includes contrast stretching, contrast analysis, thresholding, and noise removal from the document image. It works very well with most of the degradations but fails when the image suffers from bleed-through. Howe's algorithm [20] was one of the first automatic parameter tuning based complex binarization algorithm. It used graph-cut computation to minimize the energy function based on the Laplacian of pixel intensities. It is a very efficient method that performs well and autotunes itself on different kinds of images. But this method also like others performs poorly on faint text characters and sometimes even introduces background noise. A recent method that won DIBCO 2018 [13]. (Wei. et al.) [21] makes use of morphological black hat transformation to compensate the document background using a disc-shaped structuring element of a size determined by stroke width transform. Howe's binarization algorithm is then applied to form a binary image which is further enhanced by post-processing that reduces noise and also preserves the text stroke-width.

Our motivation came from the consideration of the amount of ink wasted in Xerox due to poor binarization and the challenging task of detecting faint characters, removal of bleed-through text from handwritten document images. The proposed algorithm combines several steps like background estimation, smoothing filters, thresholding (Otsu), and post-processing steps. It first estimates the background utilizing the information from neighboring pixels to get an estimate of the

current pixel value and moves similarly for the rest of the pixels. The estimated background image is then smoothened using a bilateral filter that reduces random noises from the estimated background. Then we extract the text using a global thresholding method(Otsu) and finally, the image is processed further to get rid of the remaining noise. This paper's major contribution is the algorithm's ability to detect bold and faint text characters from handwritten documents very effectively just by adjusting the parameters of the background estimation while still being able to process bleed-through, smear, and other degradations.

The rest of the paper is arranged as follows:

In Section II, we present our proposed algorithm in detail. Section III demonstrates and describes the experimental results. Finally, we conclude in Section IV.

II. PROPOSED METHOD

Our method utilizes existing techniques and combines them with our technique to form a novel document image binarization method. It includes the use of a novel iterative sliding-window based background estimation method, bilateral filter [22], Otsu thresholding, and effective post-processing to remove the remaining noise from the image. Our proposed binarization algorithm consists of the following steps:

- 1) Conversion of the Image to Grayscale
- 2) Background Estimation and Removal
- 3) Gaussian Smoothing
- 4) Otsu Thresholding
- 5) Spotting Removal

A. Conversion of the Image to Grayscale

Historical documents usually do not have the dynamic color range and hence it is not a good idea to process them in the colored form with millions of color intensities. Different documents have different colors and it is a good idea to convert all to the same color to ensure that all documents are processed in the same manner. Hence, we convert the image to a grayscale form where there are only 2^8 unique intensities.

B. Background Estimation and Removal

After we have converted the image to grayscale form, we first try to remove as much background noise as possible by forming an accurate estimation of the background. Our background estimation method is based on the assumption that the text color varies significantly from the background color. Usually, either the text is dark and the background is light or vice-versa. The neighboring pixels thus can provide a good approximation of the local background of a pixel. We utilize this information i.e. pixels with minimum and maximum intensity in the local neighbourhood to form an iterative sliding window background estimation algorithm based on the following equation:

$$I_{max}(x, y) = \max(I(x_c, y_c)) \quad (1)$$

$$I_{min}(x, y) = \min(I(x_c, y_c)) \quad (2)$$

$$x_c \in (x - w : x + w), y_c \in (y - w : y + w)$$

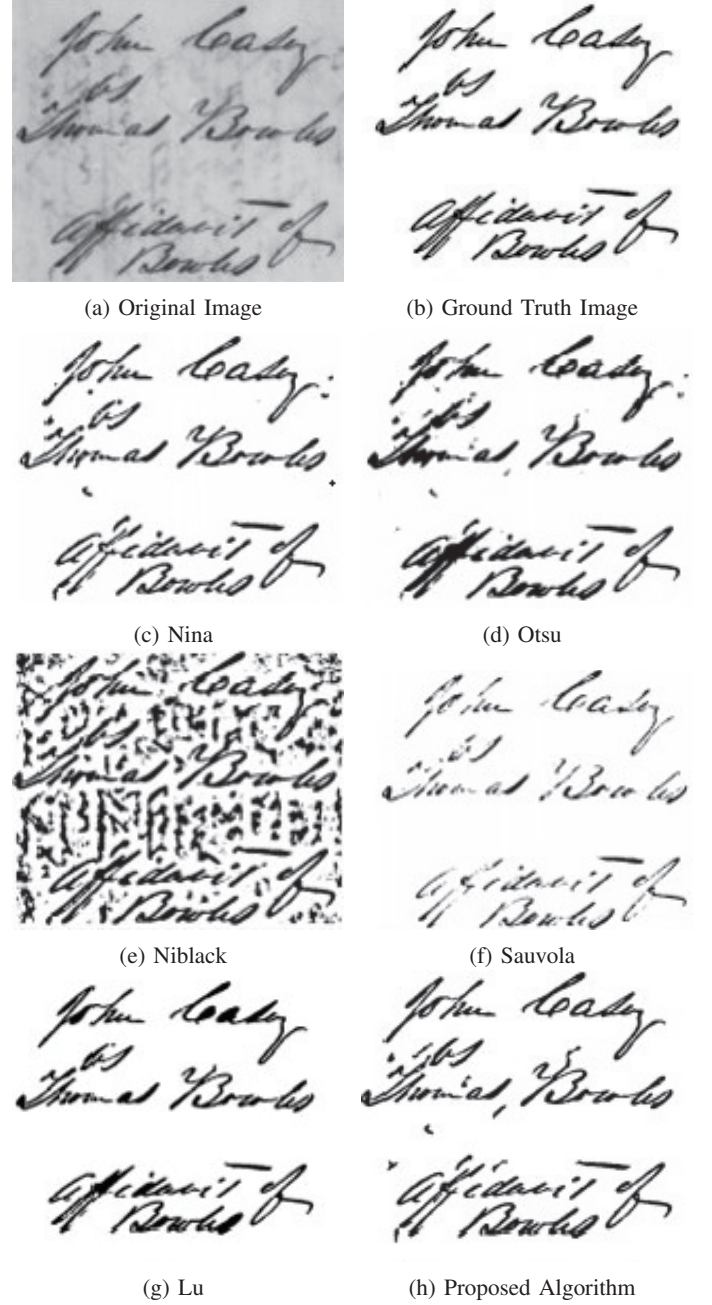


Fig. 2: Image H03 from DIBCO 2009 dataset with comparison

$$\theta = I_{min}(x, y) / I_{max}(x, y) \quad (3)$$

$$I(x, y) = I_{min}(x, y) + I_{max}(x, y)(1 - \theta)^\gamma \quad (4)$$

Equation (1) and (2) represents the calculation of maximum and minimum intensity pixel into a $(2w+1 \times 2w+1)$ sized window around the current pixel with coordinates (x, y) . Equation (3) represents the calculation of the intensity ratio factor θ . Once we calculate I_{max} , I_{min} , θ inside the window, the background estimation for the current pixel is then calculated using Equation (4) where γ is a constant. This process is

repeated iteratively for each pixel for n iterations. The reason for doing multiple iterations is the fact that bold/thick text doesn't fade away with one iteration and hence hampers the text when performing background subtraction and results in poor performance.

Algorithm:

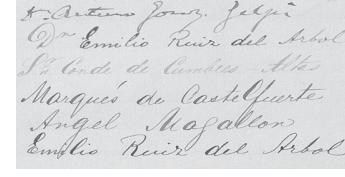
- 1) $I(x,y)$: Original Image;
- 2) w : (window size)/2;
- 3) n : number of iterations;
- 4) $C(x,y)$: Copy of Original Image
- 5) $BG(x,y)$: Estimated Background Image
- 6) For $k=1$ to n :
- 7) Slide window over each pixel and compute:
- 8) I_{min} = minimum intensity pixel(Window)
- 9) I_{max} = maximum intensity pixel(Window)
- 10) $\theta = I_{min}/I_{max}$
- 11) $BG[i][j]=I_{min} + I_{max}(1 - \theta)^\gamma$
- 12) Store background image as new copy of image.
- 13) Apply a bilateral filter to estimated background image.

Typically, the value of constant γ can vary from 1.3 to 2.5. But, we have used the $\gamma = 1.5$ so that our algorithm works well for dark as well as faint text and it also helps to reduce the number of parameters that need to be tuned. We suggest using a small window size like 5×5 for images with bold and dark visible text whereas large window sizes varying from 50×50 for faint text characters. The number of iterations is subjected to the difference between text and background intensity. A higher difference means more iterations for complete estimation. To remove the remaining noise and pixel-level irregularities inside the estimated background we use a bilateral filter over the estimated background image $BG(x,y)$. We set $\sigma_{Color} = 30$ and $\sigma_{Space} = 30$. These parameters have been tuned on H-DIBCO 2018 and perform well for H-DIBCO 2016 [23] too. The filtering effect is not so much visible on the background image but has a profound effect on binarization results. Now that we have the final estimated background image BG we move forward to the step of background removal. We remove the background by subtraction of the original image from it. Before moving on to the Gaussian smoothing we invert the image for further processing. The background removal process can be summarised as:

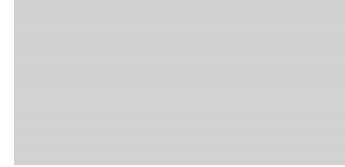
$$I_{new}(x,y) = 255 - ([BG(x,y) - I(x,y)]) \quad (5)$$

C. Gaussian Smoothing

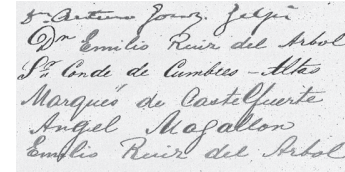
The background removal process introduces random pixel noise in the newly formed image disrupting the final output. So, before proceeding with thresholding, the image is passed through a Gaussian filter to remove these isolated black and white pixels. We have used a 5×5 gaussian kernel to filter the noise. This helps in greatly improving the binarization results.



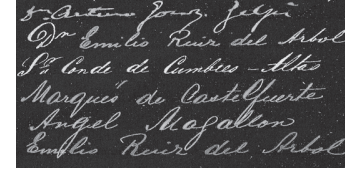
(a) Step 1: Original Image (Grayscale)



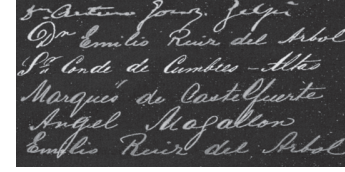
(b) Step 2.1: Background Estimation



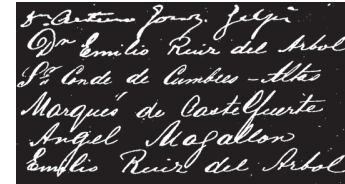
(c) Step 2.2: Background Removal



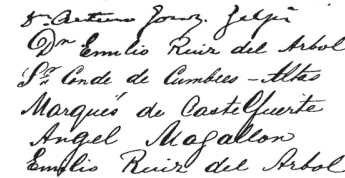
(d) Step 2.3: Inversion of Output from Step 2.2



(e) Step 3: Gaussian Smoothing



(f) Step 4: Otsu Thresholding



(g) Step 5: Spotting Removal and Inversion

Fig. 3: Overview of Binarization Algorithm. Fig 3(a) shows the input image which is already in grayscale. Fig 3(b)-3(d) represents the background estimation and removal step. Fig 3(e) shows the image after the removal of noise using gaussian smoothing. Otsu thresholding is then used to create an initial binarization in 3(f). Finally, we use spotting removal to generate the final binarization.

D. Otsu Thresholding

We use Otsu Thresholding for the binarization of our gaussian smoothened image. Otsu Thresholding is a global thresholding method used to separate text from the background. It minimizes the interclass variance between the foreground and background class or we can say maximizing the intraclass variance.

$$\sigma_B^2(T) = w_1(T)w_2(T)[\mu_1(T) - \mu_2(T)]^2 \quad (6)$$

Here σ_B^2 represents the between-class variance, we compute it for each threshold T from 0 to 255 and select the one that gives the maximum result. w_i is used to represent total pixels in the background/foreground class. μ_i represents the mean intensity for both the classes. Otsu thresholding helps us to form a pretty good approximation of the text and background but there can still be remaining noise that needs to be removed.

E. Spotting Removal

Spotting removal is the effective post-processing technique we use to remove all the remaining noise from our binary image. We do this by labelling all the connected black(text) components inside the binary image and then removing all the components of small size that are introduced while binarizing the image. Removing these small components may also remove some of the textual parts like a dot over a letter but ultimately it helps in making the binarized image clean. We select the size we want to discard manually according to the text size in an image because there is no fixed text size in handwritten documents, the dot over a character can be more than 300 pixels in one image and less than 20 pixels in another and we want to remove the maximum amount of noise possible from our binarized document. Lastly, We also take care of very large size degradations that are falsely detected as text and are much larger as compared to the text pixel components present inside the image. This can be any kind of degradation like a thick border at the corner of the document, a large ink/coffee stain, or random large spots that are often dark and hence misclassified as text. This can be easily selected according to the text size. After this, we invert the image back to make the text dark and the background white. Fig 3(f) shows the final binarization results after this post-processing(stopping removal) step. It shows how post-processing has taken care of most of the thresholding misclassification errors.

III. EXPERIMENTAL RESULTS

DIBCO datasets have images that suffer from various types of degradations and also provide the ground truth image to compare different binarization methods. These datasets consist of both degraded handwritten and printed documents. We have used images from H-DIBCO 2018 dataset to develop the algorithms as they cover a vast variety of degradations and then used it for testing the images from the H-DIBCO 2016 as well as the DIBCO 2018 dataset. We first present the qualitative results and the quantitative comparison of our method with many state-of-the-art algorithms on the H-DIBCO 2016 and H-DIBCO 2018 datasets.

A. Qualitative Results

In Fig. 2 we have compared the results of our method with Sauvola, Nina, Lu, Niblack, Otsu's method. Lu's, Nina's method and our proposed method produce good binarization results but Lu's method has made the text bold and Nina's method suffers from weak stroke detection but our method can produce an image close to the ground truth image. Fig. 4 and Fig.5 show results of our algorithm on images from H-DIBCO 2016 & 2018 datasets along with the ground truth images. Images consist of different kinds of degradations like large ink spots, thick borders, smear, severe bleed-through, faint text characters, folded pages, torn out pages, uneven illumination, smudge, shadows, low-contrast, water spilling, etc. As evident from the results, our algorithm handles all these degradations very well and can produce clean binary images with efficiently extracted text.

Before moving on to the Quantitative Results, We would first describe the evaluation measures used from DIBCO reports. DIBCO uses various benchmark evaluation methods that use mathematical calculations to measure the efficiency, accurateness of an approach in obtaining the results against the ground truth images. DIBCO 2016 and 2018 competition compute 8 statistic evaluation measures: F-measure, Pseudo F-measure, PSNR(Peak Signal to Noise Ratio) [24], DRD(Distance Reciprocal Distortion) [25], Precision, Recall, Pseudo Precision, and Pseudo Recall. It uses the four of them to decide the winner.

F-measure is calculated using two components: the precision and recall of image binarization which is obtained by counting pixels that are classified black and white correctly/incorrectly while binarizing against the ground truth images. F-measure is calculated as the harmonics mean of recall and precision.

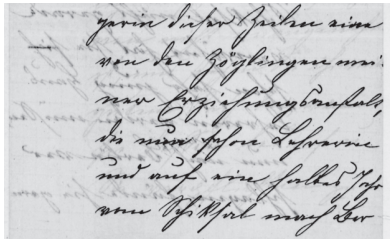
$$F_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (7)$$

where Recall = (TP)/(TP + FN) and Precision = TP/(TP + FP). TP, TN, FP, FN denotes the True Positive, True Negative, False Positive, and False Negative values, respectively.

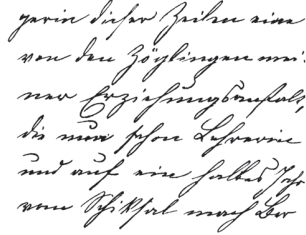
pF-measure or pseudo F-measure is a way of measuring performance similar to F-measure. The only difference being, it uses the pseudo function instead of the normal recall and precision functions. This pseudo function uses the weighted distance between the output image as boundaries of characters in the ground truth image and the boundary of characters in the extracted document. It also takes into account the local stroke width of characters.

PSNR is the relation or measure of the amount of signal in an image w.r.t amount of noise available. The higher value of PSNR implies a better signal in comparison to noise. It is used to measure the degree of reconstruction of bad lossy compression. It is really useful in the case of binarization as it can be used to measure the quality of binarization against the initial image. It helps to quantify the similarity of two images.

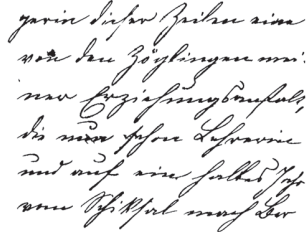
$$PSNR = 10 \log\left(\frac{C^2}{MSE}\right) \quad (8)$$



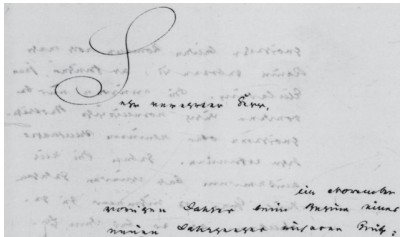
(a) Image 5 from DIBCO 2016 dataset



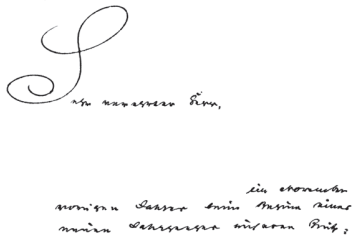
(b) Ground Truth Image



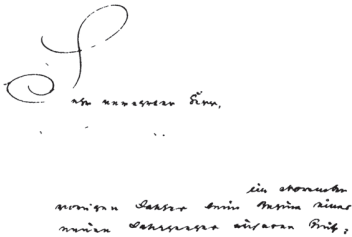
(c) Binarized Result



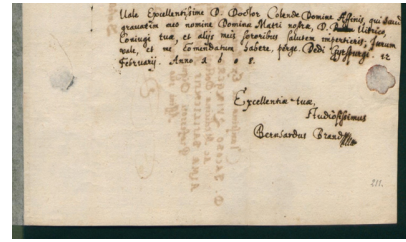
(d) Image 2 from DIBCO 2016 dataset



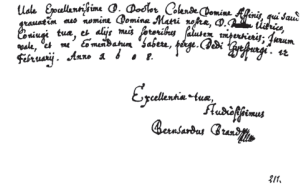
(e) Ground Truth Image



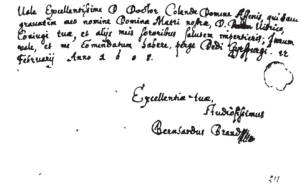
(f) Binarized Result



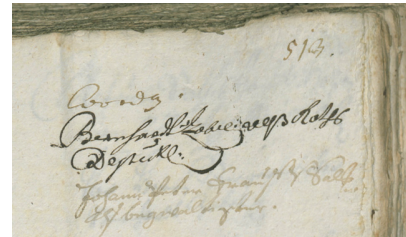
(a) Image 1 from DIBCO 2018 dataset



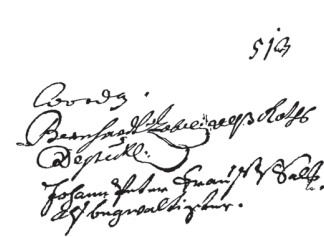
(b) Ground Truth Image



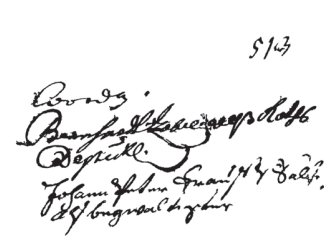
(c) Binarized Result



(d) Image 5 from DIBCO 2018 dataset



(e) Ground Truth Image



(f) Binarized Result

Fig. 4: Images from DIBCO 2016 dataset along with ground truth images and their binarized results

Fig. 5: Images from DIBCO 2018 dataset along with ground truth images and their binarized results

Where MSE is the Mean Square Error and C is a constant with a value equal to 1.

DRD is mainly used to measure distortions for binary document images. It is mainly based on the reciprocal distance. The pixels modified during binarization can have their distortions quantified using this method. DRD can be calculated as follows:

$$DRD = \frac{\sum_{k=1}^N DRD_k}{NUBN} \quad (9)$$

DRD_k represents the distortion of the k th flipped bit which is calculated using a weighted matrix of size 5×5 . Its value is given by the following formula:

$$DRD_k = \sum_{i=-2}^{i=2} \sum_{j=-2}^{j=2} |GT_k(i, j) - B_k(x, y)| * W_{Nm}(i, j) \quad (10)$$

NUBN is the total number of non-uniform pixels of the 8×8 size window inside the ground truth image. All other methods just quantitatively tell us how good our method is by checking the pixel classification but DRD tries to capture the human visual perception by measuring the perspective distortion of the binary image. Unlike other metrics lower value of DRD indicates a better binarization algorithm. All these metrics have been calculated using the DIBCO evaluation tool.

B. Quantitative Results

H-DIBCO 2016 and H-DIBCO 2018 both datasets consist of 10 handwritten document images along with their ground truth image. These images cover up a wide variety of degradations. In Table 1 and 2, we present the results of our algorithm for all the four metrics individually and averaged across all 10 Handwritten Document Images from H-DIBCO 2016 and H-DIBCO 2018 datasets respectively. Our method has been able to achieve an F-measure and pseudo-F-measure of around 90, PSNR value is greater than 18, and a DRD value of less than 4 for both of the datasets. In addition to the published results from DIBCO, we also compare our results with several traditional and state-of-the-art methods.

Image	FM	p-FM	PSNR	DRD
1	91.35	91.11	19.04	5.73
2	89.04	89.17	23.3	3.6
3	95.31	95.51	23.35	1.59
4	89.76	91.28	19.38	4.13
5	95.9	96.58	22.45	1.4
6	87.15	91.61	18.17	5.36
7	91.11	90.47	17.04	2.49
8	85.86	88.75	13.99	6.1
9	89.23	84.42	15.97	2.34
10	86.99	84.18	14.1	2.95
Average	90.17	90.31	18.68	3.57

TABLE I: Results on DIBCO 2016 dataset

In Table 3, we compare the averaged results of our method from table 1 with state-of-the-art methods like Otsu, Sauvola, Lu, Su, Howe, the winner and runner up from DIBCO 2016 dataset. Our method has achieved an FM of 90.17, p-FM of 90.31, PSNR of 18.68, DRD of 3.57, and outperformed

Image	FM	p-FM	PSNR	DRD
1	92.23	94.74	21.01	2.62
2	89.72	90.13	20	3.09
3	94.89	96.54	17.8	1.75
4	82.07	87.12	19.32	4.05
5	87.01	86.45	17.29	6.29
6	93	96.53	18.88	3.33
7	88.77	90.39	20.9	3.25
8	83.13	84.89	14.17	5.2
9	92.95	93.9	19.59	3.24
10	86.66	90.47	14.12	6.41
Average	89.04	91.12	18.31	3.92

TABLE II: Results on DIBCO 2018 dataset

both traditional and state-of-the-art methods on 3 metrics: FM, PSNR, DRD, and only its pseudo F-measure is lower than the winner's and runner up's algorithm. In Table 4, we compare our results on DIBCO 2018 dataset with Otsu, Sauvola, Winner and Runner Up from the contest. It achieves an FM of 89.04, p-FM of 91.12, PSNR of 18.31, DRD of 3.92. For this dataset as well our method outperforms all the listed methods on 3 metrics FM, p-FM, DRD and has only a lower PSNR value from Winner's method. Higher FM, lower DRD for both datasets means higher precision, recall, better pixel classification accuracy, and thus better binarization results.

Method	FM	p-FM	PSNR	DRD
Lu [16]	84.44	92.04	17.33	5.12
Su [17]	84.75	88.94	17.64	5.64
Howe [20]	87.47	92.28	18.05	5.35
Otsu [7]	86.61	88.67	17.8	5.56
Sauvola [10]	82.52	86.85	16.42	7.49
Winner's Method	87.61	91.28	18.11	5.21
Proposed Method	90.17	90.31	18.68	3.57
Lelore [26]	87.21	88.48	17.36	5.27
Runner Up	88.72	91.84	18.45	3.86

TABLE III: Comparison of performance of different methods with our method against H-DIBCO 2016 dataset

Method	FM	p-FM	PSNR	DRD
Otsu [7]	51.45	53.05	9.74	59.07
Sauvola [10]	67.81	74.08	13.78	17.69
Winner's Method [21]	88.34	90.24	19.11	4.92
Proposed Method	89.04	91.12	18.31	3.92
Runner Up	73.45	75.94	14.62	26.24

TABLE IV: Comparison of performance of different methods with our method against H-DIBCO 2018 dataset

The qualitative and quantitative results on both the datasets illustrate that our method significantly outperforms traditional and state-of-the-art algorithms. Fig 4(c),4(f),5(c) show that our method deals very well with bleed through, ink spots, spill-through. It works well even for images with faint text and preserves stroke connectivity evident from Fig. 5(f). Despite all these advantages like all methods, our method also has several limitations. Firstly, it seemed to struggle on Images 8 from DIBCO 2016 dataset and Image 10 from DIBCO 2018 dataset which has one thing in common, the background and

text have very similar colors, and hence a lot of background degradations are classified as text and result in poorly binarized images. Secondly, It also suffers when images that have black/dark colored degradation or color of degradation is similar to the text.

IV. CONCLUSION

In this paper, we have presented a novel method for binarization of degraded historical documents. The main idea is to form an approximation of the background using the information from neighboring pixels in an iterative sliding window algorithm. We use this estimated background to compensate for all the background degradations, perform binarization using Otsu, and then do effective post-processing to further clean and create a final binary image. The performance of our algorithm is greatly dependent on the selection of parameters. One is required to tune four parameters: the window size for the background estimation, the number of iterations required for background estimation, valid size range to remove maximum noise and keep maximum text(lower and upper limit). Bold text usually means more iterations for accurate approximation whereas faint text requires a larger window size to filter out the text from background. The range for text size can be selected according to the thickness and pixel size of the font present inside the image. The experimental results show that our method outperforms several traditional and state-of-the-art algorithms. But like all other methods, this method is also not perfect and suffers with images that have text color shade similar to the background, and binarization performance is affected. Our method also does not take into account the stroke width and at some places suffers in maintaining the stroke width connectivity. We would like to take this challenge in our future work to further improve our method and see applications of this background estimation technique in other fields.

REFERENCES

- [1] S. Akram, M.-U.-D. Dar, and A. Quyoum, "Document Image Processing - A Review," *International Journal of Computer Applications*, 2010.
- [2] P. K. More and D. D. Dighe, "A Review on Document Image Binarization Technique for Degraded Document Images," *International Research Journal of Engineering and Technology*, 2016.
- [3] S. Milyaev, O. Barinova, T. Novikova, P. Kohli, and V. Lempitsky, "Image binarization for end-to-end text understanding in natural images," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2013.
- [4] B. Sankur and M. Sezgin, "Image thresholding techniques: A survey over categories," *Pattern Recognition*, 2001.
- [5] P. K. Sahoo, S. Soltani, and A. K. Wong, "A survey of thresholding techniques," 1988.
- [6] P. Roy, S. Dutta, N. Dey, G. Dey, S. Chakraborty, and R. Ray, "Adaptive thresholding: A comparative study," in *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies, ICCICT 2014*, 2014.
- [7] N. Otsu, "THRESHOLD SELECTION METHOD FROM GRAY-LEVEL HISTOGRAMS," *IEEE Trans Syst Man Cybern*, 1979.
- [8] J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern Recognition*, 1986.
- [9] W. Niblack, "An introduction to digital image processing," *An introduction to digital image processing*, 1986.
- [10] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, 2000.
- [11] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 Document Image Binarization Contest (DIBCO 2009)," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2009.
- [12] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2011 Document Image Binarization Contest (DIBCO 2011)," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2011.
- [13] I. Pratikakis, K. Zagori, P. Kaddas, and B. Gatos, "ICFHR 2018 competition on handwritten document image binarization (H-DIBCO 2018)," in *Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR*, 2018.
- [14] I. K. Kim, D. W. Jung, and R. H. Park, "Document image binarization based on topographic analysis using a water flow model," *Pattern Recognition*, 2002.
- [15] B. Gatos, I. Pratikakis, and S. J. Perantonis, "Adaptive degraded document image binarization," *Pattern Recognition*, 2006.
- [16] S. Lu, B. Su, and C. L. Tan, "Document image binarization using background estimation and stroke edges," *International Journal on Document Analysis and Recognition*, vol. 13, no. 4, pp. 303–314, 2010.
- [17] B. Su, S. Lu, and C. L. Tan, "Binarization of historical document images using the local maximum and minimum," in *ACM International Conference Proceeding Series*, 2010.
- [18] O. Nina, B. Morse, and W. Barrett, "A recursive otsu thresholding method for scanned document binarization," in *2011 IEEE Workshop on Applications of Computer Vision, WACV 2011*, 2011.
- [19] O. I. Singh and O. James, "Local Contrast and Mean based Thresholding Technique in Image Binarization," *International Journal of Computer Applications*, 2012.
- [20] N. R. Howe, "Document binarization with automatic parameter tuning," *International Journal on Document Analysis and Recognition*, 2013.
- [21] W. Xiong, X. Jia, J. Xu, Z. Xiong, M. Liu, and J. Wang, "Historical document image binarization using background estimation and energy minimization," in *Proceedings - International Conference on Pattern Recognition*, 2018.
- [22] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings of the IEEE International Conference on Computer Vision*, 1998.
- [23] I. Pratikakis, K. Zagoris, G. Barlas, and B. Gatos, "ICFHR 2016 handwritten document image binarization contest (H-DIBCO 2016)," in *Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR*, 2016.
- [24] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proceedings - International Conference on Pattern Recognition*, 2010.
- [25] H. Lu, A. C. Kot, and Y. Q. Shi, "Distance-reciprocal distortion measure for binary document images," *IEEE Signal Processing Letters*, 2004.
- [26] T. Lelore and F. Bouchara, "Document image binarisation using Markov field model," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2009.

Demystifying deepfakes using deep learning

Raj Kumar Singh

Dept. of Information Technology
Delhi Technological University
New Delhi, India
rajkumarsingh_bt2k17@dtu.ac.in

Prachi Vinod Sarda

Dept. of Information Technology
Delhi Technological University
New Delhi, India
prachisarda100@gmail.com

Shruti Aggarwal

Dept. of Information Technology
Delhi Technological University
New Delhi, India
shruti_2k17it116@dtu.ac.in

Dinesh Kumar Vishwakarma

Dept. of Information Technology
Delhi Technological University
New Delhi, India
dinesh@dtu.ac.in

Abstract—Manipulation of images, videos and audios using face edit apps and web services have long been in use, since decades but recent advances in deep learning has led to rising AI generated fake images and videos with swapped faces, lip synced audios and puppet masters, popularly known as Deepfakes. Generated primarily using one of the following two approaches namely, Autoencoders and Generator Adversarial Networks which rests on trained deep neural networks, deepfakes offer unprecedented challenges. The degree of realism achieved by deep learning powered deepfakes increases with increasing amounts of data i.e. fake images and videos readily available on the internet at disposal to train GANs. Deepfake algorithms create media leaving a bare margin of difference between the authentic or original source and the forged or deepfaked targets. Thus, new mechanisms and techniques to detect and filter out such deepfakes is the need of the hour.

This paper exploits two powerful deep learning based CNN architectures namely, Inception-Resnet-v2 and XceptionNet for detecting the deepfakes. The proposed approach not only outshines the existing approaches in terms of efficiency and accuracy but also offers the best in terms of the given space and time complexity.

Keywords—Auto-encoders, DFDC, FaceForensics++, GAN, Inception-ResNet-v2, MesoNet, XceptionNet(key words)

I. INTRODUCTION

The epistemology of deepfakes derives basically from "deep learning" plus "fakes". It is a general term that covers fake videos, images, audio & other media synthesized using AI powered deep learning techniques. It can also be understood as a technique which is employed to replace the face of a target person('input') on that of a source person('output') either in a video or image [1]. The source individual thus impersonates the target individual and does actions or gives speeches which can lead to misinformation propaganda during election campaigns impacting the fair electoral process [2], hampering the social image of prominent personalities or celebrity defamation and fake news. The term first surfaced in 2017 when a video tagged as celebrity porn replaced the face of a porn actor with a celebrity through a machine learning algorithm developed by a Reddit user [3]. Ever since then, celebrity non-consensual pornography targeting famous personalities and politicians are the major commercial areas where deepfakes have been used. In 2018, a very short video about a minute proliferated every social

media where former US President Barack Obama is seen delivering an infamous hate speech which he never said [4].

Lately, the CEO of a company was duped out of \$243,000 using a deepfake audio [5]. Thus, it is clearly evident that deepfakes can very well lead to a constitutional crisis, civil & military unrest, cause religious & socio-political tensions between warring factions & countries and are a potent threat to privacy, security and national integrity. This calls for the ever-increasing need for certifying the integrity and authenticity of digital visual media.

The other side of deepfakes offers remarkable positive applications allowing for reshooting sequences of movies films in the absence of the actor as happened in the Fast & Furious series [6]. Deepfakes were employed to deliver a very high degree of photo realism in the scenes. It can also be used to provide audio to actors who have lost their voices or character voice mismatch in case of artists, deepfakes can offer realistic audio visuals with visuals of another person lip-synced with the voice of another.

It is necessary to bring out the difference between content synthesized using image manipulation tools like Adobe Photoshop and AI synthesized deepfakes. The foundations of deepfakes rests on deep neural networks trained & tested on large amounts of fake & real face images & videos data to naturally map the facial features, expressions & other face artifacts between the source and the target.

The capacity of deep learning techniques to handle complex & huge volumes of data is exploited for deepfakes generation. A large number of input samples of fake images & videos increases the photo-realism achieved in the output deepfake. The Obama deepfake was produced from a GAN which used 56 hours of sample input videos to replicate the exact lip, head, eye artifacts in the face [4]. In the case image edit apps like photoshop limited facial manipulations are allowed owing to want of complicated editing tools and domain proficiency required. Making photorealistic swaps using such apps is a very complex and time-consuming process. During incipient stage, a deepfake video could be easily identified through human eyes owing to the phenomenon of pixel collapse which gives rise to unnatural visual artifacts in the skin, face etc., and resolution inconsistency in images and others [7]. But, the recent developments in deep networks technologies and the free availability of large amounts of data produces deepfakes

that cannot be differentiated using either direct human observation skills or sophisticated computer algorithms.

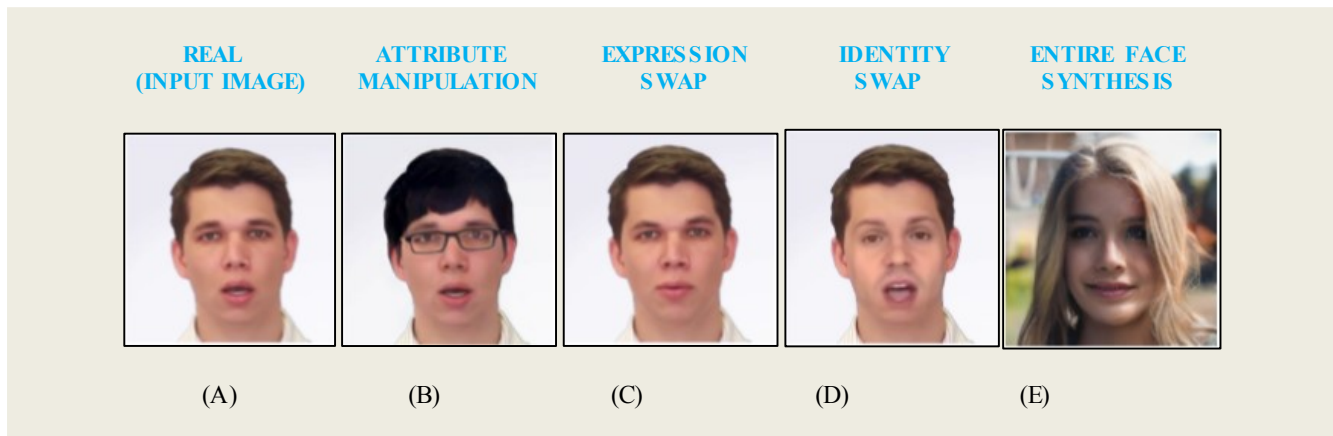


Fig. 1. Four Types of Facial Manipulations (A) Input Image. (B) Attribute Manipulation using StarGAN. (C) Expression swap using GANs. (D) Identity Swap using Face Swap and ZAO app. (E) Entire Face Synthesis using StyleGAN.

This requirement of a large quantity of image and video data for training models also explains why celebrities & politicians became and still are the first targets of deepfakes [2].

The paper has been divided into six sections - Section II revisits the existing literature that was thoroughly surveyed on the topic by us, Section III provides with our approaches, the dataset used along with the preprocessing done, vision behind the methods and a theoretical analysis of the CNN architectures employed. Section IV is the experimentation and results block where it demonstrates how proposed approach performs better than the existing state of the art techniques. Finally, Section VI concludes the paper highlighting the quintessential nuances of our approaches and gives our perspectives on the future works that intend to explore in this domain. The paper ends by mentioning the acknowledgment and references.

II. LITERATURE REVIEW

In 2018, Korshunov and Marcel [8] worked on the first generation fake videos and based their approach on the irregular disharmony measured between the lip movements and audio speech, biometric variations in images etc. For the lip sync inconsistency with speech, the audio features were represented using Mel Frequency Cepstral Coefficients (MFCCs) and visual features were represented by separation gaps between mouth landmarks. PCA was then employed for dimensionality reduction followed by Recurrent Neural Networks (RNNs). A set of 129 features related to Image Quality Measurements along with PCA with LDA, or SVM was used to detect fake videos. This proposed detection approach was tested & trained on the Deepfake TIMIT database.

Yang et al. [9] observed in 2019 that most of the deepfakes were generated by simply copying and pasting front portion of source person on target person and the inconsistencies could easily be estimated by considering 3D head poses and all round facial features. Their approach was based on the

mismatches between 360° head poses which were measured using a set of 67 visual artifacts or facial landmarks in the central facial region to classify between real and deepfake videos.

After extraction of the said features, these were normalized followed by an SVM classifier. This proposed approach against the UADFV database failed to generalize well over other databases.

Li & Lui [10] in 2019 exploited the face manipulation pipeline to automatically extract the inconsistencies in the visual artifacts. Their approach was based on the hypothesis that deepfake images are limited by their resolution capacity. It is only through proper post processing such face warping artifacts can be corrected. Such transformations owing to pixel damage and collapse result in distinctive artifacts in the detected face regions and surrounding regions which can be easily estimated using a system employing CNNs. Thus, VCGI16, ResNet50, ResNet101 and ResNet152 were trained from scratch on UADFV and Deepfake TIMIT databases. This approach which made use of four different CNN models together was able to outperform the existing approaches in terms of accuracy. By the end of the year, Li et al [11] presented modifications to the above approach improving its accuracy. It made use of a pyramid structure spatial module to match the resolution inconsistencies on the face. The approach was quite generalized achieving near optimum results over varied databases.

Mesoscopic features along with steganalysis features based approaches also exist in deepfake detection literature. In 2019, Afchar et al. [12] employed two different CNN networks, (i) a four layer CNN + fully connected layer(Meso4) and ii) a four layer CNN +MesoInception-4(improved Meso 4 using an inception module), were based on the mesoscopic properties of the images. Their approach though trained and tested on their own database was very

robust and performed similarly when FaceForensics++ dataset was used.

In early 2020s it was realized that deepfakes absolutely lacked voluntary and unconscious human eye blinking movements since the internet lacked images & videos of the public figures having their eyes shut. This was exploited by Jung et al. [13] with their very famous approach called DeepVision of combating deepfakes basing their approach on eye blinking patterns. Face to Eye aspect ratio was calculated using the visual artifacts Fast-HyperFace and Eye-Aspect-Ratio (EAR) in combination. Eye blinking period and others features were extracted to differentiate between real and deepfaked videos. They worked on their own proprietary database and achieved an accuracy of 87.5%.

Rossler et al. [14] in 2020 conducted a detailed analysis of the existing deepfake detection techniques using the flagship FaceForensics++ database. His evaluation considered five different systems. The first system was a CNN model trained using handcrafted steganalysis features. The layers of the second CNN-based system were designed as such that they mitigate the high-resolution content in the images. In the third CNN based system, statistical measures of central tendency were computed using a global pooling layer. The fourth system was same as used by Afchar et al. [12] 4 layer CNN followed by a fully connected MesoLayer(Meso4) powered with an Inception module and the final system based on XceptionNet architecture based CNN system which had been originally trained using the Imagenet dataset which was again retrained to perform the deepfake classification task. These systems were tested for their accuracy and efficiency in both deepfakes and faceswaps. The results showed that the CNN model based on the XceptionNet architecture outperformed state of the art existing approaches.

Finally, Tolosana et al. [15] in 2020 employed the deep learning CNN architecture XceptionNet. Facial regions were differentiated based on their assigned discriminative power. The experimental framework considered both 1st and 2nd generations deepfake databases. The approach performed poorly when run against the 2nd generation deepfake dataset but fairly well in case of 1st generation deepfake videos. 91.0% AUC for DFDC preview dataset and 83.6% AUC for the Celeb-DF dataset were achieved. The highlight of the research was that they had trained separate fake detection systems for each of the databases.

To sum up, although many varied approaches have been devised by researchers, all of them failed to generalize their results when run against unseen databases. While we were surveying the literature, we found that Inception-ResNet-v2 deep neural network model has not been employed yet in detection of deepfakes, though XceptionNet was used in few approaches. We base our model based on the Inception-Resnet-v2 architecture and show in this paper that it outperforms the existing approaches and shows commendable results for deepfake databases from the 2nd generation as well

III. PROPOSED METHODOLOGY

In this study, we explore two different neural networks - InceptionResNet-v2 and XceptionNet, which are popular in the image recognition applications for their good performance. Here, we employ these two models to application of deepfake detection. We train both models of DFDC dataset and observe their performance in detecting deepfakes. The results of this experiment came out to be pretty good as compared to the present state-of-the-art and are presented in the below sections.

A. XceptionNet

Xception stands for extreme Inception as it takes the principles of Inception to an extreme. XceptionNet is a traditional convolution neural network trained on ImageNet based on separable convolutions with residual connections. Forming the feature extraction base of the network, the Xception architecture has 36 convolutional layers. These convolution layers are structured into 14 modules, each of which except for the first and last module have linear residual connections around them. The 14 modules are grouped into three groups viz. the entry flow, the middle flow, and the exit flow. And each of the groups has four, eight, and two modules respectively. The final group, i.e. the exit flow, can optionally have fully connected layers at the end.

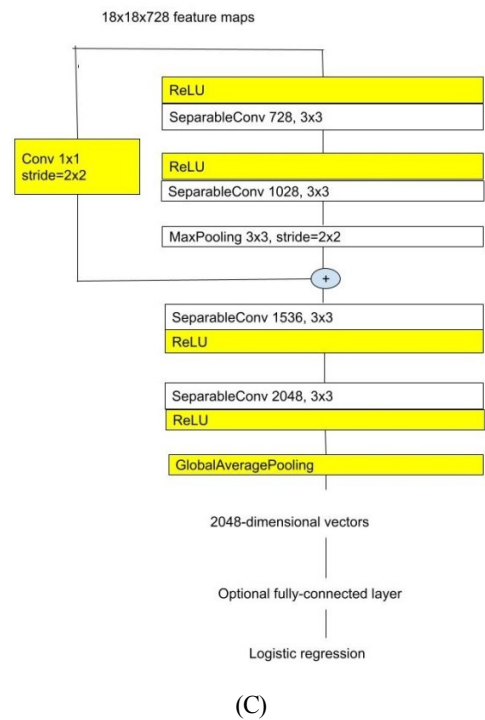
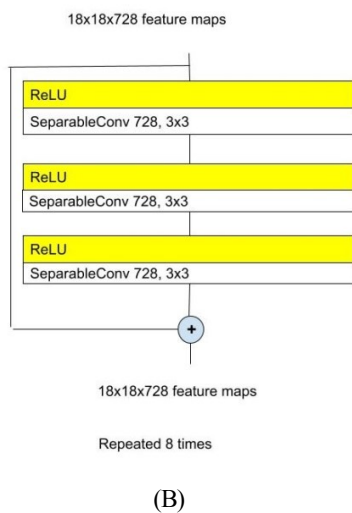
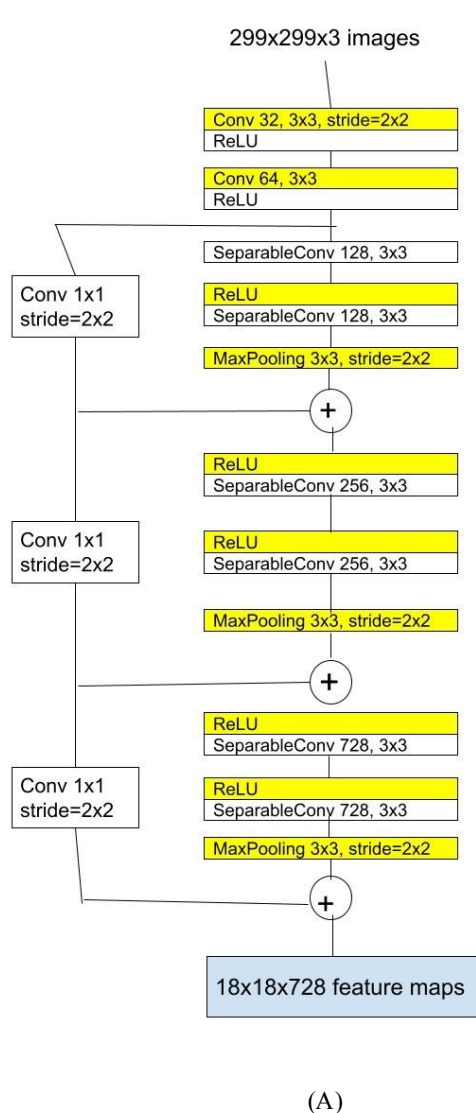


Fig. 2. The Xception architecture: the data first goes through the entry flow, then through the middle flow which is repeated eight times, and finally through the exit flow. Note that all Convolution and Separable Convolution layers are followed by batch normalization. (A) Entry Flow, (B) Middle Flow, (C) Exit Flow.

We transfer it to our task by changing the input shape from a default value of (299, 299, 3) to (128, 128, 3) as satisfied by the images in our dataset. Additional layers are added to the previously trained model - global max pooling layer, and a global average pooling layer following it, giving output probabilities for two classes - fake, not fake. The model is trained for 30 epochs and we save the best performing mode based on validation accuracy.

B. Inception-ResNet-v2

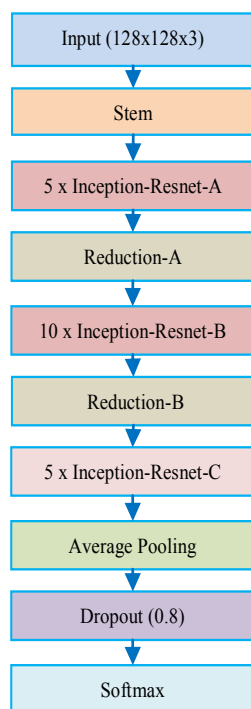


Fig. 3. Overall architecture of Inception-Resnet-v2 network.

Inception and Resnet architectures [16] have played a key role in image recognition advances in recent years, with demonstrated good performance at comparatively low computational costs. Inception-ResNet-v2 architecture combines the Inception architecture, with the concept of residual connections.

Inception-ResNet-v2 is a convolutional neural network that is based on the family of Inception architectures [17], with residual connections [18] imposed on it (in place of the filter concatenation step of the Inception network). The network is 164 layers deep, trained on images of the huge ImageNet database.

The overall architecture of Inception-Resnet-v2 network has been shown in Fig 7.

The detailed architecture of stem, inception blocks and the reduction blocks of the network can be seen in Fig 8, Fig 9 and Fig 10 respectively.

In our work, we perform transfer learning on Inception-Resnet-v2 network, by adding additional layers to the previously trained model - global max pooling layer, and a dense classifier layer following it, giving output probabilities for two classes - fake, not fake. The network is trained over input images of shape (128, 128, 3), extracted from the training set of DFDC videos. The model is trained for 30 epochs, yielding a low model validation loss on the last epoch.

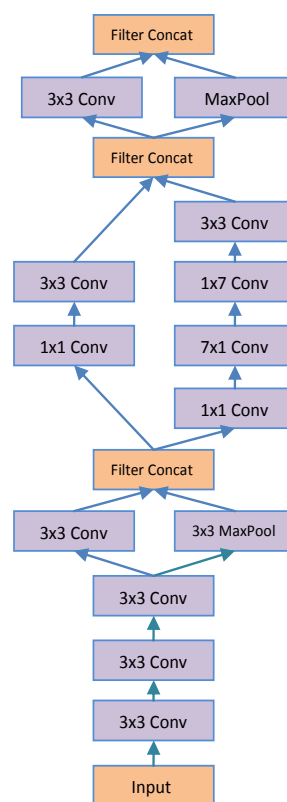


Fig. 4. Stem of Inception-Resnet-v2 network architecture.

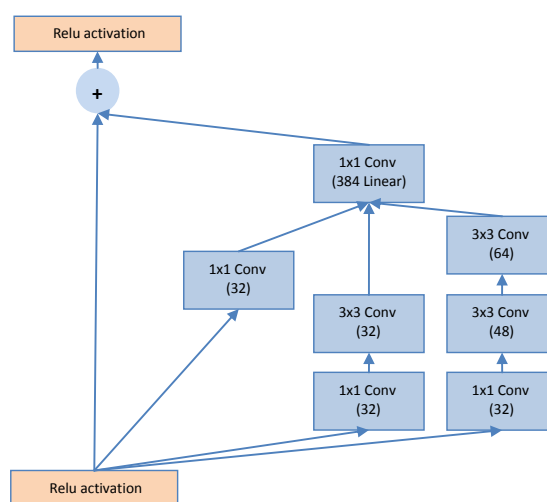


Fig. 5. Inception-Resnet-A block of Inception-Resnet-v2 network architecture.

The model's observed training and validation - loss and accuracy, are shown in Fig 12 in the next section.

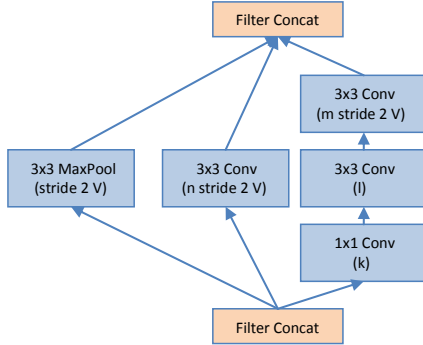


Fig. 6. Reduction-A block of Inception-Resnet-v2 network architecture with values of k,l,m,n as 256, 256, 384, 384 respectively.

IV. EXPERIMENTAL WORK

A. Tools and Libraries Used

All performed experiments (Preprocessing, Feature Extraction, Training and Testing of Various ML Models) have been performed using the Python Language (v3.8.5) on the open source Jupyter environment. The Graphic Card NVIDIA GeForce RTX 2060 (mobile) was used for training. The following packages were used: Python Imaging Library (PIL), OpenCV, Numpy and Pandas for basic calculations, data retrieval, cleaning, processing, and visualization; Scikit Learn for importing Machine Learning Models; Tensorflow and Keras for building an Artificial Neural Network; json, dlib and video capture (openCV) for feature extraction, selection, and preprocessing; and finally, Matplotlib for plotting the graphs.

B. Dataset and Preprocessing

The dataset used here for our experimental work is the DFDC (Deepfake Detection Challenge) Preview dataset. A number of datasets featuring video forgery exist, but we found DFDC to be quite diverse in terms of the gender, skin-tone, age and race of the people in the videos. To include varied head poses and lighting conditions, and yield visually diverse backgrounds, participants were allowed to record videos with any background of their choice for visual variability. The DFDC preview dataset consists of nearly 5,000 videos, including 1,131 real and 4,119 fake videos, created using two different facial modification algorithms.

We preprocess the videos from the dataset. For each video, we extract image frames out of them. To play a video from a file, we first create a Video capture object from the OpenCV library in Python. We then set the frame rate to the current position of the video file in milliseconds and then we extract the 0-based index of the frame to be decoded/captured next. After reading the frame-Id and comparing it with the frame rate, we then extract the image frame, resize to a (128, 128) and save it to a new folder thus creating a dataset of the image frames from the videos in DFDC.

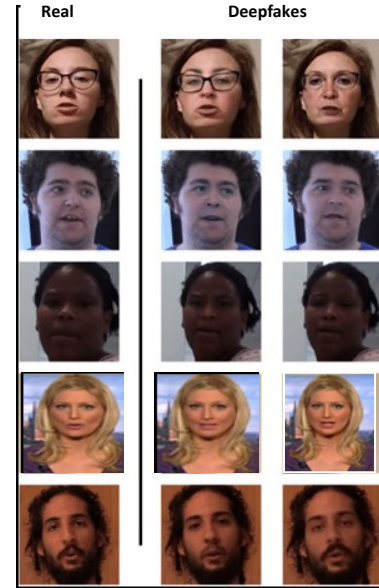
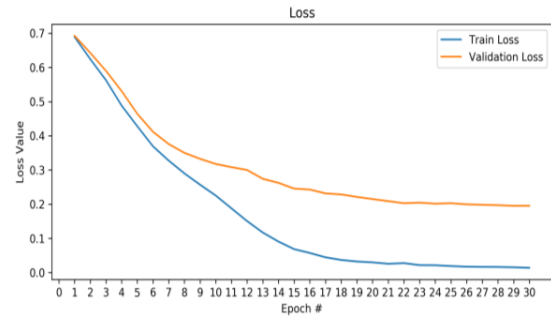
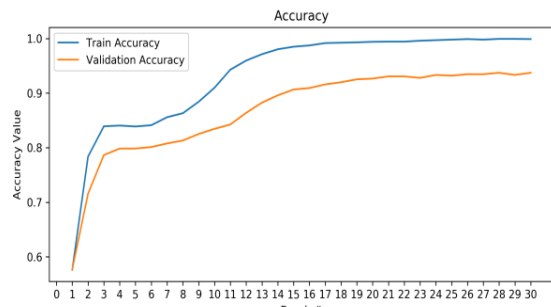


Fig. 7. Some Real and DeepFake examples from DFDC dataset

C. Training and Results

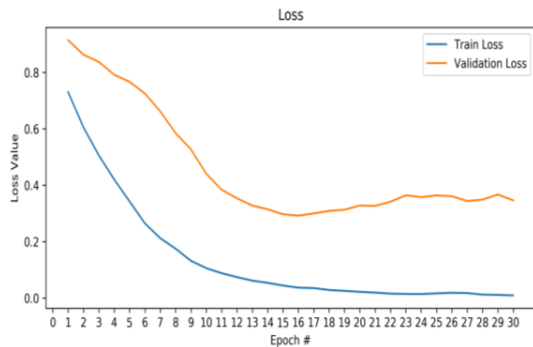


(A)

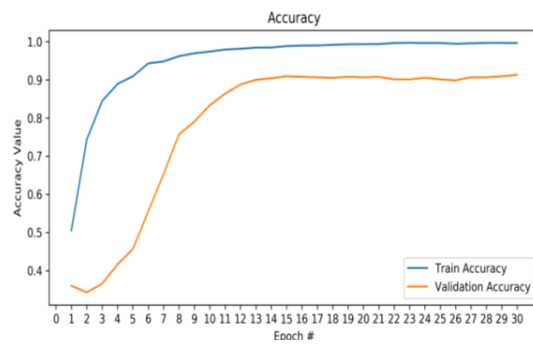


(B)

Fig. 8. Pre-trained-ResNet-v2 transfer learn with fine tuning and Image Augmentation Performance. (A) Loss Value vs. Epoch Curve, (B) Accuracy Value vs. Epoch Curve, for Training and Validation.



(A)



(B)

Fig. 9. Pre-trained Inception-ResNet-v2 transfer learn with fine tuning and Image Augmentation Performance. (A) Loss Value vs. Epoch Curve, (B) Accuracy Value vs. Epoch Curve, for Training and Validation.

TABLE I. CLASSIFICATION REPORT OF XCEPTIONNET

Class	Precision	Recall	F1-score	Accuracy
0 (DeepFake)	0.80	1.00	0.89	0.80
1 (Real)	0.90	0.06	0.11	

TABLE II. CLASSIFICATION REPORT OF INCEPTION-RESNET-V2

Class	Precision	Recall	F1-score	Accuracy
0 (DeepFake)	0.92	0.95	0.94	0.90
1 (Real)	0.79	0.71	0.75	

TABLE III. COMPARISON WITH BENCHMARK DEEPFAKE DETECTION METHODS USING DFDC PREVIEW DATASET

Study	Method	Classifier	Performance
Afchar et. al [12]	MesoNet: a Compact Facial Video Forgery Detection Network	CNN	AUC = 75.3%
Zhou et. al [19]	Two-Stream Neural Networks for Tampered Face Detection	CNN/SVM	AUC = 61.4%
Matern et. al[7]	Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations	MLP	AUC = 66.2%
Yang et. al [11]	Exposing Deep Fakes Using Inconsistent Head Poses	SVM	AUC = 55.9%
Li et. al [10]	Exposing DeepFake Videos By Detecting Face Warping Artifacts	CNN	AUC = 75.5%

Though we achieved a good accuracy of 80%, when the dataset was modeled on XceptionNet, an observation that can be made here is that in the case when an image is not as a deepfake (class 1), the XceptionNet produces an f1-score of 0.11, which is poor and Inception-Resnet-v2 having an f1-score of 0.75 out-performs XceptionNet in this case but for a case when the image is as a deepfake (0), both models produce quite competitive performance with XceptionNet producing an f1-score of 0.89 and Inception-Resnet-v2 giving an f1-score of 0.94. So both these neural networks give promising results in the case of a deepfake image.

Since our primary concern here is to identify if an image is deepfake as accurately as possible, so as to deal with the hazardous damages produced by them today, we can conclude that XceptionNet does serve the purpose upto a good extent, but Inception-ResNet-v2 performs better than XceptionNet for deepfake detection. Thus, becoming our proposed method for deepfake detection.

Another crucial comparison metric is the AUC value of the classifier. The AUC metric signifies Area under the Curve and demonstrates classifier's ability to differentiate between output classes. It is used as a summary for the classifier's ROC curve. Higher the model's AUC score better is the model's ability of differentiating between the positive and negative classes. So here we have plotted the ROC curves for both our neural networks and also calculated the AUC ROC score for both.



Fig. 10. ROC curves for XceptionNet and Inception-Resnet-v2 with respective AUC values.

Nguyen et. al [1]	Multi-task Learning For Detecting, Segmenting Manipulated Facial Images and Videos	AE + MTL	AUC = 53.3%
Dolhansky et al.[20]	The DeepFake Detection Challenge (DFDC) Preview Dataset	CNN	Precision = 93%
Tolosana et al. [15]	DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance	CNN	AUC = 91%
OUR METHODOLOGY	Demystifying DeepFakes Using Deep Learning Architecture Inception-ResNet-v2	CNN	Precision = 92% Accuracy = 90% AUC = 83%

design of more robust solutions to tackle the growing complexity in the field.

Let us now compare our models with existing state of the art approaches. While we were surveying related works, we found that it was only a few years ago that researchers started exploiting deep learning architectures to detect deepfakes. Since, detection of deepfakes is basically a classification task various approaches can be compared based on their performances calculated either from Precision score, Recall score, Accuracy and AUC. Table III compares the state of the existing approaches with the method devised by us mentioned in this work. All the models have been tested against the flagship Deepfake Detection Challenge Dataset provided by Facebook available at Kaggle. Our proposed approach outperformed all the existing approaches for each of the performance scores. Our model making use of Inception-ResNet-v2 was able to achieve performance values – 92% Precision, 95% Recall, 90% Accuracy and an AUC equals 0.83, which is one of the highest performances in state-of-the-art achieved till date.

V. CONCLUSION AND FUTURE WORK

This paper proposed efficient, near optimal models for detection of deepfakes. The proposed models were based on powerful deep neural network architectures and therefore find universal applicability in other scenarios as well. We also provided the readers with the insight of the vision, a theoretical analysis of the various techniques and methods employed in the detection of deepfakes. Our models were able to produce very good accuracy scores, with XceptionNet at 80% and Inception-Resnet-v2 at 90%. The highlight of the paper was that we worked with the best real life video dataset available today, curated by Facebook Inc. in collaboration with Microsoft Corp. for their million-dollar DeepFake Detection Challenge. In future works, we intend to extend the applicability of our methodology to solving other image and video forgery activities and experiment with other versions of the proposed CNN architectures to improve the accuracy further. We also are working on other flagships datasets like FaceForensics++ and CelebDF to verify the applicability of our proposed architecture. The limitations of using Inception-ResNet-v2 and in detecting deepfakes is that when training on regular images and testing on negative images, the model accuracy is significantly lower than when it is tested on regular images. Therefore, current training methods do not effectively train models to generalize the concepts. It would certainly not be an exaggeration to say that our analysis will aid in the

VI. REFERENCES

- [1] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep learning for deepfakes creation and detection," *arXiv*, pp. 1–12, 2019.
- [2] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," vol. 31, no. 2, pp. 211–236, 2017.
- [3] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A Survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, 2020, doi: 10.1016/j.inffus.2020.06.014.
- [4] "How faking videos became easy and why that's so scary," *Bloomberg*.
- [5] J. Damiani, "A voice deepfake was used to scam a CEO out of \$243,000," 2019.
- [6] B. Marr, "The best (and scariest) examples of AI-enabled deepfakes," 2019.
- [7] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," *Proc. - 2019 IEEE Winter Conf. Appl. Comput. Vis. Work. WACVW 2019*, pp. 83–92, 2019, doi: 10.1109/WACVW.2019.00020.
- [8] P. Korshunov and S. Marcel, "DeepFakes: A new threat to face recognition? Assessment and detection," *arXiv*, pp. 1–5, 2018.
- [9] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2019-May, pp. 8261–8265, 2019, doi: 10.1109/ICASSP.2019.8683164.
- [10] Y. Li and S. Lyu, "Exposing DeepFake Videos by Detecting FaceWarping Artifacts," *CVPRW*, 2019.
- [11] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3204–3213, doi: 10.1109/CVPR42600.2020.00327.
- [12] D. Afchar *et al.*, "MesoNet : a Compact Facial Video Forgery Detection Network To cite this version : HAL Id : hal-01867298 MesoNet : a Compact Facial Video Forgery Detection Network," 2018.
- [13] T. Jung, S. Kim, and K. Kim, "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern," *IEEE Access*, vol. 8, pp. 83144–83154, 2020, doi: 10.1109/ACCESS.2020.2988660.
- [14] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," *arXiv*, 2019.
- [15] R. Tolosana, S. Romero-Tapiador, J. Fierrez, and R. Vera-Rodriguez, "DeepFakes evolution: Analysis of facial regions and fake detection performance," *arXiv*, 2020.
- [16] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017, pp. 4278–4284.
- [17] C. Szegedy, V. Vanhoucke, S. Ioffe, and J. Shlens, "Rethinking the Inception Architecture for Computer Vision."

- [18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." Accessed: Feb. 07, 2021. [Online]. Available: <http://image-net.org/challenges/LSVRC/2015/>.
- [19] I. Demir and U. A. Ciftci, "Where Do Deep Fakes Look? Synthetic Face Detection via Gaze Tracking," pp. 1–14, 2021.
- [20] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," *arXiv*, 2019.



Design and Analysis of a Bandpass Filter Using Dual Composite Right/Left Handed (D-CRLH) Transmission Line Showing Bandwidth Enhancement

Priyanka Garg¹ · Priyanka Jain¹

Accepted: 13 April 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

A compact, low-profile, Band Pass Filter (BPF) based on balanced Dual Composite Right/Left Handed (D-CRLH) Transmission Line (TL) is proposed in this article. A balanced D-CRLH TL can be used to provide wideband filter characteristics due to no frequency separation between the RH and LH frequency bands. The proposed D-CRLH TL is designed using U-shaped complementary split-ring resonator (UCSRR). The extraction of equivalent circuit model of the proposed structure is also performed. The proposed filter provides a 3 dB passband range from 2.44 to 5.58 GHz. Further, the bandwidth is enhanced by introducing a slot in UCSRR, which resulted in a 3 dB passband range from 1.43 to 5.56 GHz. The proposed via less BPF has a compact size of $15 \times 15 \text{ mm}^2$ designed on an FR-4 substrate with dielectric constant (ϵ_r) = 4.3. The design analysis of the proposed bandpass filter is presented in terms of reflection coefficient, transmission coefficient, propagation constant and group delay.

Keywords Bandpass filter (BPF) · Dual Composite Right/Left Handed (D-CRLH) Transmission Line (TL) · Metamaterial · U-shaped complementary split ring resonator (UCSRR) · Microstrip line

1 Introduction

Bandpass filters play an important role in various types of radio frequency (RF) and microwave systems in order to accept a particular band of frequencies. With recent advancement in wireless communication technology, requirement of highly efficient, miniaturised and low cost bandpass filter is also increasing. Several approaches have been investigated till date to design high performance filters. Amongst them microstrip

✉ Priyanka Jain
priyjain2000@rediffmail.com

Priyanka Garg
priyankagarg_phd2k16@dtu.ac.in

¹ Department of Electronics and Communication Engineering, Delhi Technological University, New Delhi 110042, India

technology offers the advantage of compact dimensions, easy integration with circuit elements as well as self tuning capabilities.

The requirement of wideband bandpass filters has increased exponentially in past few decades. Various techniques are proposed in literature to achieve wideband filter characteristics such as coupled line structure [1, 2], ring resonators [3], step impedance resonator [4]. In [5], Sassi et al. developed bandpass filter by linking hexagonal-omega resonators to microstrip lines. However in these cases the performance achieved had a trade-off with the overall size of the structure.

Further, size reduction is achieved when the filters are realised by etching slots either on the ground plane or on the microstrip line termed as Defected ground structure (DGS) and Defected microstrip line (DML) respectively. In [6], CSRRs are etched on the microstrip line to achieve low pass characteristics which further extended in [7] to achieve bandpass filter characteristics by etching CSRR on the ground plane and interdigital capacitor on the microstrip line to operate on K-band. Although the bandwidth achieved here is narrow.

In [8], the concept of metamaterial transmission lines (TL) is introduced in 2002 to achieve wideband structures. These TL metamaterials typically exhibit a LH band at lower frequencies and a right-handed (RH) band at higher frequencies and are termed as Composite Right/Left Handed Transmission lines (CRLH TL) [9, 10]. Mirroring to conventional CRLH behaviour, a novel metamaterial with LC parallel-tank impedance and LC series-tank admittance, termed Dual-CRLH TL (D-CRLH) was proposed by [11]. Gonzalez et al. [12] proved that the new structure results in larger bandwidth and lower losses compared to conventional CRLH TL. While, an unbalanced D-CRLH TL can be used to design dual band filters with high rejection between the pass bands [13] or bandpass filter with notch bands [14], a balanced D-CRLH TL can be used to design a filter with wide transmission bandwidth and low losses [15]. In the case of a balanced D-CRLH TL, the right handed passband (at lower frequency) is followed by the left handed passband (at higher frequency) without any frequency separation between them.

To the authors knowledge, few state of the art has been available that utilize D-CRLH TL as a balanced TL to design bandpass filters. Belenguer et al. [15] have designed a balanced dual-CRLH TL just by modifying the well known Split ring resonator (SRR) to show wider transmission bandwidth. Cano et al. [16] have extended the same work by enhancing the design to provide reconfigurable bandwidth and propagation characteristics. The present work deals with improving the design complexity as well as the passband bandwidth within a compact structure.

In this paper a compact sized balanced D-CRLH TL based wideband bandpass filter utilizing U-shaped Complementary Split Ring Resonator (UCSRR) is proposed. The paper starts with extraction of the material parameters followed by designing and analysis of D-CRLH TL bandpass filter in the Sec. 2. Then an equivalent circuit model is presented in Sect. 3 followed by discussion on bandwidth enhancement technique and dispersion characteristics. The proposed design is a via less, D-CRLH TL based wideband bandpass filter designed to provide a 3 dB passband from 2.44 to 5.58 GHz which further is increased from 1.43 to 5.56 GHz on cutting a slot between the two U-shaped resonators (details are discussed in later sections). Finally the measured results are compared with the simulated ones in Sect. 4. All the design simulations are performed using Computer Simulation Tomography (CST) Microwave Studio [17] and circuit simulations are carried out using Advance Design System (ADS) [18].

2 Design and Simulation Approach

2.1 Extraction of Material Parameters of UCSRR

The design of proposed bandpass filter is based on U-shaped complementary split ring resonator (UCSRR) which exhibits negative refractive index near its resonant frequency. This type of structure provides 180° rotational symmetry, thus avoiding cross polarization and also resulting in smaller electrical size as compared to conventional CSRR [19]. The extraction of material parameters is performed using time domain solver of CST Microwave studio. The initial dimensions of the UCSRR are taken to be according to the sub-wavelength rule of metamaterial i.e. less than $\lambda/4$. As the centre frequency is nearly 4 GHz for the intended band so the major dimension of the UCSRR is chosen to be 11.5 mm which is optimized further. The UCSRR is designed on a FR-4 substrate with relative permittivity = 4.3, and thickness = 1.6 mm, followed by assigning appropriate boundaries. The structure is designed along XY plane where electric boundary is assigned to x-axis, magnetic boundary to y-axis and z-axis is used to assign ports (direction of propagation). Figure 1 illustrates the geometry and port assignment of UCSRR unit cell.

Now, effective medium parameters are extracted from the scattering parameters using the modified Nicolson-Ross-Weir (NRW) relations presented in [20]. Taking two arbitrary variables given by following equations:

$$V_1 = S_{21} + S_{11} \quad (1)$$

$$V_2 = S_{21} - S_{11} \quad (2)$$

Complex normalised wave impedance(z) can be obtained by following equations:

$$z = \sqrt{\frac{(1 + S_{11})^2 - S_{21}^2}{(1 - S_{11})^2 - S_{21}^2}} \quad (3)$$

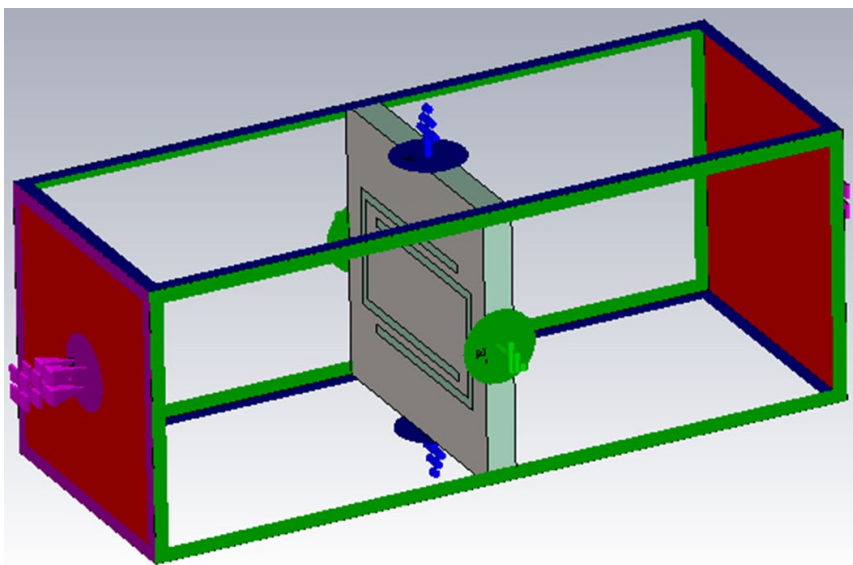


Fig. 1 Geometry of UCSRR unit cell showing boundary assignment

Now, permeability(μ_r), permittivity(ϵ_r) and refractive index(n) are obtained simply by:

$$\mu_r = \frac{2}{jkd} \cdot \frac{1 - V_2}{1 + V_2} \quad (4)$$

$$\epsilon_r = \frac{2}{jkd} \cdot \frac{1 - V_1}{1 + V_1} \quad (5)$$

$$\epsilon = n/z \quad (6)$$

$$\mu = nz \quad (7)$$

where k is the free space propagation constant and d is the unit cell dimension. Figure 2a, b respectively shows the real and imaginary part of the extracted material parameters. It can be observed that the proposed U-CSRR exhibits negative real effective permittivity and permeability in the common frequency range 4.3 GHz to 5.2 GHz, and positive value of imaginary permittivity and permeability for the same range. This results in negative refractive index which means that the proposed unit cell behaves as a left handed material (LHM) in the frequency range 4.3 GHz to 5.2 GHz.

2.2 Design of D-CRLH TL Bandpass Filter

Proposed D-CRLH TL based wideband bandpass filter is designed on an FR-4 dielectric substrate having compact dimensions $15 \times 15 \times 1.6 \text{ mm}^2$. Falcon et al. [21] presented that a microstrip transmission line incorporated by CSRR on the ground plane shows bandstop characteristics, further, on loading the microstrip line with capacitive gaps, the bandstop behaviour switches to bandpass one. Thus, the proposed structure of BPF consists of a 50- Ω microstrip line with capacitive gap on top side and U-shaped CSRR etched on the bottom ground plane. The dimensions of microstrip line are chosen to deliver 50- Ω characteristic impedance. Figure 3 illustrates the geometry and dimensions of the proposed bandpass filter. The slot, shown in Fig. 3c is introduced between the two U-shaped slots

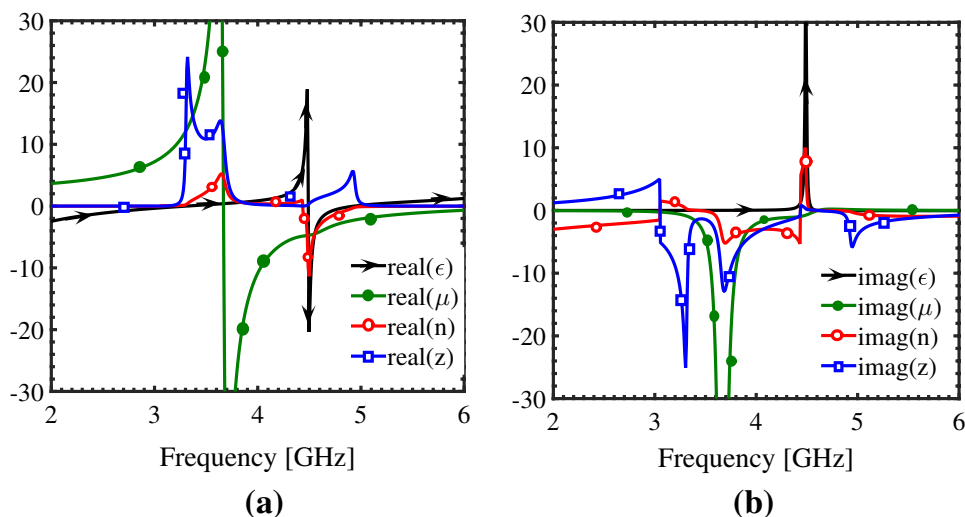


Fig. 2 a Real part of extracted material parameters. b Imaginary part of the extracted material parameters

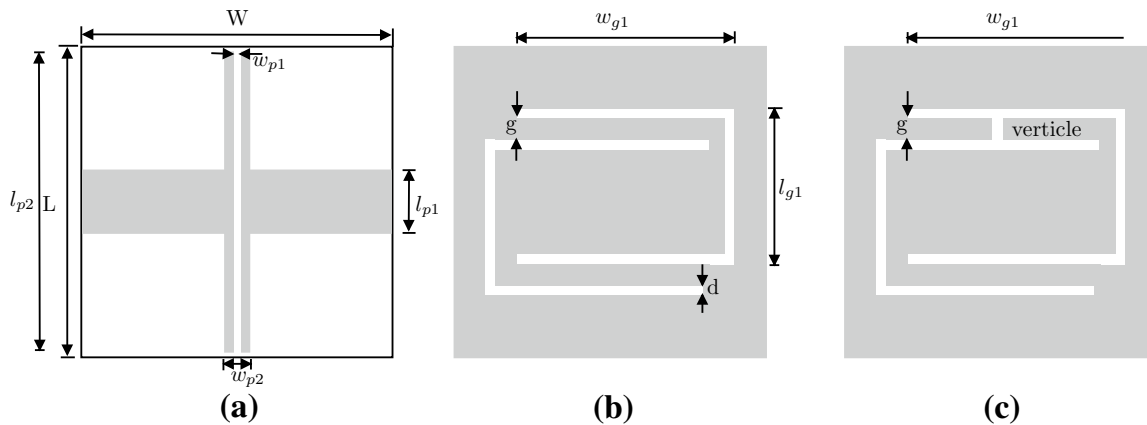


Fig. 3 **a** Top capacitively loaded microstrip line, **b** UCSRR etched bottom ground plane without slot, **c** with vertical slot

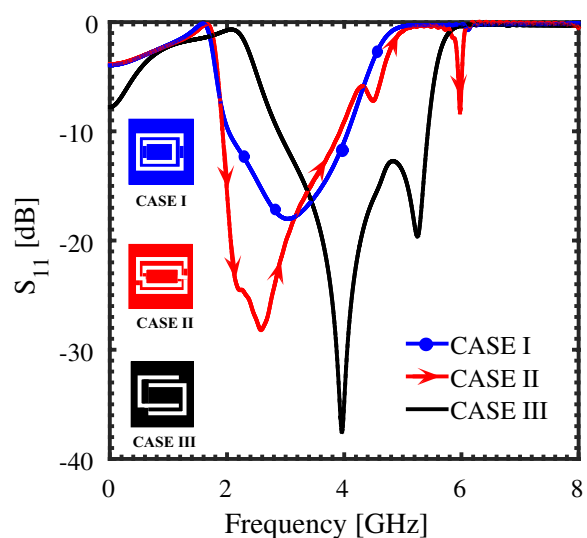
to improve the impedance bandwidth of the filter (discussed in Sect. 3). Table 1 gives the design parameters of proposed bandpass filter shown in Fig. 3.

The U shaped CSRR structure is a modification of conventional CSRR structure. The Fig. 4 shows the evolution steps of the structure. CASE I is the conventional CSRR geometry whose corresponding result in terms of reflection coefficient is indicated by the blue graph which shows $S_{11} < -20$ dB at the resonant frequency with 3-dB bandwidth of 2 GHz. The CASE II is a modification of conventional CSRR geometry (providing 180° symmetry to the structure) as specified in [15] whose result is indicated by red graph which shows $S_{11} < -30$ dB at the resonant frequency with no significant improvement in

Table 1 Design parameters of the proposed BPF

Parameters	Unit (mm)	Parameters	Unit (mm)
L	15	W	15
l_{p1}	3	w_{p1}	0.3
l_{p2}	14	w_{p2}	0.8
l_{g1}	6.5	w_{g1}	10.5
g	0.5	d	0.5

Fig. 4 Stepwise evolution of proposed geometry



bandwidth. Now, the CASE III is a simplified version of CASE II obtained by removing the discontinuities in the left and right arm, it resulted in a significant improvement in the reflection coefficient at resonant frequency as well as 3-dB bandwidth.

3 Results and Discussion

Figure 5 shows the S-parameters of the proposed BPF. It provides a 3 dB passband from 2.44 to 5.58 GHz with a resonant frequency at 4 GHz. Also, two transmission zeros can be observed, one in the lower stopband and other in the upper stopband region, therefore, increasing the selectivity of filter and providing good out-of-band rejection level.

3.1 Extraction of Equivalent Circuit of Proposed Filter

The equivalent circuit extraction is important to study the electrical behaviour of the planar design in order to provide ease of integration with any external electrical circuit. In a split ring resonator (SRR), each ring can be modelled as an inductor and the gap between the rings can be modelled as capacitor. CSRR, being the dual of SRR, shows complementary effect where the inductance is substituted by capacitance of the disk and the gap capacitance is substituted by inductance between the slotted rings [19]. The proposed design consists of U-shaped slots that can be modelled as an equivalent capacitor (C_c) whereas the copper between slots where the current flows can be represented as inductor (L_c). The capacitively loaded transmission line on the top of the substrate can be modelled as a series combination of an inductor (L_l) followed by a gap capacitor (C_g) and an inductor (L_l). The equivalent circuit model of proposed filter is shown in Fig. 6.

Here, the UCSRR equivalent L_c and C_c are divided into two LC tank circuits joined by a weak inductance between the upper and lower halves of the structure (L). The parasitic capacitances C_{lg} and C_{cg} are due to the coupling of line with the ground and capacitive gap with the middle portion of the bottom structure respectively. The parameters are obtained using the procedure described by [15]. The approximate expression for the capacitance (C_c) is give by:

Fig. 5 S-parameters of the proposed BPF without slot

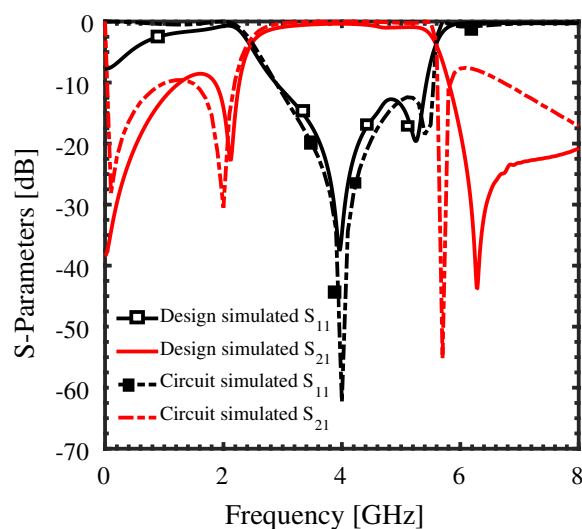
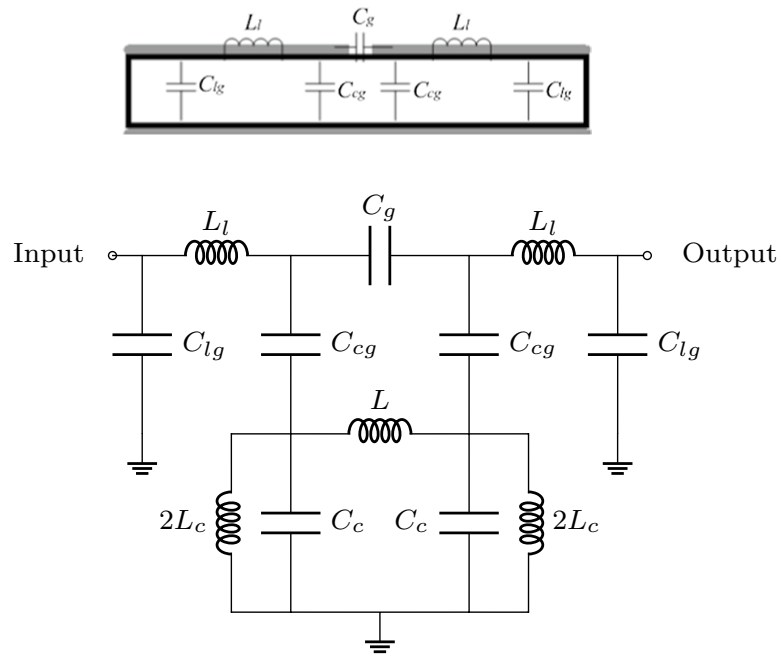


Fig. 6 Equivalent circuit of the proposed filter



$$C_c = (a_{avg} - t/2)C_{pul} \quad (8)$$

where a_{avg} is the average length of the U-shaped structure by considering it as a rectangular ring, t is the gap between the edges of the U-shaped structure and C_{pul} is the distributed capacitance per unit length and is obtained as:

$$C_{pul} = \frac{\sqrt{\epsilon_e}}{c_0 Z_0} \quad (9)$$

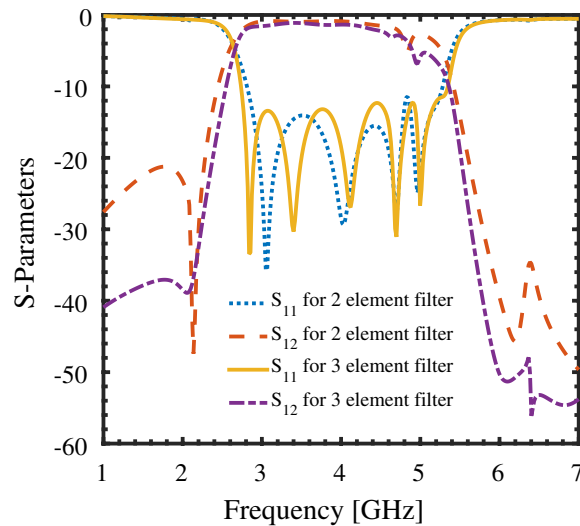
where ϵ_e and c_0 are the effective permittivity of the medium and velocity of light in free space and Z_0 is the characteristic impedance of the CPW line. ϵ_e and Z_0 are obtained using the equations give by [22]. Now, the approximate expression for the inductance (L_c) is give by [22]:

$$L_c = \frac{\mu_0}{2} \frac{b_{avg}}{4} \left[\ln \left(\frac{b_{avg}}{g} \right) - 2 \right] \quad (10)$$

where b_{avg} is the length of the line between the U-shaped slots whose thickness is represented as g in the Fig. 3. The circuit is designed in Advanced Design System (ADS). The parameters are optimised to match the simulation results as shown in Fig. 5. The optimised value of parameters are: $L_l = 2.195$ nH, $C_g = 0.625$ pF, $C_{lg} = 0.33$ pF, $C_{cg} = 1.124$ pF, $L = 1.1$ nH, $L_c = 1.88$ nH, $C_c = 1.39$ pF.

The out-of-band rejection level of the filter can be further improved by using a periodic arrangement of the single element. Figure 7 indicates the S-parameter response obtained after using two and three elements of the filter, showing improved out-of-band rejection level as the number of elements are increased.

Fig. 7 Simulated S-parameters obtained after periodic arrangement of the filter

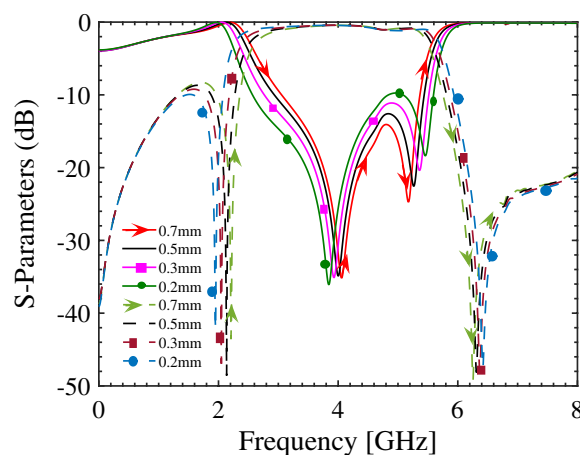


3.2 Bandwidth Enhancement

Now, the separation ‘ g ’ between the two U-shaped slots is varied in order to observe the coupling behaviour between the two adjacent slots. It was observed that on reducing the separation distance ‘ g ’ between the two lateral slots, the passband bandwidth of the proposed filter is increased i.e. the lower cut-off frequency shifts more towards the lower side and the higher cut-off frequency shifts more towards the higher side, as illustrated in Fig. 8. Also, one can say that the separation between the two transmission zeros on the lower and upper stop bands increases with decrease in the distance between the two lateral slots which is due to increase in the inductance between them.

Furthermore, when a vertical slot is introduced between the two lateral slots, the direction of propagation of current changes, as depicted in Fig. 9, specifically at 1.5 GHz (Fig. 9c, d). Initially (Fig. 9c), the current enters the port 1 and is equally distributed on the upper and lower half of the left U-shaped slot and is not allowed to propagate further. Whereas, after the introduction of a vertical slot (Fig. 9d), the current changes its path and concentrates on the lower half of the structure. The increase in concentration of current between the two lateral slots indicate an increase in effective inductance due to which the lower cutoff frequency shifts towards lower frequency side, thus increasing the passband bandwidth of the filter. Figure 10 shows the simulated

Fig. 8 Parametric analysis of the separation between the two U-shaped slots i.e. g



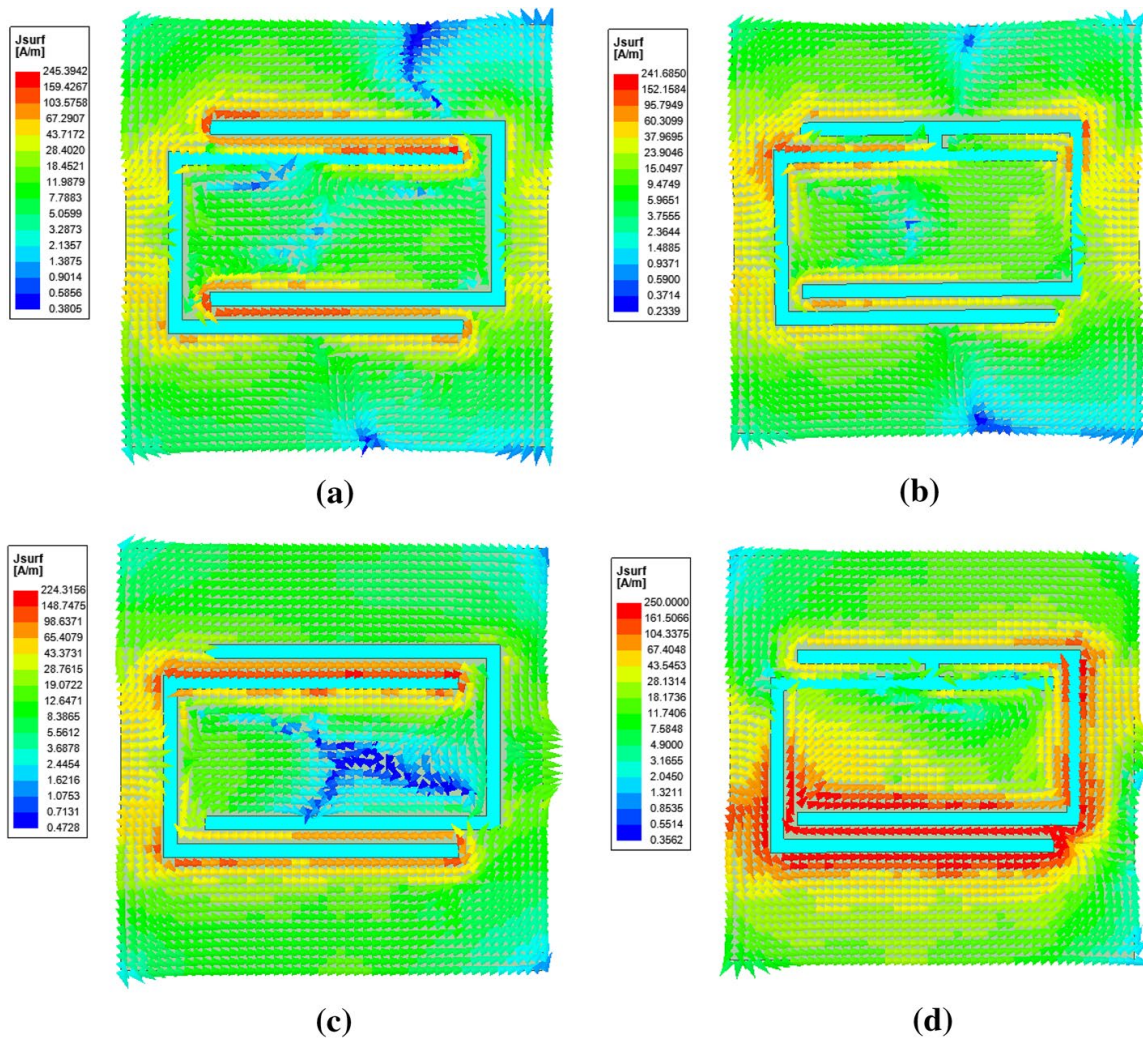
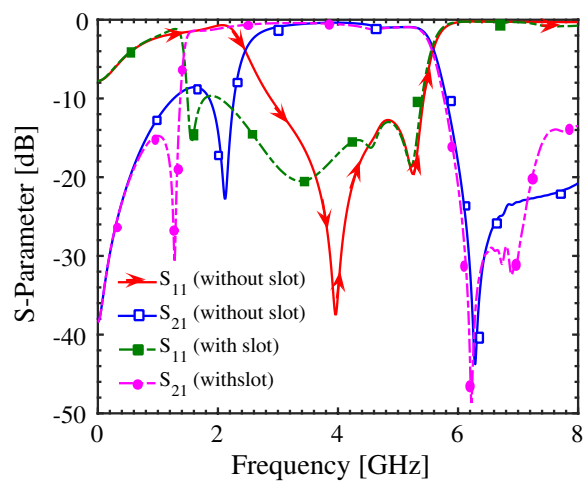


Fig. 9 Surface current density plot at the bottom surface of proposed filter with slot and without slot respectively at 3.75 GHz (a, b) and at 1.5 GHz (c, d)

Fig. 10 S-parameters of proposed design with (shown by solid lines) and without vertical slot (shown by dotted lines)



S-parameters of the filter after cutting the slot. The effective length (L_{eff}) of the path followed by the current at 1.5 GHz (Fig. 9d) can be calculated as,

$$L_{eff} = l_{g1}/2 + 3w_{g1} + l_{g1} + w_{g1}/4 = 43.87\text{mm}. \quad (11)$$

Now, the frequency can be calculated as,

$$f = \frac{c}{2L_{eff}\sqrt{\epsilon_r}} \approx 1.58\text{GHz} \quad (12)$$

The theoretically obtained value is close to the simulated value. Thus, the concept is validated.

So, now the effect of increase in effective inductance can also be incorporated in the equivalent circuit model to demonstrate bandwidth enhancement, as shown in Fig. 11. The value of inductance L_c obtained after tuning is 5.066 nH.

So we conclude that the bandwidth of the proposed wideband filter can be increased further just by connecting the two U-shaped slots with a vertical slot. Using this technique, the 10 dB passband of the filter now includes the 2.4 GHz WLAN band also which was earlier confined to 3.5 GHz WiMAX band and 5.2 GHz WLAN band. Therefore, this results in the miniaturization of the structure by two times with respect to the lowest 10 dB cutoff frequency.

$$\gamma = \frac{1}{l} \cosh^{-1} \left(\frac{(1 + S_{11})(1 - S_{22}) + S_{12}S_{21} + \left(\frac{Z_{01}}{Z_{02}}\right)(1 - S_{11})(1 + S_{22}) + S_{12}S_{21}}{4S_{21}} \right) \quad (13)$$

3.3 Dispersion Characteristics

Figure 12 shows the dispersion diagram of proposed wideband bandpass filter with and without vertical slot. The propagation factor here is $\exp^{\gamma l}$ where $l = 15$ mm is the period of the structure and $\gamma = \alpha + j\beta$ is the complex propagation constant in the direction of propagation given by Eq. (13) [23].

Figure 12 shows that the slope ($d\omega/d\beta$) of dispersion curve is negative in region 1. This means that the wave has anti-parallel phase velocity (v_p) and group velocity (v_g) in this region. Whereas, region 2 shows positive slope of dispersion curve exhibiting parallel

Fig. 11 S-parameters of the proposed BPF with slot

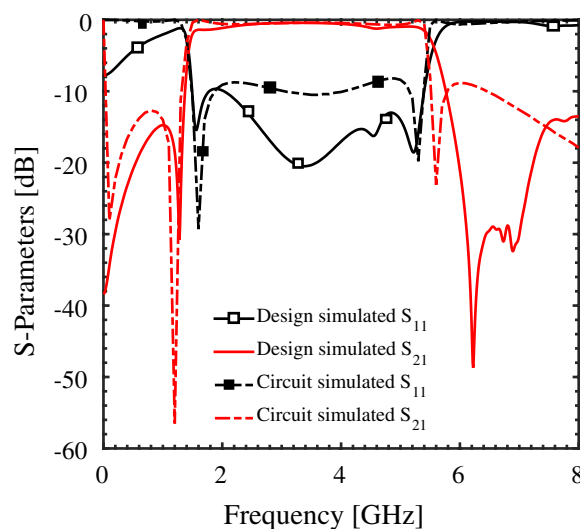
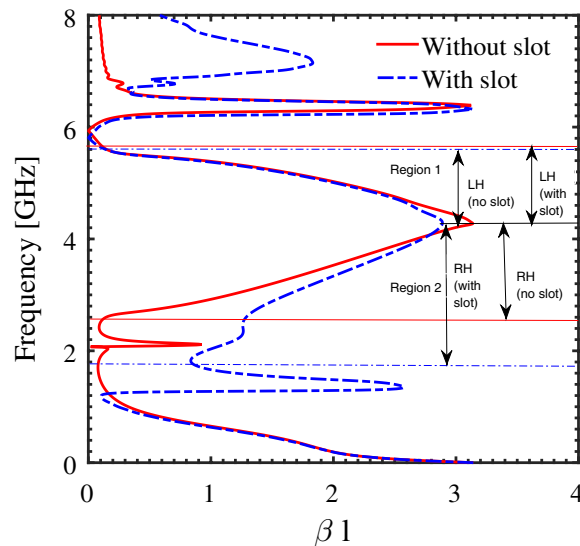


Fig. 12 Dispersion diagram of bandpass filter without slot (solid line) and with slot (dashed line)



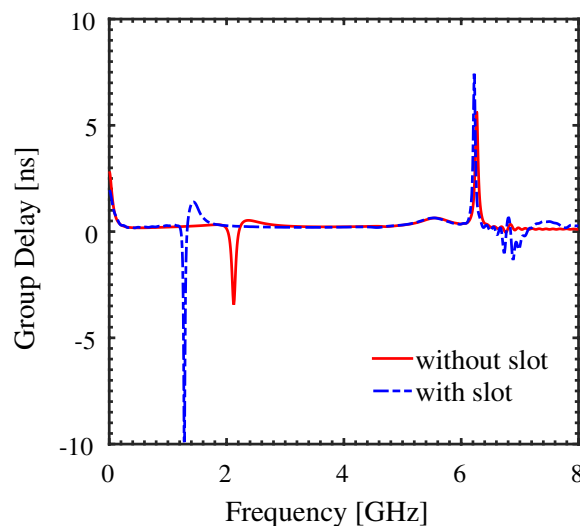
v_p and v_g . Therefore the proposed D-CRLH TL supports backward waves (LH mode) in region 1 (above the resonant frequency) and forward wave in region 2 (below the resonant frequency). A smooth transition is observed between the two regions at the transition frequency approximately 4.2 GHz in both cases, thus, it is said to be a balanced D-CRLH TL.

Further, the filter characteristics are also studied in terms of group delay, as shown in Fig. 13. Proposed bandpass filter exhibits a maximum delay time of 0.63 ns in the pass band.

3.4 Design Synthesis and Parametric Analysis

The parametric analysis is done and design synthesis equations are presented with respect to the most significant dimensions of the filter (shown in Fig. 14) in order to help the designer choose the optimal dimensions for the filter to synthesize a filter with given specifications. On increasing g from 0.2 mm to 0.7 mm and keeping all the other dimensions constant, the transmission zeros shift away from each other increasing the operating bandwidth of the filter, as shown in Fig. 14a. Further, increasing l_{g1} from 5.5 mm to 8 mm again results in similar shift of transmission zeros (shown in Fig. 14c) whereas, on varying w_{g1}

Fig. 13 Group delay of proposed bandpass filter without slot (solid line) and with slot (dashed line)



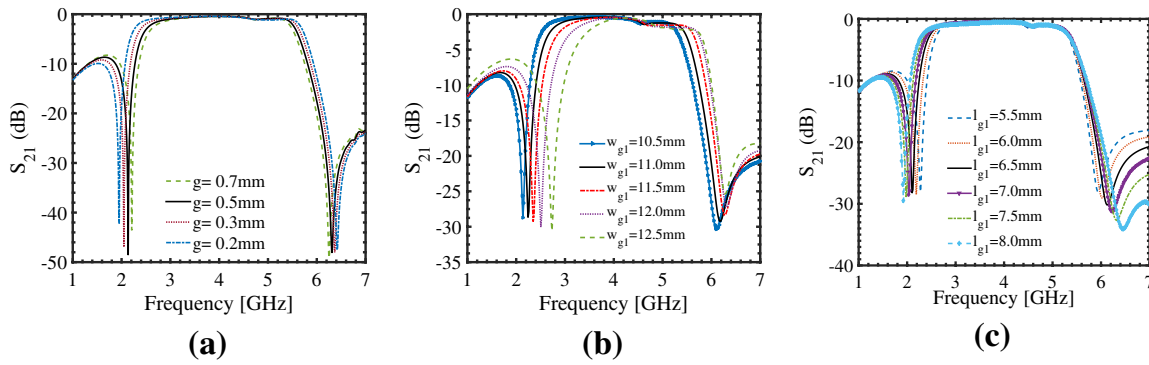


Fig. 14 Variation with respect to **a** g , **b** w_{g1} , **c** l_{g1}

from 10.5 mm to 12.5 mm, both the transmission zeros shift towards the upper frequency side.

Table 2 indicates the synthesis equations of the filter, where f_1, f_2, f_3 and f_4 indicate the frequency at first transmission zero, first 3-dB cut off, second 3-dB cut off and second transmission zero respectively. Using the equations given in Table 2, one can obtain the dimensions of the filter according to given frequency of operation, however, minute tuning is required to be done.

3.5 Performance with Different Substrates

The performance of proposed filter is analysed by taking different substrates of same height. The Fig. 15 below shows the variation of S_{11} in dB with respect to frequency for three different substrates i.e. Roger 4350 with $\epsilon_r = 3.48$ FR-4 with $\epsilon_r = 4.4$ and Duroid 6010 with $\epsilon_r = 10.7$. Since the resonant frequency is inversely proportional to the square root of the dielectric constant so the operating band shifts towards the lower side as the substrate with higher dielectric constant is used and vice versa. However, there is no significant change in the reflection coefficient at resonant frequency when the higher dielectric constant substrate is used. It can also be observed that the reflection coefficient throughout the band degrades in the case of both Roger 4350 and Duroid 6010 as compared to FR-4.

4 Experimental Results and Discussion

Both designs are finally fabricated and results are measured using Keysight N9914A vector network analyzer. Figure 16a, b shows the measured and simulated S-parameters of the proposed filter before and after applying the bandwidth enhancement technique respectively.

Table 2 Synthesis equations based on parametric analysis

Parameters	Values (mm)						Synthesis equations
g	0.2	0.3	0.5	0.7	—	—	$0.55f_1 + 0.96f_2 + 1.985f_3 - 2.75f_4 + 3.236$
w_{g1}	10.5	11	11.5	12	12.5	—	$2.99f_1 - 0.62f_2 + 1.29f_3 + 1.185f_4 - 19.327$
l_{g1}	5.5	6	6.5	7	7.5	8	$-4.97f_1 - 1.86f_2 - 2.97f_3 + 1.011f_4 + 25.386$

Fig. 15 Performance analysis of proposed filter with different substrates of same height

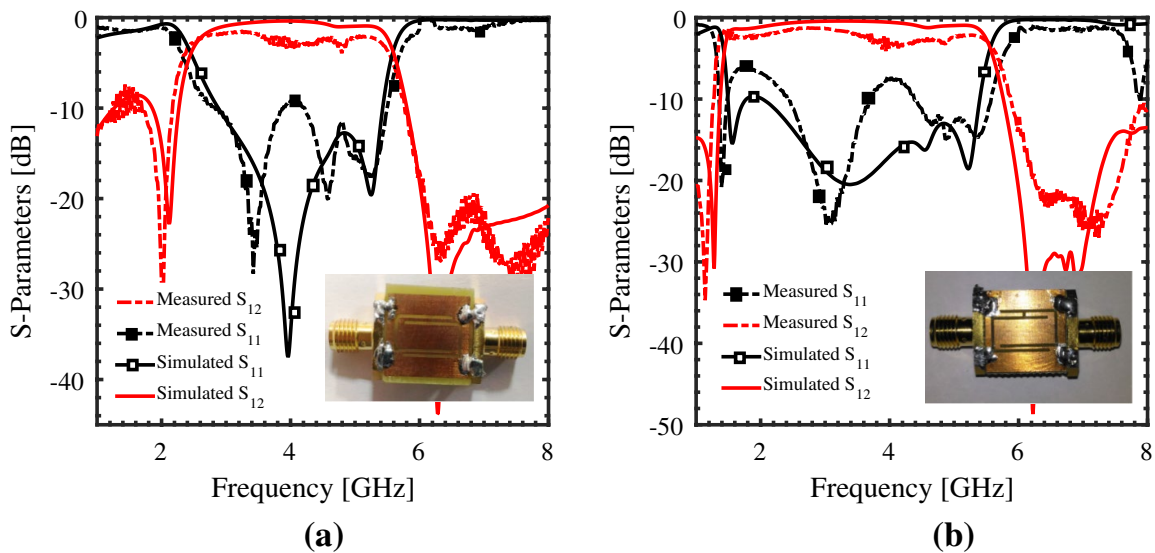
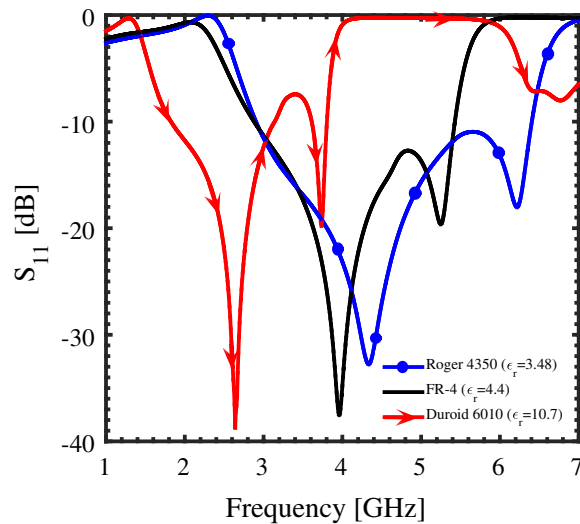


Fig. 16 **a** Measured and simulated S-parameters of proposed filter without slot. **b** Measured and simulated S-parameters of proposed filter after bandwidth enhancement

Simulated and measured results seem to match fairly well. The measured 3-dB passband bandwidth of the proposed filter is 3.29 GHz (82.25 % fractional bandwidth) ranging from 2.33 GHz to 5.62 GHz which is increased to 4.4 GHz (125.7% fractional bandwidth) ranging from 1.3 to 5.7 GHz, after applying the bandwidth enhancement technique i.e. joining the two U-shaped slots by etching out a vertical slot between them.

5 Comparison with Previously Proposed BPF

Use of metamaterials based structures in planar technology provide exceptional properties that are generally beyond the scope of conventional structures. A comparative study is performed which shows better performance of the proposed metamaterial based bandstop filter as compared to previously proposed metamaterial based BPF as well as conventional bandstop filters available in literature, given in Table 3.

Table 3 Comparative analysis of proposed BPF with previously proposed ones

References	Size (w.r.t. λ_g)	3-dB bandwidth (GHz)	Fractional Bandwidth (%)	Insertion loss (dB)	Center frequency (GHz)
[1]	0.48×0.24	1.33	60	0.6	2.05
[4]	0.3×0.1	2.5	55	1.2	4.2
[10]	0.12×0.22	> 2.64	80.48	0.35	3.25
[24]	0.26×0.30	—	3.5	2.3	2.45
[25]	—	5.9	86.76	1.9	6.8
Prop	0.15×0.15	4.4	125.7	0.45	3.5

The comparison table shows that the proposed design provides wider band and compact sized filter as compared to the previously proposed ones.

6 Conclusions

A dual composite right/left handed transmission line based bandpass filter with high stop-band rejection level has been designed and simulated. Two closely coupled U-shaped resonators are used to fulfil the purpose. A slot is further introduced between the two U-shaped slots to increase the bandwidth on account of change in the direction of flow of current and increase in effective inductance of the design. The filter provided a 3 dB pass band from 2.44 to 5.58 GHz which further increased from 1.43 to 5.56 GHz when the vertical slot was placed resulting in miniaturization of the structure by 2 times. The propagation characteristics of the filter have been studied that demonstrated the dual balanced CRLH line behaviour of the filter. Very small amount of group time delay (0.63 ns) is provided by the filter in the passband region. Also, a high return loss of 37.5 dB is achieved at the pass band resonant frequency. Finally the designs are fabricated and measured to show close proximity to the simulated results.

Funding None.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Xu, K. D., Li, D., & Liu, Y. (2019). High-selectivity wideband bandpass filter using simple coupled lines with multiple transmission poles and zeros. *IEEE Microwave and Wireless Components Letters*, 29(2), 107–109. <https://doi.org/10.1109/LMWC.2019.2891203>.
2. Garg, P., Awasthi, S., & Jain, P. (2018). A survey of microwave bandpass filter using coupled line resonator—Research design and development. In *2018 International conference on sustainable energy, electronics, and computing systems (SEEMS)* (pp. 1–9). Greater Noida. <https://doi.org/10.1109/SEEMS.2018.8687341>.

3. Feng, W., Gao, X., Che, W., & Xue, Q. (2015). Bandpass filter loaded with open stubs using dual-mode ring resonator. *IEEE Microwave and Wireless Component Letters*, 25(5), 295–297. <https://doi.org/10.1109/LMWC.2015.2410174>.
4. Liu, L., Zhang, P., Weng, M. H., Tsai, C. Y., & Yang, R. Y. (2019). A miniaturized wideband bandpass filter using quarter-wavelength stepped-impedance resonators. *Electronics*, 8(12), 1540. <https://doi.org/10.3390/electronics8121540>.
5. Sassi, I., Talbi, L., & Hettak, K. (2016). Compact bandpass filters based on linked hexagonal-omega resonators. *Microwave and Optical Technology Letters*, 58(5), 1049–1052. <https://doi.org/10.1002/mop.29720>.
6. Nasraoui, H., Mouhsen, A., El Aoufi, J., & Taouzari, M. (1999). Novel microstrip low pass filter based on complementary split-ring resonators. *International Journal of Modern Communication Technologies and Research*, 2(10), 265760.
7. Bonache, J., Posada, G., Carchon, G., De Raedt, W., & Martn, F. (2007). Compact ($< 0.5 \text{ mm}^2$) K-band metamaterial bandpass filter in MCM-D technology. *Electronics Letters*, 43(5), 288–290. <https://doi.org/10.1049/el:20073891>.
8. Caloz, C., & Itoh, T. (2005). *Electromagnetic metamaterials: Transmission line theory and microwave applications*. John Wiley & Sons.
9. Yang, S., Chen, Y., Yu, C., Lu, G., Li, B., Wang, L., et al. (2019). Super compact and ultra-wide-band bandpass filter with a wide upper stopband based on a SCRLH transmission-line unit-cell and two lumped capacitors. *Journal of Electromagnetic Waves and Applications*, 33(3), 350–366. <https://doi.org/10.1080/09205071.2018.1552538>.
10. Choudhary, D. K., & Chaudhary, R. K. (2018). A compact via-less metamaterial wideband bandpass filter using split circular rings and rectangular stub. *Progress In Electromagnetics Research*, 72, 99–106. <https://doi.org/10.2528/pier117092503>.
11. Caloz, C. (2006). Dual composite right/left-handed (D-CRLH) transmission line metamaterial. *IEEE Microwave and Wireless Components Letters*, 16(11), 585–587. <https://doi.org/10.1109/LMWC.2006.884773>.
12. Gonzalez-Posadas, V., Jimnez-Martn, J. L., Parra-Cerrada, A., Garca-Munoz, L. E., & Segovia-Vargas, D. (2010). Dual-composite right left-handed transmission lines for the design of compact diplexers. *IET Microwaves, Antennas & Propagation*, 4(8), 982–990. <https://doi.org/10.1049/iet-map.2009.0571>.
13. Kholodnyak, D., Turgaliev, V., & Zameshaeva, E. (2015, March). Dual-band immittance inverters on dual-composite right/left-handed transmission line (D-CRLH TL). In *2015 German microwave conference* (pp. 60–63). IEEE. <https://doi.org/10.1109/GEMIC.2015.7107752>.
14. Wu, G. C., Wang, G., & Wang, Y. W. (2013). Novel simplified dual-composite right/left-handed transmission line and its application in bandpass filter with dual notch bands. *Progress in Electromagnetics Research*, 44, 123–131. <https://doi.org/10.2528/pierc13082602>.
15. Belenguer, A., Cascon, J., Borja, A. L., Esteban, H., & Boria, V. E. (2012). Dual composite right/left-handed coplanar waveguide transmission line using inductively connected split-ring resonators. *IEEE Transactions on Microwave Theory and Techniques*, 60(10), 3035–3042. <https://doi.org/10.1109/TMTT.2012.2210438>.
16. Cano, L. M., Borja, A. L., Boria, V. E., & Belenguer, A. (2016). Highly versatile coplanar waveguide line with electronically reconfigurable bandwidth and propagation characteristics. *IEEE Transactions on Microwave Theory and Techniques*, 65(1), 128–135. <https://doi.org/10.1109/TMTT.2016.2613526>.
17. Computer Simulation Technology Microwave Studio (CST MWS). Available at <http://www.cst.com> (online).
18. Advanced Design System (ADS), Keysight EEs of EDA. (2011). Available at <http://www.keysight.com> (online).
19. Baena, J. D., Bonache, J., Martn, F., Sillero, R. M., Falcone, F., Lopetegi, T., et al. (2005). Equivalent-circuit models for split-ring resonators and complementary split-ring resonators coupled to planar transmission lines. *IEEE Transactions on Microwave Theory and Techniques*, 53(4), 1451–1461. <https://doi.org/10.1109/TMTT.2005.845211>.
20. Smith, D. R., Vier, D. C., Koschny, T., & Soukoulis, C. M. (2005). Electromagnetic parameter retrieval from inhomogeneous metamaterials. *Physical Review E*, 71(3), 036617. <https://doi.org/10.1103/PhysRevE.71.036617>.
21. Falcone, F., Lopetegi, T., Laso, M. A. G., Baena, J. D., Bonache, J., Beruete, M., et al. (2004). Babinet principle applied to the design of metasurfaces and metamaterials. *Physical Review Letters*, 93(19), 197401. <https://doi.org/10.1103/PhysRevLett.93.197401>.
22. Bahl, I. J., & Bhartia, P. (2003). *Microwave solid state circuit design*. John Wiley & Sons.

23. Ying, X., & Alphones, A. (2005). Propagation characteristics of complimentary split ring resonator (CSRR) based EBG structure. *Microwave and Optical Technology Letters*, 47(5), 409–412. <https://doi.org/10.1002/mop.21185>.
24. Tang, M. C., Shi, T., & Tan, X. (2016). A novel triple-mode hexagon bandpass filter with meander line and central-loaded stub. *Microwave and Optical Technology Letters*, 58(1), 9–12. <https://doi.org/10.1002/mop.29483>.
25. Becharef, K., Nouri, K., Kandouci, H., Bouazza, B. S., Damou, M., & Bouazza, T. H. C. (2020). Design and simulation of a broadband bandpass filter based on complementary split ring resonator circular “CSRRs”. *Wireless Personal Communications*, 111(3), 1341–1354. <https://doi.org/10.1007/s11277-019-06918-6>.

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



Priyanka Garg was born in Dehradun, India. She received her B.Tech degree (Electronics & communication) in 2013 from Uttarakhand Technical University, Uttarakhand, India, and M.Tech. (Digital Signal Processing) in 2016 from G.B. Pant Engineering College, Pauri, Uttarakhand, India. She is currently pursuing Ph.D from Delhi Technological University, Delhi. Her research is in the area of metamaterial based microwave components.



Priyanka Jain has received her Ph.D in Signal Processing and is currently working as an Assistant Professor, in Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India since 2011. She received her B.E. degree (Electronics & Telecommunication) from Amravati University, Maharashtra, India, in 1998 and M.Tech. (Microwave Engg.) from Delhi University, India. From 2001 to 2002, she worked as Lecturer at Guru Prem Sukh Memorial College of Engineering (GGSIP University), Delhi, India. From 2002 to 2011 she worked as Lecturer at India Gandhi Institute of Technology, New Delhi. Her teaching and research are in signal processing, analog electronics and microwave engineering. She has published many articles in international and national journals in fields of signal processing, microwave and communication.



Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Design and evaluation of stand-alone solar-hydrogen energy storage system for academic institute: A case study

Alfred John^a, Srijit Basu^a, Akshay^a, Anil Kumar^{a,b,*}^a Department of Mechanical Engineering, Delhi Technological University, Delhi 110042 India^b Centre for Energy and Environment, Delhi Technological University, Delhi 110042 India

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Solar energy
Electrification
Energy economics
Simulation
HOMER

ABSTRACT

Energy demand is increasing with population and technical advancements. Hence the need for solar energy for electrification has increased tremendously due to the abundance of sunlight. Solar Energy is intermittent and can be associated with a hydrogen energy to stabilize the grid. This paper focuses on the design of electrification by solar energy using hydrogen energy storage for the science block of Delhi Technological University, Delhi (India). It provides an economic analysis using HOMER software. Levelized cost has been \$0.6050 kW/hr with a total annual production of 436.020 MW/yr. Therefore, solar energy with other renewable energy source for electrification in universities can be economical and sustainable.

© 2021 Elsevier Ltd. All rights reserved.

Selection and peer-review under responsibility of the scientific committee of the Technology Innovation in Mechanical Engineering-2021.

1. Introduction

The demand for energy in the world is increasing enormously due to rapid urbanization. In one century alone, the population has grown by 2 billion and that developed countries have made significant contributions. Preventing an energy shortage is one of the 21st century's most trivial issues. Therefore, energy market is expanding rapidly in line with the needs of the increasing world population. The various countries in the world are formed up in the world by their own tactics, schemes, legislation and close monitoring. World's available resources are diminishing with population growth and development strategies [1]. Therefore, understanding energy sources is necessary, as they play a major role in addressing the world's and people's needs. For various factors, accessible energy is not available to citizens, such as the nation's development profile, people's fiscal situation, and technological advancements. Ecosystem is highly contaminated by the pollution of multiple fuel based combustion gases that are easily accessible and used worldwide to fulfill energy demand. Developing nations are also being pressured to look for energy options, as their population has increased significantly and economic prosperity is being sought [2]. World Bank and the International Energy

Agency (IEA) predicted that in the next 40 years, the world would need to double installed energy capacity to satisfy the expected demands of developed countries [3]. IEA claimed that grid energy was blocked by 1.3 billion people in developed countries who live far from towns [4]. Though grid propagation or construction are the first electrification solutions, the massive investments involved in these areas make this approach unaffordable [5]. Eighty percent of people living in rural areas of developed countries historically use wood to fulfill their energy needs, which has since made deforestation one of the worst environmental problems in the world [6]. World's leading sources of energy today include finite energy sources such as natural gas, coal and unprocessed crude. Due to an exponential increase in population and energy consumption, the world cannot depend solely on small traditional supplies to satisfy the demand [7]. Inexhaustible energy supplies, also known as green energy, are available in large numbers at no cost [8]. Due to the decline in the supply of fossil fuels and coal, volatile oil price, the rising demand for power, and the risks of global warming, alternate energies have received significant attention as the source for electricity generation in recent decades. While development of green energy seems to be a positive course ahead, there are also significant drawbacks, including sources which are available over time, are not sufficiently powerful in all areas and which entail a high cost of capital. Renewable energy systems have a significantly greater upfront cost than fossil fuel systems. This is why fossil fuels

* Corresponding author.

E-mail address: anilkumar76@dtu.ac.in (A. Kumar).<https://doi.org/10.1016/j.matpr.2021.04.461>

2214-7853/© 2021 Elsevier Ltd. All rights reserved.

Selection and peer-review under responsibility of the scientific committee of the Technology Innovation in Mechanical Engineering-2021.

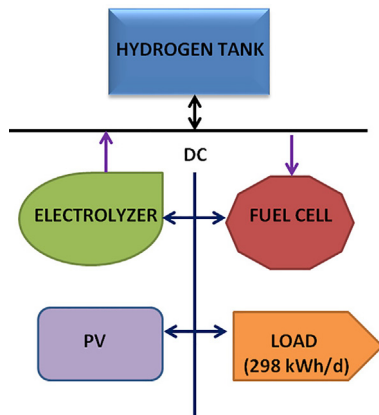


Fig. 1. System Layout

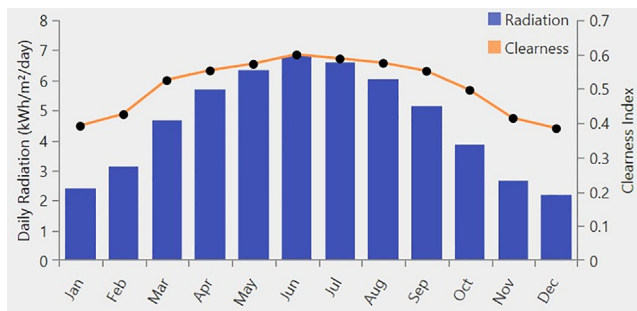


Fig. 2. Solar Radiation

now meet 80 percent of global energy demand, leading to a substantial environmental impact.

In comparison with other sources of green energy, there has been significant development in solar energy towards practicability since the oil crises of the 1970 s. Sun is a direct or indirect source of limitless energies and is known for using solar radiation from the Sun [9]. Solar energy has a leading role in reducing hazardous environmental gases from power generation in response to environmental emissions questions. IEA estimated that 100Gt (Gigaton) of CO₂ emissions from Solar PV technologies could be avoided between 2008 and 2050 [10]. The generation of solar energy has little impact on cultivated soil, decreases the costs of the distribution of grid transmission lines and enhances life quality in remote regions [11]. Many solar power sources worldwide have been deployed as standalone or hybrid systems for electrifying

Table 1
Solar Radiation and Clearness Index

Month	Clearness Index	Daily Radiation (kWh/m ² /day)
January	0.391	2.392
February	0.425	3.111
March	0.524	4.655
April	0.552	5.672
May	0.571	6.339
June	0.598	6.809
July	0.587	6.575
August	0.574	6.050
September	0.551	5.141
October	0.495	3.848
November	0.413	2.637
December	0.384	2.200

remote areas in the form of an annual monthly solar radiation spectrum of 376 kWh/m² [12].

Contrary to traditional energy sources, renewable sources cannot provide consistent energy to satisfy energy requirements. Such sources differ abundantly over the season (e.g., solar and wind) (e.g., hydroelectric). However, using solar energy in a hybrid device will remove green power sources from their downside [6]. Hybrid technology is the best way to generate power in rural areas to reduce the rising cost of fuel and the cost of grid propagation. It is also the cheapest replacement for a generator [13]. Because of its global availability, most hybrid systems locations have prioritized solar energy. In cases where solar radiation is insufficient, hybridization would resolve the possible issue of solar-PV reliability. Nature of hybrid solar power systems depends on environmental conditions and the available energy sources at the site and consider the most cost-effective and stable source mixture to reduce unsustainable investment and satisfy demand [5]. It is desirable to use the more reliable hydrogen storage method, more effective and more dependable in long-term energy storage [14]. In the second half of the 21st century, water hydrogen output will eventually overtake fossil fuels and become the main power carrier [15]. Water is commonly regarded by the use of term green energy as a natural and safe source for hydrogen production [16]. Energy conservation is a good way to solve daylight and storage challenges by balancing water demand with electricity supply. Electrochemical energy storage technologies have gained more attention because of flexibility, reliability, grid efficiency and high quality between electricity storage technology [17]. Hydrogen generation is now one of the most common choices for storing chemical energy for its high value, energy density, and negligible or close to zero emissions based on water electrolysis [16].

The current study focuses on designing and evaluating the hybrid solar- hydrogen energy system for the science block of Delhi Technological University, Delhi (India). The block has an elec-

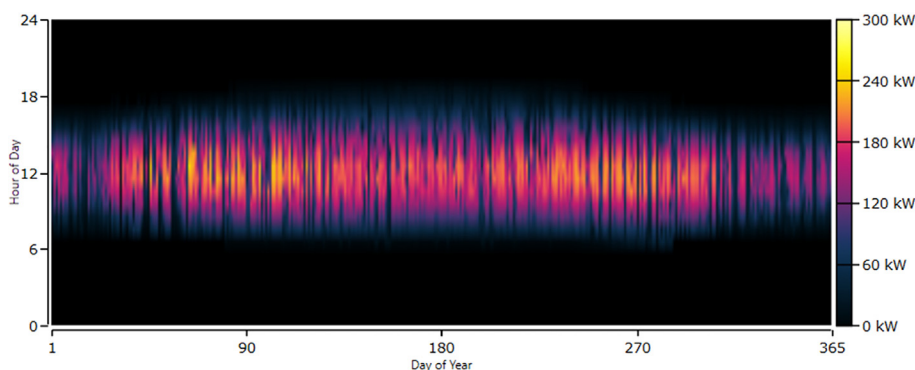


Fig. 3. PV Power Output

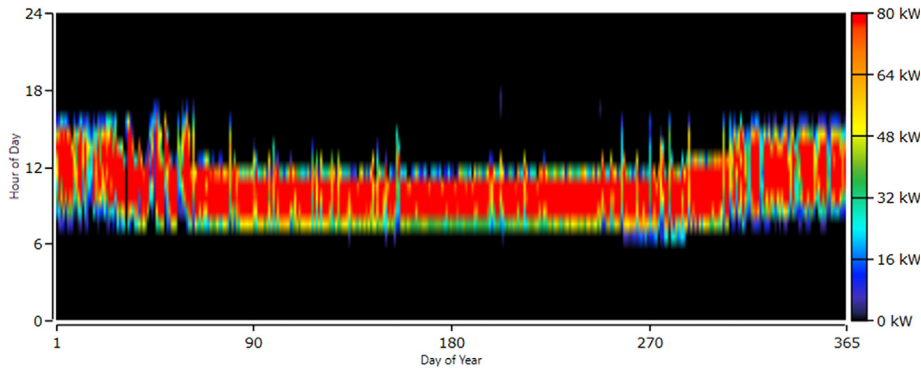


Fig. 4. Electrolyzer Input Power

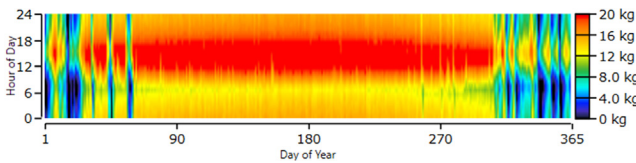


Fig. 5. Hydrogen Tank level (Hourly)

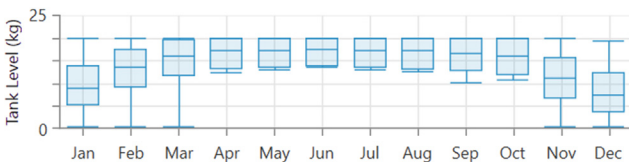


Fig. 6. Hydrogen Tank Level (Monthly)

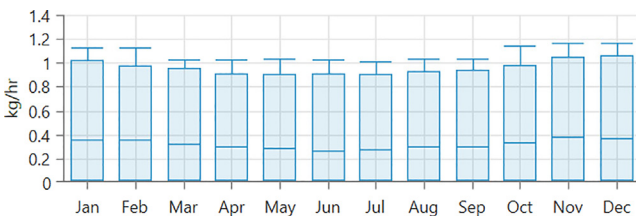


Fig. 7. Monthly Fuel Consumption

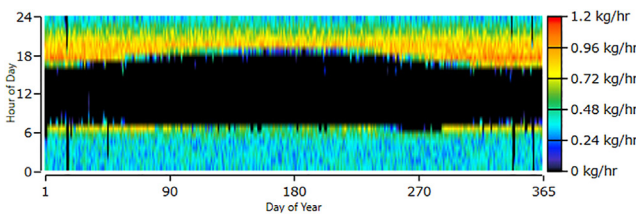


Fig. 8. Hourly Fuel Consumption

tric load demand of 108.129 MWh/yr. Simulation in HOMER software optimized a PV system with a capacity of 240 kW and integrated electrolyzer and fuel cell for hydrogen production and grid stability.

2. System description

Fig. 1 shows the schematic diagram of the system proposed for the science block of Delhi Technological University. Hybrid energy

system consists of a PV cell with a capacity of 240 kW to produce electricity to meet the demand. Electrolyzer uses the excess electricity produced by PV to make hydrogen by splitting water into hydrogen and oxygen. Fuel cell produces electricity using stored hydrogen when the electric load demand of the science block is not met by PV cell. This ensures the complete stability of the electricity requirements of the block.

2.1. Solar PV radiation

Solar radiation is reliability in the context of a highly dispersed energy supply of energy [18]. Their regular cycle can vary widely and be strongly affected by weather, storm, hazy weather, and fog. Input data such as solar radiation, local weather conditions and other techniques of the proposed PV systems are essential for the modelling of an output simulation [19]. HOMER software obtains solar data from NASA Prediction of Worldwide Energy Resources. Solar radiation for Delhi Technological University is given in Fig. 2. The average daily radiation and clearness index for every month is shown in Table 1.

3. Result and discussion

HOMER software has simulated the most optimal design with a net present cost of \$1,030,406. Solar photovoltaic produces energy of 389.865 MWh/yr with a rated capacity of 240 kW and mean output of 44.5 kW, which is also mean output of 1,068 kWh/day. The maximum output of PV system is 253 kW with a PV penetration of 358% when operated for 4,344 hrs/yr shown in Fig. 3.

Excess energy produced by PV is used to obtain hydrogen through electrolysis. An electrolyzer with a rated capacity of 80 kW, mean input of 14.7 kW and maximum input of 80 kW is used to produce hydrogen. Electrolyzer has a capacity factor of 18.4% with an operation time of 2328 hr/yr and thus having total input energy of 128,687 kWh/yr as shown in Fig. 4. Thus, electrolyzer produces hydrogen with a mean output of 0.317 kg/hr and a maximum output of 1.72 kg/hr. The total production is 2,773 kg/yr with a specific consumption of 46.4 kWh/kg.

A hydrogen tank with the autonomy of 53.7 hr and a capacity of 20 kg is used as described in Fig. 5. Simulation has found the content at the beginning of year to be 2 kg and at the end of year to be 5.83 kg with a storage capacity of 667 kWh. A total of 2,769 kg of hydrogen is consumed after production through electrolysis. The average consumption of fuel per day is 7.59 kg and the average consumption of fuel per hour is 0.316 kg as shown in Fig. 6.

Fuel cell is operated for 5,207 hrs/yr to produce electricity of 46.155 MWh/yr with a capacity factor of 26.3%. Mean electrical output has been found as 8.86 kW and maximum electrical output to be 19.4 kW. Specific fuel consumption of fuel cell is 0.06 kg/kWh

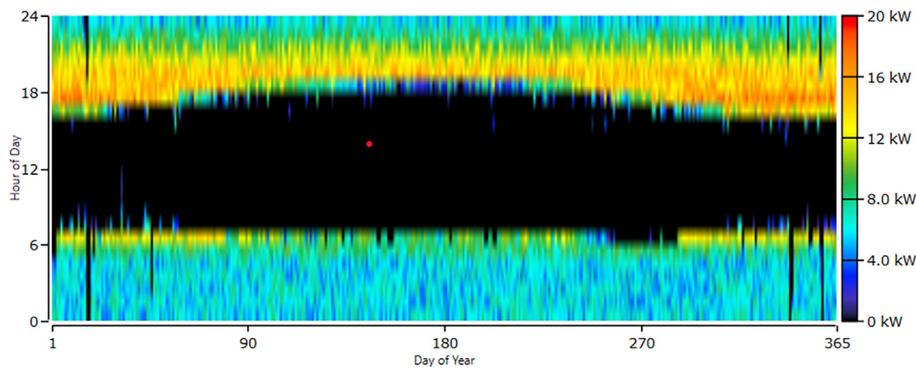


Fig. 9. Generator Power Output

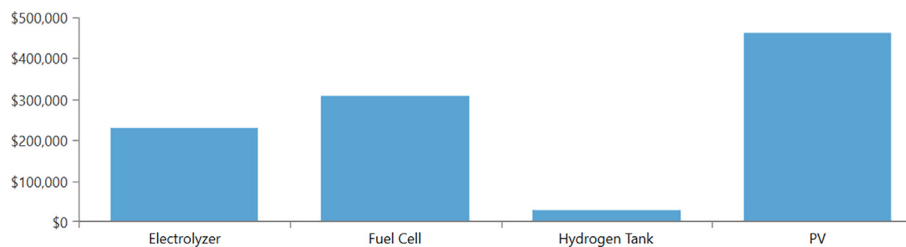


Fig. 10. Cost Summary

with a fuel energy input of 92,310 kWh/yr. Hence, contributing to a mean electrical efficiency of 50 percent. Fig. 7 shows the monthly fuel consumption, while Fig. 8 shows hourly fuel consumption.

Thus, proposed system produces electricity of 436.020 MWh/yr in which PV produces 389.865 MWh/yr that contributes to 89.4% of total production and fuel cell produces electricity of 46.155 MWh/yr, which contributes to 10.6% of total production, as shown in Fig. 9. The produced energy is used to satisfy the DC primary load of 108,129 kWh/yr for the science block of DTU, Delhi (India).

The total net present cost of the system is \$ 1,030,406 and the operating cost of the system is \$ 26,688.37. The cost distribution graph of various components of the system is given in Fig. 10.

4. Conclusion

Current study looks at the scientific and economic viability of using solar photovoltaic energy sources to feed the Science Block of Delhi Technological University, Delhi (India). Hybrid energy system uses hydrogen energy storage to stabilize the intermittency of solar energy to provide a stable electrical current. Total electricity produced by PV system is 389.865 MWh/yr, with an excess electricity production of 199.204 MWh/yr. Electrolyzer uses the excess electricity to produce hydrogen and produce electricity when needed via fuel cell. The designed energy system uses a PV cell with a rated capacity of 240 kW and a hydrogen tank of 20 kg. Total capital of the proposed system is \$ 610,000 with a replacement cost of \$ 359,252.87. Operation and maintenance cost of the system is \$ 201,851.21. Thus, all cost leads to a total net present cost of \$ 1,030,406 with a lifetime of 25 years. Levelized Cost of energy is obtained as 0.6050 for the PV-hydrogen system. Although cost per kW for the proposed system is much higher than the cost per kW of electricity obtained from the grid. It is justified with the motive of using a cleaner form of energy with no harmful emission. Integrating hydrogen energy storage with photovoltaic aids in standalone use of clean energy and support energy transitions. Proposed system's benefits include climate change reduction,

increased energy sector reliability, and increased power source stability.

CRediT authorship contribution statement

Alfred John: Data curation, Conceptualization, Writing - original draft. **Srijit Basu:** Visualization, Methodology. **Akshay:** Investigation. **Anil Kumar:** Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Acknowledgment

The authors are highly thankful to the Centre for Energy and Environment, Delhi Technological University, for providing basic infrastructure for compiling this work.

References

- [1] S. Shafiee, E. Topal, When will fossil fuel reserves be diminished?, *Energy Policy*. 37 (1) (2009) 181–189, <https://doi.org/10.1016/j.enpol.2008.08.016>.
- [2] J. Asafu-Adjaye, The relationship between energy consumption, energy prices and economic growth: Time series evidence from Asian developing countries, *Energy Econ.* 22 (6) (2000) 615–625, [https://doi.org/10.1016/S0140-9883\(00\)00050-5](https://doi.org/10.1016/S0140-9883(00)00050-5).
- [3] B.E. Türkay, A.Y. Telli, Economic analysis of standalone and grid connected hybrid energy systems, *Renew. Energy*. 36 (7) (2011) 1931–1943, <https://doi.org/10.1016/j.renene.2010.12.007>.
- [4] IEA, *World Energy Outlook 2012 en francais*, (2013) 12.
- [5] P. Díaz, C.A. Arias, R. Peña, D. Sandoval, FAR from the grid: A rural electrification field study, *Renew. Energy*. 35 (12) (2010) 2829–2834, <https://doi.org/10.1016/j.renene.2010.05.005>.
- [6] N.M. Ijumba, C.W. Wekesah, P.O. Box, *Sources I N Rural Electrification* (1996) 720–723.
- [7] O. Erdinc, M. Uzunoglu, Optimum design of hybrid renewable energy systems: Overview of different approaches, *Renew. Sustain. Energy Rev.* 16 (3) (2012) 1412–1425, <https://doi.org/10.1016/j.rser.2011.11.011>.

- [8] S.G.J. Ehnberg, M.H.J. Bollen, Reliability of a small power system using solar power and hydro, *Electr. Power Syst. Res.* 74 (1) (2005) 119–127, <https://doi.org/10.1016/j.epsr.2004.09.009>.
- [9] G.R. Timilsina, L. Kurdgelashvili, P.A. Narbel, Solar energy: Markets, economics and policies, *Renew. Sustain. Energy Rev.* 16 (1) (2012) 449–465, <https://doi.org/10.1016/j.rser.2011.08.009>.
- [10] X. Zhang, X. Zhao, S. Smith, J. Xu, X. Yu, Review of R&D progress and practical application of the solar photovoltaic/thermal (PV/T) technologies, *Renew. Sustain. Energy Rev.* 16 (2012) 599–617, <https://doi.org/10.1016/j.rser.2011.08.026>.
- [11] A. Bahadori, C. Nwaoha, A review on solar energy utilisation in Australia, *Renew. Sustain. Energy Rev.* 18 (2013) 1–5, <https://doi.org/10.1016/j.rser.2012.10.003>.
- [12] M.A. Elhadidy, Performance evaluation of hybrid (wind/solar/diesel) power systems, *Renew. Energy.* 26 (3) (2002) 401–413, [https://doi.org/10.1016/S0960-1481\(01\)00139-2](https://doi.org/10.1016/S0960-1481(01)00139-2).
- [13] P. Nema, R.K. Nema, S. Rangnekar, A current and future state of art development of hybrid energy system using wind and PV-solar: A review, *Renew. Sustain. Energy Rev.* 13 (8) (2009) 2096–2103, <https://doi.org/10.1016/j.rser.2008.10.006>.
- [14] O.V. Marchenko, S.V. Solomin, The future energy: Hydrogen versus electricity, *Int. J. Hydrogen Energy.* 40 (10) (2015) 3801–3805, <https://doi.org/10.1016/j.ijhydene.2015.01.132>.
- [15] J.O'M. Bockris, The hydrogen economy: Its history, *Int. J. Hydrogen Energy.* 38 (6) (2013) 2579–2588, <https://doi.org/10.1016/j.ijhydene.2012.12.026>.
- [16] Y. Li, D.W. Chen, M. Liu, R.Z. Wang, Life cycle cost and sensitivity analysis of a hydrogen system using low-price electricity in China, *Int. J. Hydrogen Energy.* 42 (4) (2017) 1899–1911, <https://doi.org/10.1016/j.ijhydene.2016.12.149>.
- [17] T.M.I. Mahlia, T.J. Saktisahdan, A. Jannifar, M.H. Hasan, H.S.C. Matseelar, A review of available methods and development on energy storage; Technology update, *Renew. Sustain. Energy Rev.* 33 (2014) 532–545, <https://doi.org/10.1016/j.rser.2014.01.068>.
- [18] M.d. Alam Hossain Mondal, A.K.M. Sadrul Islam, Potential and viability of grid-connected solar PV system in Bangladesh, *Renew. Energy.* 36 (6) (2011) 1869–1874, <https://doi.org/10.1016/j.renene.2010.11.033>.
- [19] A.K. Shukla, K. Sudhakar, P. Baredar, Design, simulation and economic analysis of standalone roof top solar PV system in India, *Sol. Energy.* 136 (2016) 437–449, <https://doi.org/10.1016/j.solener.2016.07.009>.

Design of Compact Circular Microstrip Patch Antenna using Parasitic Patch

Richa Sharma
ECE Department
Delhi Technological University
Delhi, India
sharmaricha@akgec.ac.in

N.S.Raghava
ECE Department
Delhi Technological University
Delhi, India
nsraghava@gmail.com

Asok De
ECE Department
Delhi Technological University
Delhi, India
asok.de@gmail.com

Abstract—In the present study circular microstrip patch antenna is designed for the Ultra High frequency (UHF) band of 470MHz-806MHz. Different studies have been published in this area for the amelioration of the patch antenna. We can choose any antenna application depending upon our requirements. There is a big disadvantage of microstrip antenna i.e. narrow bandwidth, low efficiency and low gain. To enhance antenna gain multilayer are used in the proposed design. In this study circular microstrip patch antenna is designed and the resonant frequency of the proposed antenna is determined for different parasitic patch locations, Feeding Point, Shorting Pin, and Substrate thickness. Main Patch is shorted with the pin from center. Resonant frequencies of the purposed antenna are investigated by changing the antenna parameters like parasitic patch antenna location, Feeding Point, shorting pin location, and substrate height (h). It is observed that the proposed design can reduce size up to 60%.

Keywords—Parasitic Layer, Circular microstrip patch antenna (CMSPA), Compact size

I. INTRODUCTION

As the growth of communication engineering increases everyone wants to stay connected. This connection can be done by only communication link which can be done by Antenna only. The height of the antenna is inversely proportional to the frequency of operation [1]. For the UHF band the height of the antenna will be in meters only. For compact size, microstrip patch antenna is a good choice. Due to the growing demand for compact antennas for the reduced size of products used in personal communication microstrip patch antennas received much attention. The dimensions of the antenna can be minimized at the fixed resonant frequency by the use of substrate of high permittivity as larger permittivity substrate can result in smaller physical dimensions at the fixed resonant frequency. The electrical properties of an antenna are degraded by the use of dielectric as it extracts a part of the surface wave produced for direct radiation (space waves). It has been found in the literature that the dimensions of the antenna can be downsized effectively by shorting ground with the patch[2]-[5]. In literature, various methods are available to increased antenna gain and bandwidth. The substrate in the microstrip patch antenna is required for providing robustness to the antenna [6]. In the literature, many techniques are used for designing the Microstrip Patch Antenna for dual resonance frequency operation such as multilayer patch antenna [7], [8], slotted microstrip patch antenna [9] square patch with the introduction of notches [10], loading of shorting pin [11] or varactor diode[12] and feeding by inclined slot [13]. The

antenna size reduction is possible by shorting pin technique because of transfer of the null-voltage point from the center to the edges of the microstrip patch. The circular patch antenna dimensions can be effectively minimized, and this reduction is limited by the distance between the null-voltage point at the center of the patch and the edge of the patch.

In the present work, a compact circular microstrip patch antenna (CMSPA) with a multilayer is proposed. The feed location, shorting point location, parasitic patch position, and height of the substrate are varied and their effects on the performance parameters of the antenna are analyzed. Major contributions of the proposed work are i) The proposed antenna can work in the UHF TV band. ii) The relation between antenna parameters such as permittivity, substrate height, shorting pin position, feeding point, and location of a parasitic patch with the compactness of antenna is also described. iii) It is found that when shorting pin point location is transferred from center of the patch to the peripheral of the driven patch return loss is deteriorated. v) Shorting pin at center of ground significantly reduce the size of the antenna.

This paper is organized into four sections. The second section of the paper describes the design of proposed antennas while demonstrating the effect of parasitic patch location, feeding point, shorting point, and substrate thickness in the UHF TV band. The third section of the paper discusses the effect of these variations on antenna size reduction. The final section concludes the findings and observations in the presented work. The proposed antenna is designed with optimized feed location and centre of the driven patch is shorted with the ground by pin to minimize the dimensions of antenna.

II. PROPOSED ANTENNA DESIGN

A Compact CMSPA is designed with a fixed working frequency of 480MHz. The proposed design of circular microstrip patch antenna is shown in Fig.1.below. Here a is the radius of the circular Patch.

The radius of CMSPA is calculated from the design equation of circular patch antenna i.e.

$$a = \frac{F}{\left\{1 + \frac{2h}{\pi \epsilon_r F \left[\ln\left(\frac{\pi F}{2h}\right) + 1.7726\right]}\right\}^{1/2}} \quad (1)$$

$$\text{where } F = \frac{8.791 \times 10^9}{f_r \sqrt{\epsilon_r}}.$$

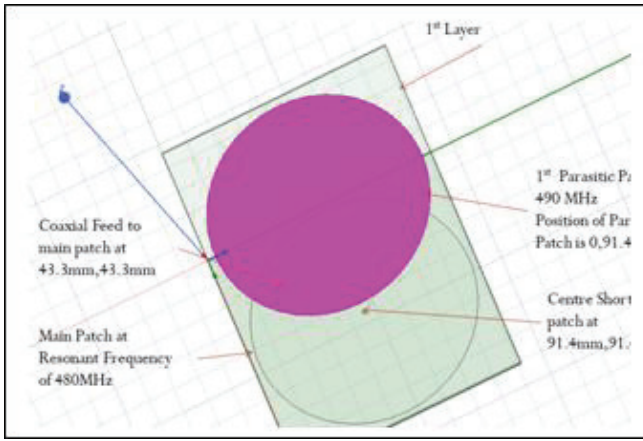


Fig. 1. Proposed Circular Microstrip Patch Antenna

In the proposed design substrate FR4 with permittivity 4.4 and substrate height, 1.6mm is used. The dimensions of the antenna can be minimized by shorting the patch with the ground. The center of CMSPA is shorted with the ground. Further, the size is reduced by using a parasitic layer. In the proposed design one parasitic layer is used on the main patch. The radius of the patch used on the first parasitic layer is also calculated from the design equation of circular patch antenna with a working frequency of 490MHz. Feed point varies at various location to get the optimized location of feeding. The position of a parasitic patch is varied from center to peripheral of the main patch to get the minimization of an antenna. Variation of substrate height and permittivity also studied here. High-Frequency Structure Simulator (HFSS) is used for designing of antenna and driven patch is feed by coaxial cable.

III. RESULTS

Location of feed point is optimized by taking values at different points and from Fig.2 it is cleared that when the feeding location is moving from center of the patch on its diagonal to the peripheral of the driven patch we are getting improved S_{11} towards resonant frequency. This optimized feed point is used for feeding of the proposed antenna.

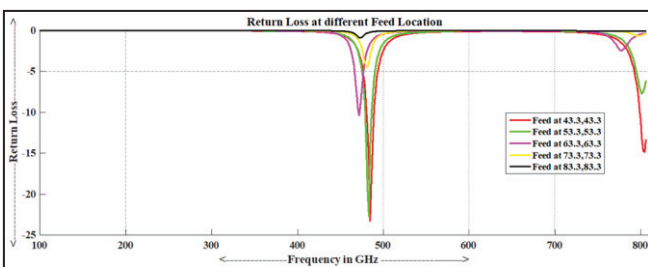


Fig. 2. Return Loss at different feed location

The S_{11} graph of the main CMSPA is shown in Fig.3 from the figure it is found that there is a difference of 4.3778MHz in the resonant frequency of calculated value and designed value. The gain of the proposed designed is shown in Fig. 4.

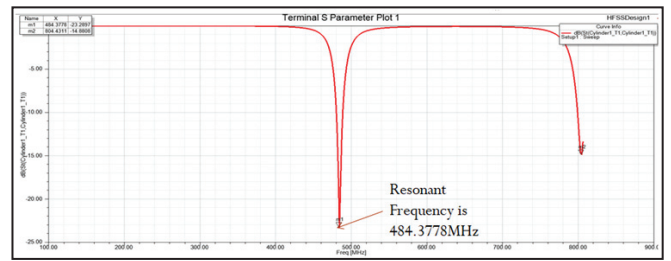


Fig. 3. Return Loss

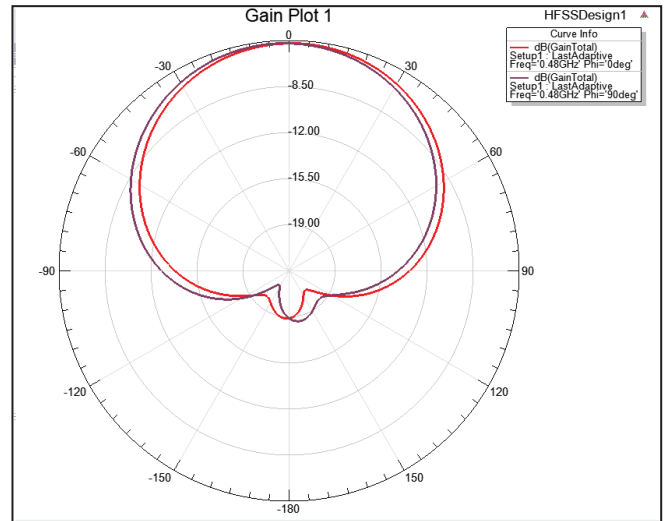


Fig. 4. Gain Plot

Now to optimize the location of shorting pin various points have been simulated and after simulation, it is verified from Fig. 5, that when shorting pin location is moving from center to the peripheral of the driven patch return loss is going to deteriorate. Size reduction is possible when center of the driven patch is shorted with the ground.

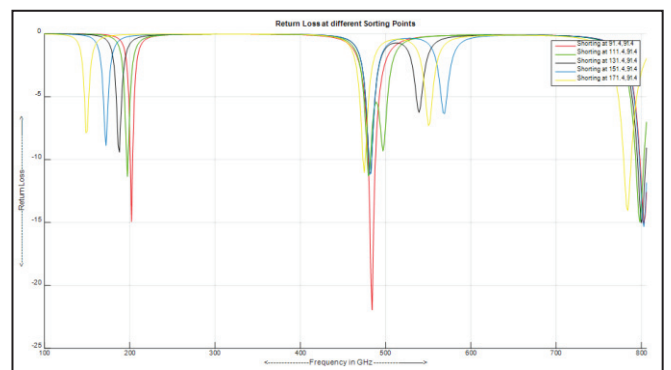


Fig. 5. Return Loss at different location of shorting pin

It is found that without shorting pin circular patch antenna resonates at its lowest frequency i.e. 484.3778 MHz. It is shown in Fig. 6 that, when the center of the circular patch is shorted with a pin, the resonant frequency is significantly reduced. It can be seen that positions of shorting pin affect the resonant frequency. 58 % size reduction can be possible by shorting pin technique.

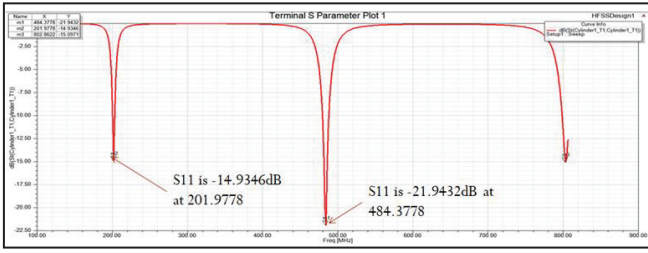


Fig. 6. Return Loss of center short main CMSPA

As shown in Fig.1 the location of 1st parasitic patch is also varied to get optimized results. The location of the 1st parasitic patch is varied from center to peripheral of the main patch and it is cleared from Fig 7 that the resonant frequency shifted from 484.3778MHz to 192.56MHz. S_{11} is also improving and resonant frequency is decreasing as the parasitic patch position is moving from center to peripheral of the driven patch. Corresponding values are shown in Table 1 which shows that at center position proposed antenna is working at a resonant frequency of 478.10MHz and a lower frequency of 192.56MHz at the peripheral position. 60% size reduction is achieved by using a parasitic patch at the peripheral with optimized feed location and shorting pin position on the driven patch.

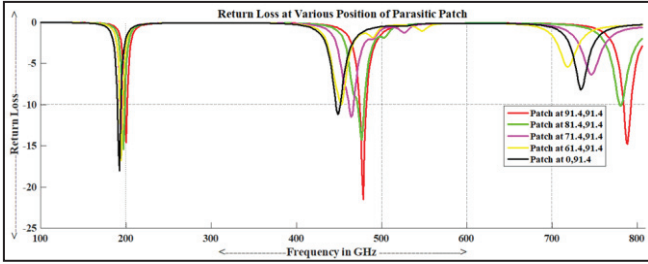


Fig. 7. Return Loss at the various position of parasitic patch

TABLE I. RETURN LOSS AT DIFFERENT FEED LOCATION

Location	S_{11} at Lower Frequency	S_{11} at Resonant Frequency
91.4,91.4	-14.58dB(200.40MHz)	-21.57 dB(478.10MHz)
81.4,91.4	-15.41dB(197.27MHz)	-14.29 dB(476.53MHz)
71.4,91.4	-16.27dB(194.13MHz)	-11.49 dB(463.98MHz)
61.4,91.4	-16.76dB(194.10MHz)	-11.34 dB(453MHz)
0,91.4	-18.03dB(192.56MHz)	-11.23 dB(448.29MHz)

Substrate height also affects the miniaturization of an antenna. From Fig. 8 and 9, it is found that for size reduction height of the parasitic substrate as well as the main substrate should minimum.

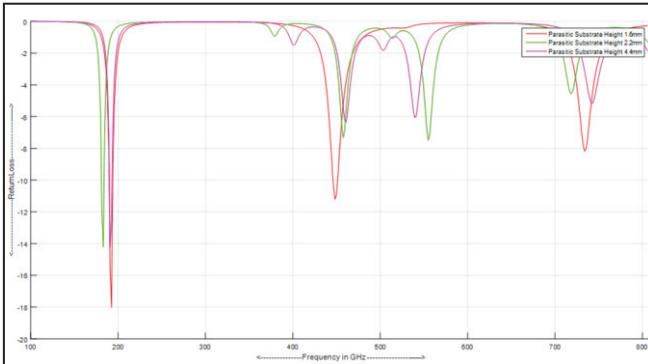


Fig. 8. Return loss for different height of the parasitic patch

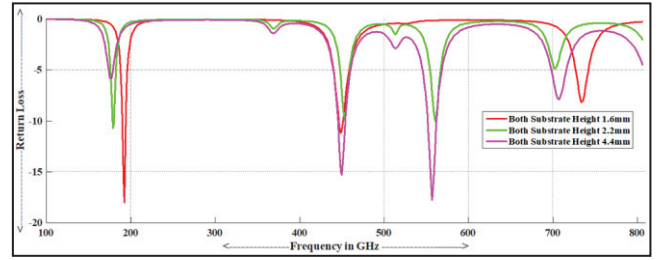


Fig. 9. Return Loss for different height of both substrate

IV. CONCLUSION

In this paper, the circular patch antenna is designed. The relation between resonant frequencies of a shorting pin-loaded circular patch antenna is described. The relation between antenna parameters such as permittivity, substrate height, shorting pin position, feeding point and, the location of parasitic patch with the compactness of antenna is also described. The proposed antenna is designed with optimized feed location and center of the driven patch is shorted with the ground by pin. In the proposed design the size can be reduced up to 60%. The error in the operating frequency of the designed and simulated antenna is approximately 0.4%. The proposed design is having return loss of -21.5dB.

REFERENCES

- [1] Balanis C.A., "Antenna Theory: Analysis and Design", John Wiley & Sons, Inc, 1997.
- [2] S.Dey and R. Mittra, "Compact microstrip patch antenna", Microwave Opt. Technol. Lett. 13, 12-14, Sept. 1996.
- [3] R. Waterhouse, "Small microstrip patch antenna", Electron. Lett. 31, 604-605, April, 1995.
- [4] C.L. Tang, H.T. Chen and K.L. Wong, "Small circular microstrip antenna with dual frequency operation", Microwave Opt. Lett. 33, 1112-1113, June 19, 1997.
- [5] K. L. Wong, C. L. Tang and H. T. Chen, "A Compact meandered circular microstrip patch antenna with a shorting pin", Microwave Opt. Technol. Lett. 15, 147-149, June 20, 1997.
- [6] Ramesh Garg, Prakash Bhartia, Inder Bahl, Apisak Ittipiboon, "Microstrip Antenna Design Handbook", Artech House, Boston, London, pp. 759-768, 2001.
- [7] J. S. Dahele, K. F. Lee, and D. P. Wong, "Dual-frequency stacked annular-ring microstrip antennas", IEEE Trans. Antennas Propag., vol. AP-35, no. 11, pp. 1281-1285, Nov. 1987.
- [8] S. A. Long and M. D. Walton, "A dual-frequency stacked circular disc antenna", IEEE Trans. Antennas Propag., vol. AP-27, no. 2, pp. 270-273, Mar. 1979.
- [9] S. Maci, G. B. Gentili, and G. Avitabile, "Single-layer dual-frequency patch antenna", Electron. Lett., vol. 29, pp. 1441-1443, 1993.
- [10] H. Nakano and K. Vichien, "Dual-frequency square patch antenna with rectangular notch", Electron. Lett., vol. 25, pp. 1067-1068, 1989.
- [11] D. H. Schaubert, F. G. Farrar, A. Sindoris, and S. T. Hayes, "Microstrip antennas with frequency agility and polarization diversity", IEEE Trans. Antennas Propag., vol. AP-29, no. 1, pp. 118-123, Jan. 1981.
- [12] R. B. Waterhouse and N. V. Shuley, "Dual-frequency microstrip rectangular patches", Electron. Lett., vol. 28, 1992.
- [13] [Y. M. M. Antar, A. I. Ittipiboon, and A. K. Bhattacharyya, "A dualfrequency antenna using a single patch and an inclined slot", Microw. Opt. Technol. Lett., vol. 8, pp. 309-311, 1995.
- [14] Khan, Taimoor, Asok De, and Moin Uddin. "Prediction of Slot-Size and Inserted Air-Gap for Improving the Performance of Rectangular Microstrip Antennas Using Artificial Neural Networks", IEEE Antennas and Wireless Propagation Lett., vol. 12, pp. 1367-1371, 2013.
- [15] Asok De, NS Raghava, Sagar Malhotra, Pushkar Arora, Rishik Bazaz, "Effect of different substrates on Compact stacked square Microstrip

- Antenna", Journal of Telecommunications, Volume 1, Issue 1, pp63-65, February 2010.
- [16] NS Raghava, A De, "Photonic bandgap stacked rectangular microstrip antenna for road vehicle communication", IEEE Antennas and wireless propagation letters, Vol-5, pp. 421-423, 2006.
- [17] N. S. Raghava and Asok De, "A Novel High-Performance Patch Radiator", International Journal of Microwave Science and Technology, vol-2008, pp.-4, 2008.



Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Design of photonic crystal OR gate with multi-input processing capability on a single structure

Chandan Kumar^a, Punit^a, Praveen Kumar^a, Preeti Rani^{b,*}, Yogita Kalra^a

^aTIFAC – Centre of Relevance and Excellence in Fiber Optics and Optical Communication, Department of Applied Physics, Delhi Technological University (Formerly Delhi College of Engineering), Bawana Road, Delhi 110042, India

^bSharda University, Knowledge Park III, Greater Noida, UP 201310, India

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Logic gate
Optical circuit

ABSTRACT

In this paper, the design of the cavity-based photonic crystal (PhC) 'OR' Gate has been proposed. The structure consists of three waveguides and two square resonators in a PhC composed of cylindrical rods of GaAs in air. The designed logic gate possesses multi input processing capability. The processing of multiple inputs on a single structure (Hardware) makes it idiosyncratic. Owing to its petiteness, fastness, and multi-input processing ability, the proposed structure overcomes constraints of modern electronics.

© 2021 Elsevier Ltd. All rights reserved.

Selection and peer-review under responsibility of the scientific committee of the National Conference on Recent Advances in Functional Materials-2020.

1. Introduction

The speed of light is the fastest speed humans have recorded yet. So it is an intelligent practice to make use of optical waves in order to make current devices faster. For instance, using optical signals in Fibre-optic communication boosted data transmission speed. Owing to its high speed, no radio frequency (RF) interference and low power requirements, optical signals are far more significant than electrical signals. Consequently, the concept of using optical signals for processing purposes has recently prevailed. The devices based on optical signal processing have a large bandwidth and high speed. If optical processors are developed there will be no need of converting optical signals from optical fibre back to electrical signals. Furthermore, a natural processor i.e. brain uses electrical signals, hence using electrical signals in processing devices will end up with a processor similar to the brain but using optical signals will lead to significantly different processors [1].

The most fundamental component of processing devices are logic gates. Hence optical logic gates are a very essential component for future optical processors and integrated optical circuits. Thus a lot of research work has been done on optical logic gates [2–11].

An optical logic PhC 'OR' logic gate, with a provision of processing multiple inputs on a single structure has been designed and

analysed. If the inputs of different frequencies are given to the proposed optical 'OR' gate, then output of respective frequencies is obtained at the output port which can be separated by optical filters. The PhC consists of GaAs rods assembled in a square lattice, with air in their gaps. Two square resonators and three waveguides have been drafted on a photonic crystal structure. The square resonator is highly frequency-selective, making the design effective. In this paper, two optical signals of different frequencies have been used to provide two sets of input signals [(0,1)&(0,0)], and different output for both frequencies i.e. '1' for the first frequency and '0' for the second frequency which can be separated at the end. Hence, two input sets are processed on a single structure, which will double the processing speed of the logic gate.

2. Design of the proposed 'OR' logic gate

A two-dimensional PhC has been used to design and analyse the optical 'OR' logic gate. It consists of $24a \times 10a$ two dimensional square lattice where 'a' is lattice constant with value $0.595 \mu\text{m}$. In the proposed structure radius 'r' and the refractive index of GaAs rods have been taken as $0.09 \mu\text{m}$ and 3.5 respectively.

Fig. 1 illustrates the design of the proposed logic gate. It consists of three waveguides of width $2a$ and two square resonators having inner and outer side length $3a$ and $5a$ respectively. The structure has six ports. Port A & Port B are input ports and Port C is the output port. Port 'C' & 'F' are Backward and Forward drop Output and Port 'D' & 'E' are Transmission port [12].

* Corresponding author.

E-mail address: preeti1703.soni@yahoo.in (P. Rani).

<https://doi.org/10.1016/j.matpr.2021.04.252>

2214-7853/© 2021 Elsevier Ltd. All rights reserved.

Selection and peer-review under responsibility of the scientific committee of the National Conference on Recent Advances in Functional Materials-2020.

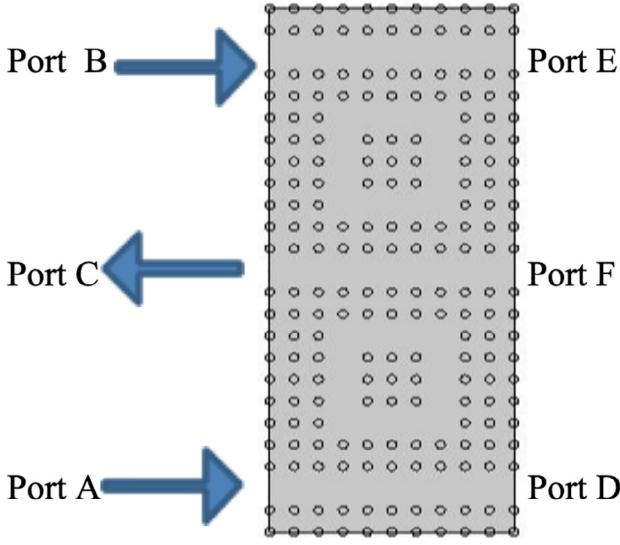


Fig. 1. The proposed structure of PhC OR gate.

3. Operating principle

3.1. Resonant wavelength and operating wavelength

At a certain wavelength, an optical wave coming at the input port reaches the backward drop port through the square resonator. This wavelength is called resonant wavelength. Optical waves having wavelength equal to resonant wavelength interfere constructively in square loop. Hence the round trip path difference in the square loop is an integral multiple of wavelength. Theoretical calculation of Resonant wavelength of the square resonator is as follows:

Path difference of the wave in a round trip inside the cavity is equal to the perimeter of the square which lies between $4 \times 3a$ to $4 \times 5a$ i.e. $12a < \Delta x < 20a$. Average path difference is $4 \times 4a$, $\Delta x = 4 \times 4a$.

Let λ be the resonant wavelength, hence it satisfies the condition

$$\Delta x = n\lambda, \text{ where } n \text{ is an integer}$$

$$\Rightarrow 16a = n\lambda$$

$$\Rightarrow \lambda = 9.5, 4.7, 3.1, 2.3, 1.9, 1.5, 1.3, 1.1 \dots \mu\text{m}, \text{ For } n = 1, 2, 3 \dots$$

The operating wavelength is the wavelength which satisfies resonating conditions as well as lie in a photonic band gap (1265 nm to 1750 nm) as shown in Fig. 2. Hence the operating wavelength 1.5 μm lies in the middle of the gap for proper propagation.

To choose operating wavelength precisely, optical waves having wavelength in the operating range are given at input port and the electric field strength is measured at Backward drop port (Port C) and Transmission Port (Port D). As shown in Fig. 3 & 4, a sharp peak as well as a sharp trough is observed at Port C and Port D respectively when input wavelength is 1550 nm. Hence this wavelength is chosen as an operating wavelength. Fig. 5 compares electric field distribution at wavelength 1540 nm and 1550 nm and shows that an optical wave having wavelength equal to 1550 nm reaches port C through the square resonator.

3.2. Impact of Rod's radius variation on operating wavelength

Fig. 6 shows the variation of Electric Field strength at the output port with wavelength of the input optical wave, for various values of radius 'r'. The operating wavelength i.e. the wavelength corresponding to the maximum Electric field strength, increases with

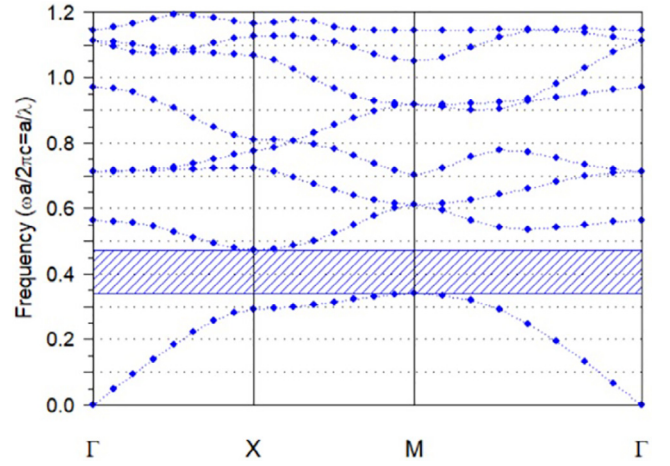


Fig. 2. Band gap of proposed design for TM mode.

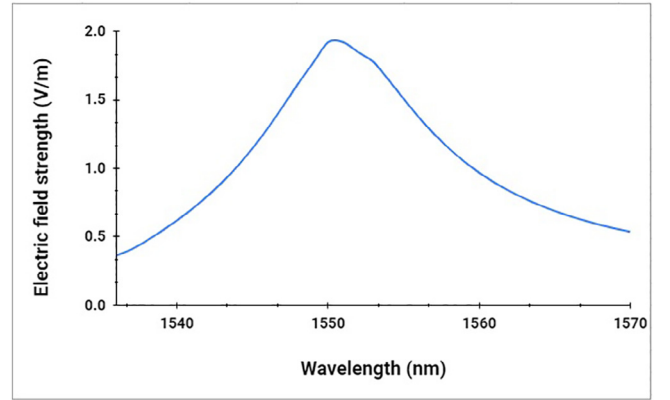


Fig. 3. Electric field strength at Backward drop Port.

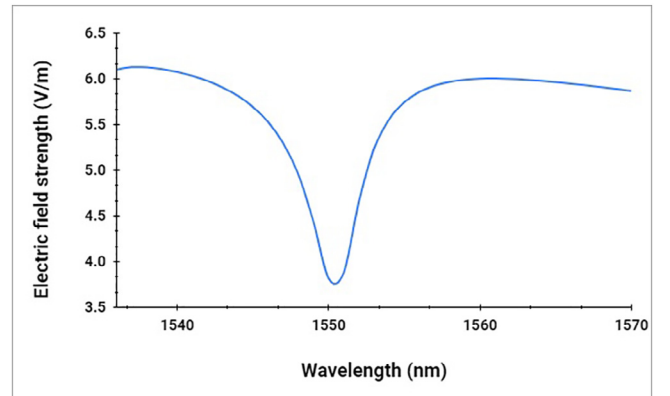


Fig. 4. Electric field strength at Transmission port.

the increase in 'r'. For $r = 80$ nm its value is 1520 nm and as 'r' increases to 100 nm, its value increases to 1574 nm.

Fig. 7 shows the variation of operating wavelength with the radius of the rod. The proposed structure gives freedom to choose the radius of the rod, which makes structure compatible with the manufacturing process. The desired operating wavelength can be achieved too, by opting the rod's radius appropriately.

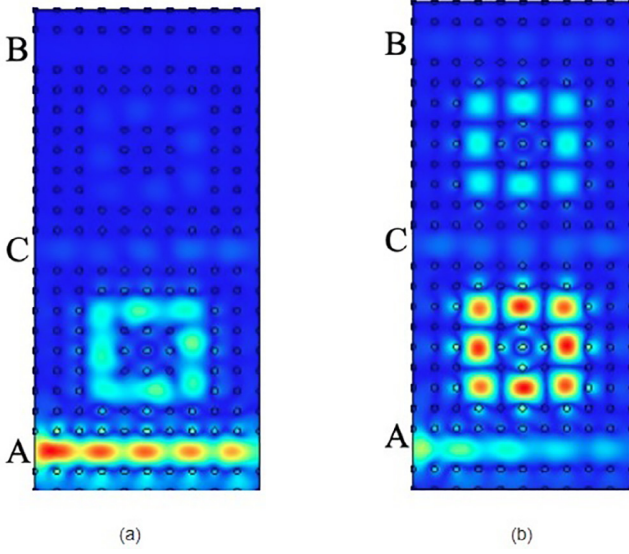


Fig. 5. Electric Field distribution for input wavelength (a) 1540 nm & (b) 1550 nm.

4. Results and discussion

The two-dimensional PhC optical logic 'OR' gate structure has been simulated on Comsol Multiphysics, Version 5.5 using Electro-magnetic Wave Frequency Domain (EWFD) Wave Optics Module. Finite element method is used to analyse the Logic gate performance.

4.1. Device operation as an 'OR' gate

Fig. 8 shows working states of the proposed logic OR gate. When logic '1' is given at Port A and logic '0' at Port B, electric field strength at Port C is obtained to be 1.92 V/m. Hence the output is logic '1'. If logic '1' is given at both ports i.e. Port A & Port B, then electric field strength at Port C is 3.78 V/m as shown in Fig. 8. Hence, it has been analysed that if any one input is logic '1' the output is also logic '1' i.e. the logic gate behaves as 'OR' Gate. Table 1 represents the truth table of the logic gate.

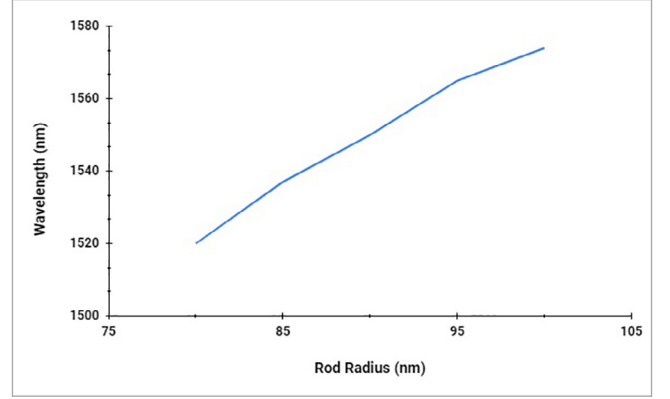


Fig. 7. Operating wavelength variation with rod radius.

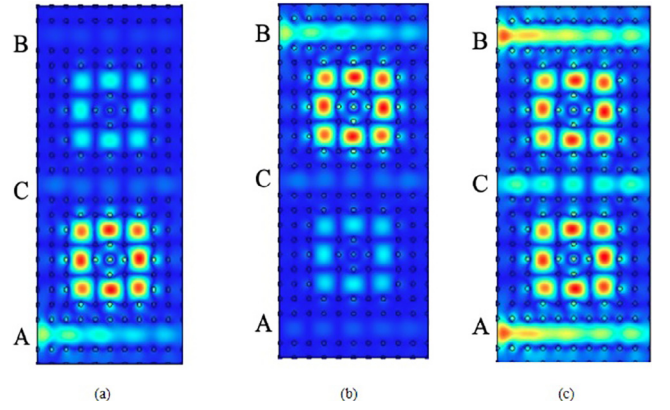


Fig. 8. Electric field distribution for input (a) A = 1, B = 0 (b) A = 0, B = 1 (c) A = 1, B = 1.

5. Features of multiple input processing

Fig. 9 shows that an electric field strength of 3.17 V/m is obtained at the output port when a wave having a wavelength of 1657 nm is used. If both waves of wavelength 1550 nm &

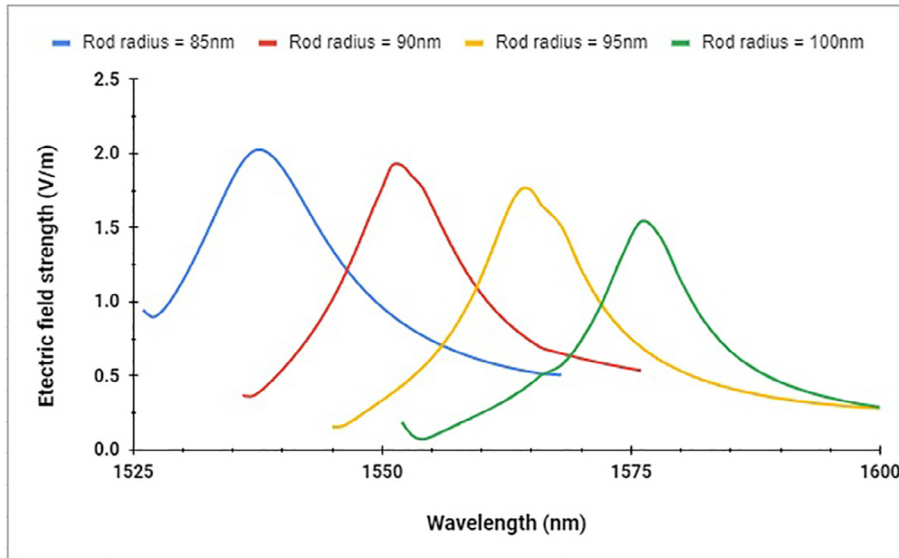


Fig. 6. Electric field strength at Port C vs Wavelength, for various values of 'r'.

Table 1
Truth Table of PhC 'OR' gate.

Port A	Port B	Port C	E field at output
0	0	0	0.00 V/m
0	1	1	1.92 V/m
1	0	1	1.92 V/m
1	1	1	3.78 V/m

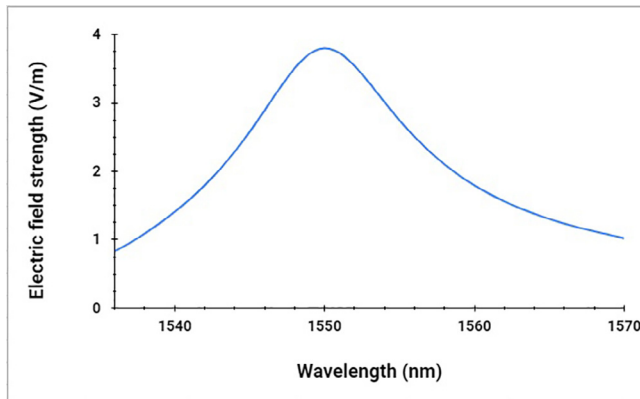


Fig. 9. Electric field strength at Port C when input is given at both Port A & Port B.

1657 nm are given at the input Port A simultaneously, then at Port C there will be the corresponding output for both. Since their frequencies are significantly different hence they can be separated by using an optical filter. In this way, two inputs can be processed on a single structure. By adjusting the perimeter of the square, radius of rods and lattice constant, the wave of desired frequencies can be used. In electronic circuits, the approach to process multiple inputs is not feasible as the response of electronic devices such as BJT, FET, etc varies significantly with frequency. The proposed method of using multiple wavelengths in a single structure not only makes the logic gate faster but also reduces the size and the cost of the device.

6. Conclusion

In this paper, we have depicted the design of the PhC optical logic 'OR' gate and discussed its performance based on simulation results. The device is performing as an 'OR' gate as depicted by the Table 1. The performance has been optimized by varying parameters like the lattice constant, the radius of rods and square side length. As the device is frequency selective therefore it is far less affected by the noise as compared to electronic devices. The device is capable of processing multiple inputs simultaneously which makes the device faster. The analysis of performance had been done by the Finite Element method on the Comsol Multiphysics Version5.5 platform. Non-Linear effects were not included hence

it can operate at low power although the variation of refractive index with frequency is included. If the feature of multi-input processing in further devices like Flip-Flop, Mux, Adder, etc is implemented then the vision of a very fast and compact processor can be accomplished.

CRediT authorship contribution statement

Chandan Kumar: Formal analysis, Investigation, Conceptualization, Writing - original draft. **Punit:** Data curation. **Praveen Kumar:** Visualization, Investigation. **Preeti Rani:** Project administration, Writing - review & editing, Supervision. **Yogita Kalra:** Funding acquisition, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Authors gratefully acknowledge the support of "TIFAC-Center of Relevance and Excellence in Fiber Optics and Optical Communication" at Delhi College of Engineering, Delhi, through Mission Reach Program of Technology Vision 2020, Government of India and Sharda University for providing various resources."

References

- [1] M. Sumathi, K. Venkatraman, R. Narmadha, C. Chhaya, Study of optical transmission and digitization of analog signals, in: 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), IEEE, 14833120, 2014, pp. 133–136.
- [2] R. Preeti, K. Yogita, R.K. Sinha, Realization of AND gate in y shaped photonic crystal waveguide, *Opt. Commun.* 298–299 (2013) 227–231.
- [3] R. Preeti, K. Yogita, R.K. Sinha, Design of all optical logic gates in photonic crystal waveguides, *Optik* 126 (9–10) (2015) 950–955.
- [4] P. Singh, D. Tripathi, S. Jaiswal, H. Dixit, All-optical logic gates: designs, classification, and comparison, *Adv. Optical Technol.* (2014) 1–13.
- [5] P. Fariborz, M. Mitra, Designing and simulation of 3-input majority gate based on two-dimensional photonic crystals, *Optik* 216 (2020) 164930.
- [6] L. Weijia, Y. Daquan, S. Guansheng, T. Huiping, J.i. Yuefeng, Design of ultra compact all-optical XOR, XNOR, NAND And OR gates using photonic crystal multi-mode interference waveguides, *Opt. Laser Technol.* 50 (2013) 55–64.
- [7] G. Kiyanoosh, M. Ali, C. Iman, G. Dariush, All-Optical XOR and OR logic gates based on line and point defects in 2-D photonic crystal, *Opt. Laser Technol.* 78 (2016) 139–142.
- [8] A.-B. Hamed, S. Somaye, M. Farhad, All optical NOR and NAND gate based on nonlinear photonic crystal ring resonators, *Optik* 125 (19) (2014) 5701–5704.
- [9] F. Parandin, Malmir M. Reza, M. Naseri, A. Zahedi, Reconfigurable all-optical NOT, XOR, and NOR logic gates based on two dimensional photonic crystals, *Superlattices Microstruct.* 113 (2018) 737–744.
- [10] A. Srinivasulu, Modified optical OR and AND gates, *Semicond. Phys., Quantum Electr. Optoelectr.* 5 (4) (2002) 428–430.
- [11] Li Zhangjian, Chen Zhiwen, Li Baojun, Optical pulse controlled all-optical logic gates in SiGe/Si multimode interference, *Optics Express*, 13(3) (2005) 1033.
- [12] T. Sreenivasulu, Kolli V. Rao, T.R. Yadunath, T. Badrinarayana, A. Sahu, G. Hegde, S. Mohan, T. Srinivas, Photonic crystal-based force sensor to measure sub-micro newton forces over a wide range, *Curr. Sci.* 110 (10) (2016) 1989.

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/351428375>

Designing and Analyzing the Brake Master Cylinder for an ATV vehicle

Article · January 2020

DOI: 10.35121/ijapie202001143

CITATIONS

0

READS

14

6 authors, including:



Rakesh Chander Saini

Maharaja Agarsain Institute of Technology

13 PUBLICATIONS 5 CITATIONS

[SEE PROFILE](#)



Ramakant Rana

Maharaja Agarsain Institute of Technology

35 PUBLICATIONS 65 CITATIONS

[SEE PROFILE](#)



Roop Lal Rana

Delhi Technological University

17 PUBLICATIONS 94 CITATIONS

[SEE PROFILE](#)

Designing and Analyzing the Brake Master Cylinder for an ATV vehicle

Shubham Upadhyaya¹, Divyam Raj¹, Kaushal Gupta¹, Rakesh Chander Saini^{2,*}, Ramakant Rana², Roop Lal³

(¹Student, Mechanical and Automation Engineering Department, Maharaja Agrasen Institute of Technology, Delhi India, ²Assistant Professor, Mechanical and Automation Engineering Department, Maharaja Agrasen Institute of Technology, Delhi India, ³Assistant Professor, Mechanical Engineering Department, Delhi Technological University, Delhi India)

*Email: rakeshchandersaini@gmail.com

ABSTRACT: Braking system is a means of converting momentum into heat energy by creating friction in the wheel brakes. The braking system which works with the help of hydraulic principles is known as hydraulic braking systems. The most frequently used system operates hydraulically, by pressure applied through a liquid. These are the foot operated brakes that the driver normally uses to slow or stop the car. Our special interest in hydraulics is related to the actions in automotive systems that result from pressure applied to a liquid. This is called hydraulic pressure. Since liquid is not compressible, it can transmit motion. A typical braking system includes two basic parts. These are the master cylinder with brake pedal and the wheel brake mechanism. The other parts are the connecting tubing, or brake lines, and the supporting arrangements. The present paper is about designing of Twin master cylinder system for an all-terrain vehicle and doing a feasibility study of its strength using ANSYS. Our work is focused on reducing weight which is one of the factors to increase the efficiency. Reduction in weight and space, due to its compactness. The twin Master cylinder system is a great advancement in braking system for an ATV. 3-D CAD modeling is done using SOLIDWORKS 2017, whereas the analysis of its strength is done using ANSYS.

Keywords: Hydraulic System, Brake, Master Cylinder, Analysis, Design, Twin Master Cylinder

I. INTRODUCTION

Master cylinder is a component of hydraulic braking system and it is just a simple piston inside a cylinder. Master cylinder is the key element of braking system which initiates and controls the braking action. A reservoir is attached to the master cylinder to store brake fluid. A master cylinder having a reservoir and a cylinder formed from a single piece of molded material. Master cylinder is a component of hydraulic braking system and it is just a simple piston inside a cylinder. Master cylinder is the key element of braking system which initiates and controls the braking action. A reservoir is attached to the master cylinder to store brake fluid. A master cylinder having a reservoir and a cylinder formed from a single piece of molded material [1-3]. The master cylinder displaces hydraulic pressure to the rest of the brake system. It holds the most important fluid in your car, the brake fluid. It actually controls two separate subsystems which are jointly activated by the brake pedal. This is done so that in case a major leak occurs in one system, the other will still function. The two systems may be supplied by separate fluid reservoirs, or they may be supplied by a common reservoir. Some brake subsystems are divided front/rear and some are diagonally separated. When you press the brake pedal, a push rod connected to the pedal moves the "primary piston" forward inside the master cylinder. The primary piston activates one of the two subsystems [4-6]. The hydraulic pressure created, and the force of the primary piston

spring, moves the secondary piston forward. When the forward movement of the pistons causes their primary cups to cover the bypass holes, hydraulic pressure builds up and is transmitted to the wheel cylinders. When the brake pedal retracts, the pistons allow fluid from the reservoir to refill the chamber if needed. Electronic sensors within the master cylinder are used to monitor the level of the fluid in the reservoirs, and to alert the driver if a pressure imbalance develops between the two systems. If the brake light comes on, the fluid level in the reservoir(s) should be checked. If the level is low, more fluid should be added, and the leak should be found and repaired as soon as possible [7-11].

The master cylinder displaces hydraulic pressure to the rest of the brake system. It holds the most important fluid in your car, the brake fluid [12]. It actually controls two separate subsystems which are jointly activated by the brake pedal. This is done so that in case a major leak occurs in one system, the other will still function [13-15]. The master cylinder displaces hydraulic pressure to the rest of the brake system. It holds the most important fluid in your car, the brake fluid. It actually controls two separate subsystems which are jointly activated by the brake pedal. This is done so that in case a major leak occurs in one system, the other will still function [16-19]. The two systems may be supplied by separate fluid reservoirs, or they may be supplied by a common reservoir. Some brake subsystems are divided front/rear and some are diagonally separated. When you press the brake pedal, a push rod connected to the pedal moves the "primary piston" forward inside the master cylinder. The primary piston activates one of the two subsystems [20-22]. The hydraulic pressure created, and the force of the primary piston spring, moves the secondary piston forward. When the forward movement of the pistons causes their primary cups to cover the bypass holes, hydraulic pressure builds up and is transmitted to the wheel cylinders [23-26].

II. DESIGN CONSIDERATIONS OF MASTER CYLINDER

The basic information about brake system and its master cylinder, function, purpose, working principle, different shape and size of master cylinder, failure considerations has been taken from automotive brake system. The work done by brake system parts manufacturers tells that cost mold brake master cylinder made of cast iron was used universally in all the old car and light trucks and after that there has been increased research done on improving the mileage of the vehicle by reducing the weight. The research made a way to concentrate on reducing the weight of brake master cylinder by changing the materials [27, 28].

The manufacturers came up with new idea of composite master cylinder having integral body made of aluminum and reservoir made of plastic material and thus reducing the weight when compare to cost mold master cylinder made of cast iron. Those manufacturers are concentrating on reducing weight of master cylinder by changing the material and by changing the type of manufacture [29]. This information gives basic steps for this project in taking reduction of weight further and considering plastic material to design brake master cylinder. The second edition of brake design and safety gives basic design considerations to design safer brakes and its components. The standard of quality of brake technology as changed over the last two decades. The new design can only be achieved through proper research, through the use of sound engineering concepts and testing the results of small design changes. The information provided by the author has helped in considering engineering design concepts, safety considerations, material selection, guides, standards and practices for the project [30].

III. Experiment Calculations

Important Parameters:

Pedal Force applied by driver (F_p) = 250 N Pedal Leverage = 4.5

Wheel Torque (T_c) = 161 Nm

Brake caliper piston diameter (D_c) = 32 mm Maximum piston travel of caliper (L_c) = 1.5mm Radius of disc (R) = 190 mm

Assumptions:

Deceleration = 0.8g

Coefficient of friction between tire and ground = 0.78 Coefficient of friction between pads and

Disc = 0.35 Dynamic weight transfer = 75.66 kg

Piston Diameter Calculations:

F_M = Force on master cylinder

F_m = $F_p \times l$

F_C = Force on caliper

F_c = T_c/R

A_c = Area of caliper piston

A = $(\pi/4) \times D_c^2$

P = Pressure in the system

P = F_c/A_c

A_m = Area of piston

A_m = F_m/P

M = Master cylinder bore diameter

D_m = $\sqrt{(A_m \times 4/\pi)}$

Stroke Length Calculations:

V = Volume displaced by caliper piston

V = $\pi \times D_c^2 \times L_c / 4$

L_m = Stroke length of master cylinder

L_m = $4 \times V / \pi \times D_m^2$

IV. CAD MODELING

Finite Element Analysis is a practical application of Finite Element Method (FEM). FEM is a numerical technique for finding approximate solutions to boundary value problems for partial differential equations. It uses subdivision of a whole problem domain into simpler parts, called finite elements, and variational methods from the calculus of variations to solve the problem by minimizing an associated error function. Analogous to the idea that connecting many tiny straight lines can approximate a larger circle, FEM encompasses methods for connecting many simple element equations over many small subdomains, named finite elements, to approximate a more complex equation over a larger domain.

A simple structural analysis was performed as the first step to see if components were structurally strong. If a component failed with the loadings, then no need to continue stress or fatigue analysis since the component is

not strong enough to be used. The analysis of the various components of the master cylinder was done in ANSYS 16.0 WORKBENCH for meshing as well as solving.

Meshing of all the parts was done in ANSYS. The mesh is generated by using tetrahedron elements of 1 mm size. Mesh quality is further improved by using proximity and curvature function. This improves mesh density where curvature is small or edges are closed in proximity.

Material used is Al 6061 with $S_{yt}=350$ Mpa,

Poisson's ratio=0.33 and Density=2700 kg/m³.

The boundary conditions applied are pressure generated in cylinder casing and the axial force applied through the push rod. The casing is fixed at the mounting points. For the braking system consider which is for an ATV the applied braking force is assumed to be 350 N. The force is magnify by the leverage of 4.5 provided by the pedal assembly and 1575 N force is applied by the push rod. Also the maximum pressure generated in system is applied on inner surfaces of casing.

The results of maximum stress and deformation shows that the master cylinder is safe for designed shell and mounting thickness.

Maximum Stress (Cylinder casing) = 177.6 Mpa

Maximum Deformation (Cylinder casing) = 0.02 mm

Maximum Stress (Piston) = 138.52 Mpa

Maximum Deformation (Piston) = 0.0108 mm

V. ANALYSIS

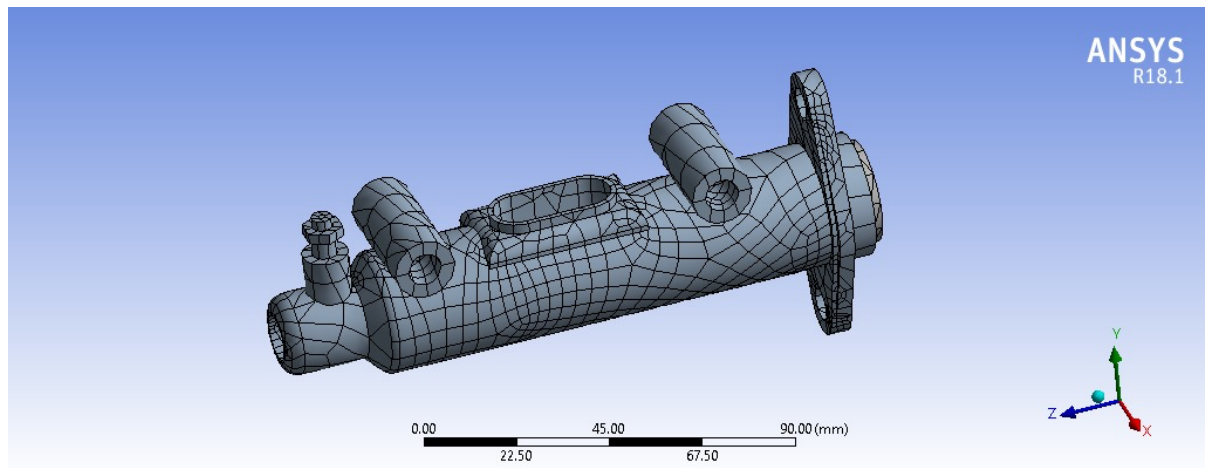


Figure 1: FEM Design analysis step 1

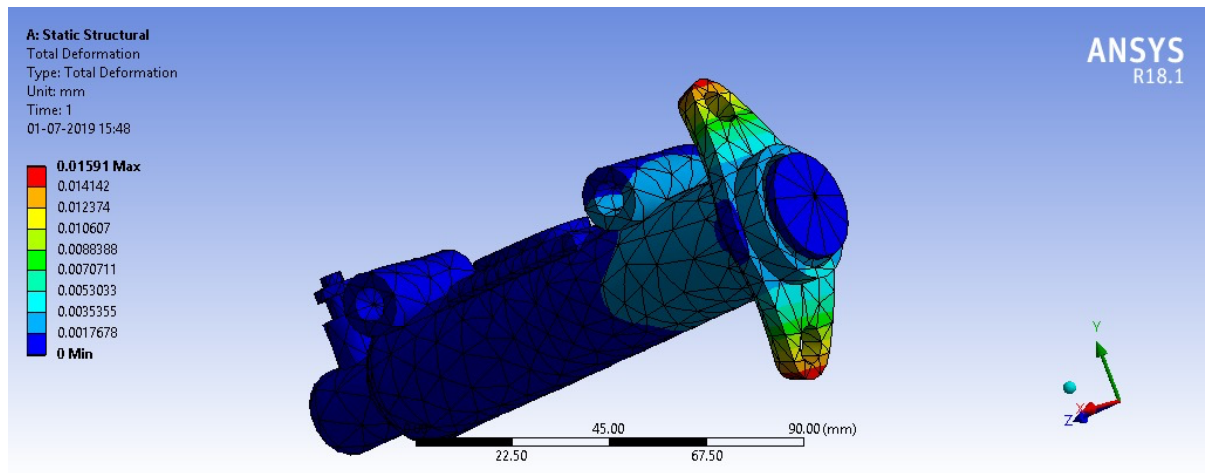


Figure 2: FEM Design analysis step 2

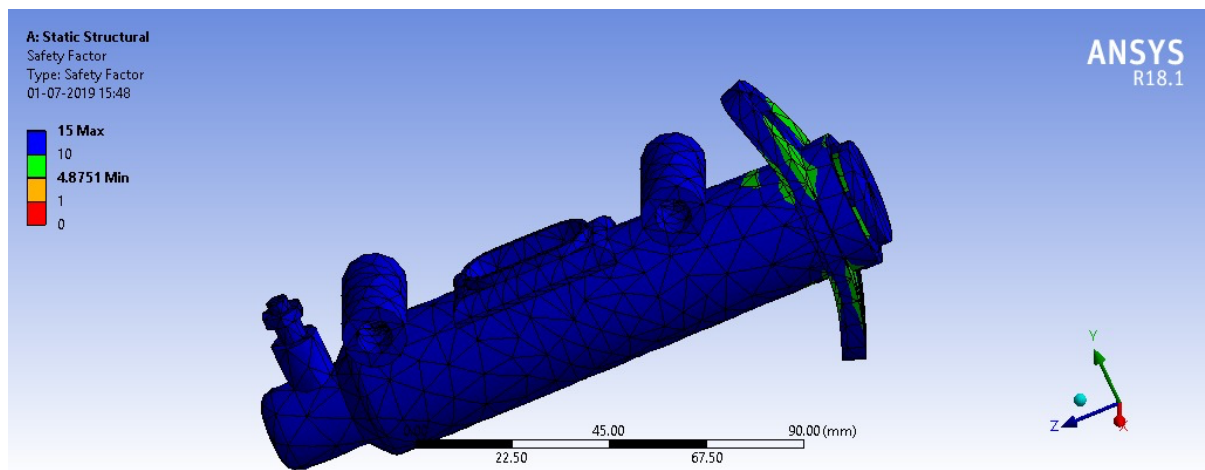


Figure 3: FEM Design analysis step 3

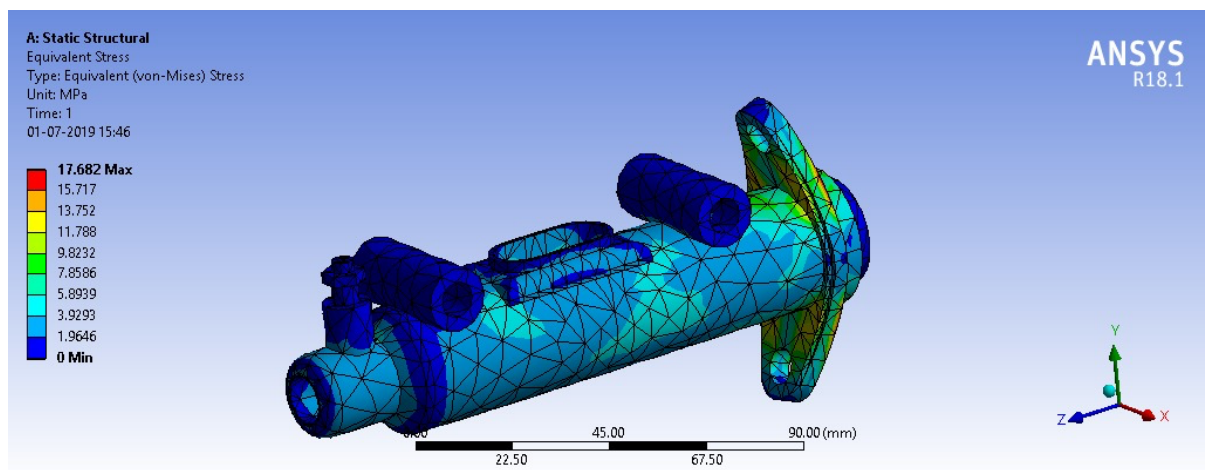


Figure 4: FEM Design analysis step 4

VI. CONCLUSIONS

Vehicle dynamics have been carefully studied. It included design of rear and front suspension, load transfer calculations, design of springs, selection of bearings and analysis in ANSYS Workbench. The purpose of the paper is not only the designing of suspension and steering of hybrid tricycle but also to provide in depth study to increase the performance of the vehicle in terms of vehicle dynamics. Design features have been proven effective in terms of vehicle dynamics and the results from FEA indicate the real track performance is quite safe.

VII. ACKNOWLEDGEMENT

Authors extend their regards to the Centre for Advanced Production and Industrial Engineering Research (CAPIER) of Delhi Technological University, New Delhi, India, for providing the layout for this research. Authors would also like to acknowledge and give special thanks to the support of “Metrology Lab” and “Research and Development Lab” of Maharaja Agrasen Institute of Technology, New Delhi, India, for providing the facilities for the completion of this work.

REFERENCES

- [1]. Ranganath, M. S. "Vipin, Optimization of Process Parameters in Turning Operation Using Taguchi Method and Anova: A Review." *international journal of advance research and innovation* 1 (2013): 31-45.
- [2]. Lata, Surabhi, Ankur Pandey, Ankit Sharma, Kuldeep Meena, Ramakant Rana, and Roop Lal. "An experimental study and analysis of the mechanical properties of titanium dioxide reinforced aluminum (AA 5051) composite." *Materials Today: Proceedings* 5, no. 2 (2018): 6090-6097.
- [3]. MS, Ranganath, and R. S. Vipin. "Neural Network Process Modelling for Turning of Aluminium (6061) using Cemented Carbide Inserts." *International Journal of Advance Research and Innovation*, 1, no. 3 (2013): 211-219.
- [4]. Rana, Srikant, Sumit Kumar, and Ramakant Rana., "Optimization of Temperature variations on Steel Grade EN-18 using Pin-on-disc Method", *International journal of advanced production and industrial engineering*, Delta 171, Vol 3 (1), 21-26.
- [5]. Madan, A. K., and M. S. Ranganath. "Application of selective inventory control techniques for cutting tool inventory modeling and inventory reduction-A case study." In *International Conference of Advance Research and Innovation (ICARI)*, pp. 127-135. 2014.
- [6]. Khanna, Rachit, Raghav Maheshwari, Anish Modi, Shivam Tyagi, Anupam Thakur, and Ramakant Rana. "A review on recent research development on Electric Discharge Machining (EDM)." *International Journal of Advance Research and Innovation*, Vol, 5, no. 4 (2017): 444-445.
- [7]. Kaplish, Akshit, Anurag Choubey, and Ramakant Rana. "Design and Kinematic Modelling Of Slave Manipulator For Remote Medical Diagnosis", *International Journal of Advanced Production and Industrial Engineering*, (2017): 19-22.
- [8]. Saxena, Himanshu, R. C. Singh, Rajiv Chaudhary, and Ranganath MS. "Experimental investigation of defective ball bearings with vibration analyzer." In *International Conference of Advanced Research and Innovation*. 2014.
- [9]. Rana, Ramakant, Walia, R. S. and Manik, Singla, "Effect of friction coefficient on En-31 with different pin materials using pin-on-disc apparatus." In *International conference on recent advances in mechanical engineering (RAME-2016)*, pp. 619-624. 2016.

- [10].Rana, Ramakant, Walia, R. S., Qasim, Murtaza and Mohit. Tyagi, "Parametric optimization of hybrid electrode EDM process." In TORONTO'2016 AESATEMA International Conference "Advances and Trends in Engineering Materials and their Applications, pp. 151-162. 2016.
- [11].Jain, Siddharth, Aggarwal, Vidit, Tyagi, Mohit, Walia, R. S. and Rana, Ramakant, "Development of aluminium matrix composite using coconut husk ash reinforcement." In International conference on latest developments in materials, manufacturing and quality control (MMQC-2016), pp. 12-13. 2016.
- [12].Rana, Ramakant. "Development of Hybrid EDM Electrode for Improving Surface Morphology." PhD diss., 2016.
- [13].Lata, Surabhi, Ashish Gupta, Aditya Jain, Sonu Kumar, Anindya Srivastava, Ramakant Rana, and Roop Lal. "A Review on Experimental Investigation of Machining Parameters during CNC Machining of OHNS." International Journal of Engineering Research and Applications 6 (2016): 63-71.
- [14].Lal, Roop, and Rana Ramakant. "A Textbook of Engineering Drawing", IK International Publishing House Pvt. Ltd., (2015) 1, 452.
- [15].Ramakant, Rana, Mani Adarsh, Anmol Kochhar, Shrey Wadhwa, Sandeep Kumar Daiya, Sparsh Taliyan, and Roop Lal,—An Overview On Process Parameters Improvement In Wire Electrical Discharge Machining." International Journal of Modern Engineering Research, Vol 5, Issue 4, (2015); 22-27.
- [16].Rana, Ramakant, Kunal Rajput, Rohit Saini, and Roop Lal. "Optimization of tool wear: a review." Int J Mod Eng Res 4, no. 11 (2014): 35-42.
- [17].Rana, Ramakant, Mitul Batra, Vipin Kumar Sharma, and Aditya Sahni. "Wear Analysis of Brass, Aluminium and Mild Steel by using Pin-on-disc Method.", 3rd International Conference on Manufacturing Excellence – MANFEX, (2016): 17-20
- [18].Lal, Roop, and R. C. Singh. "Investigations of tribodynamic characteristics of chrome steel pin against plain and textured surface cast iron discs in lubricated conditions." World Journal of Engineering, Vol. 16, No. 4, (2019): 560-568.
- [19].Singh, R. C., R. K. Pandey, M. S. Ranganath, and S. Maji. "Tribological performance analysis of textured steel surfaces under lubricating conditions." Surface Topography: Metrology and Properties 4, no. 3 (2016): 034005.
- [20].Singh, R. C., Roop Lal, M. S. Ranganath, and Rajiv Chaudhary. "Failure of piston in IC engines: A review." International Journal of Modern Engineering Research 4, no. 9 (2014): 1-10.
- [21].Lal, Roop, and R. C. Singh. "Experimental comparative study of chrome steel pin with and without chrome plated cast iron disc in situ fully flooded interface lubrication." Surface Topography: Metrology and Properties 6, no. 3 (2018): 035001.
- [22].Ranganath M. S. , Vipin, Mishra, R. S., "Effect of Cutting Parameters on MRR and Surface Roughness in Turning of Aluminium (6061)." International Journal of Advance Research and Innovation, Vol. 2, no. 1 (2014): 32-39.
- [23].Lal, Roop, R. C. Singh, M. S. Ranganath, and S. Maji. "Friction and Wear of Tribo-Elements in Power Producing Units for IC Engines-A Review." International Journal of Engineering Trends and Technology (IJETT)—Volume 14 (2014).
- [24].Ranganath, M. S. "Vipin,"Experimental Investigation and Parametric Analysis of Surface Roughness in CNC Turning Using Design of Experiments". International Journal of Modern Engineering Research 4, no. 9 (2014): 1-8.
- [25].Lal, Roop, R. C. Singh, Vaibhav Sharma, and Vaibhav Jain. "A Study of Active Brake System of Automobile." International Journal 5, no. 2 (2017): 251-254.
- [26].Chaudhary, Rajiv, M. S. Ranganath, and Vipin RC Singh. "Experimental investigations and Taguchi analysis with drilling operation: A review." International Journal of Innovation and Scientific Research, Vol. 13 No. 1, (2015): 126-135.

- [27].Lal, Roop, Mohd Shuaib, and Vikal Paliwal. "Comparative Study of Mechanical Properties of TIG Welded Joints of Similar and Dissimilar Grades of Stainless Steel Material." *International Journal* 6, no. 3 (2018): 205-208.
- [28].Ranganath, M. S., and Harshit Vipin. "Surface Roughness Prediction Model for CNC Turning of EN-8 Steel Using Response Surface Methodology." *International Journal of Emerging Technology and Advanced Engineering* 5, no. 6 (2015): 135-143.
- [29].Lal, Roop, R. C. Singh, and Davendra Singh. "Stress Analysis at Contact Region of Rail-Wheel.", Vth International Symposium on "Fusion of Science & Technology", New Delhi, India, January 18-22, (2016): 75-85.
- [30].Singh, Devendra, R. C. Singh, and Roop Lal. "Computational Static Analysis of Rail-Wheel Model of Indian Railways.", Vth International Symposium on "Fusion of Science & Technology", New Delhi, India, January 18-22, (2016): 106-113.

Detection of Cyberbullying on Social Media Using Machine learning

Varun Jain
Department of Information Technology
Delhi Technological University
Delhi, India
varunjn652@gmail.com

Vishant Kumar
Department of Information Technology
Delhi Technological University
Delhi, India
kumar.vishant229@gmail.com

Vivek Pal
Department of Information Technology
Delhi Technological University
Delhi, India
vpal9052@gmail.com

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

Abstract— Cyberbullying is a major problem encountered on internet that affects teenagers and also adults. It has led to mishappenings like suicide and depression. Regulation of content on Social media platforms has become a growing need. The following study uses data from two different forms of cyberbullying, hate speech tweets from Twitter and comments based on personal attacks from Wikipedia forums to build a model based on detection of Cyberbullying in text data using Natural Language Processing and Machine learning. Three methods for Feature extraction and four classifiers are studied to outline the best approach. For Tweet data the model provides accuracies above 90% and for Wikipedia data it gives accuracies above 80%.

Keywords—Cyberbullying, Hate speech, Personal attacks, Machine learning, Feature extraction, Twitter, Wikipedia

I. INTRODUCTION

Now more than ever technology has become an integral part of our life. With the evolution of the internet. Social media is trending these days. But as all the other things misusers will pop out sometimes late sometime early but there will be for sure. Now Cyberbullying is common these days.

Sites for social networking are excellent tools for communication within individuals. Use of social networking has become widespread over the years, though, in general people find immoral and unethical ways of negative stuff. We see this happening between teens or sometimes between young adults. One of the negative stuffs they do is bullying each other over the internet. In online environment we cannot easily said that whether someone is saying something just for fun or there may be other intention of him. Often, with just a joke, "or don't take it so seriously," they'll laugh it off. Cyberbullying is the use of technology to harass, threaten, embarrass, or target another person. Often this internet fight results into real life threats for some individual. Some people have turned to suicide. It is necessary to stop such activities at the beginning. Any actions could be taken to avoid this for example if an individual's tweet/post is found offensive then maybe his/her account can be terminated or suspended for a particular period.

So, what is cyberbullying??

Cyberbullying is harassment, threatening, embarrassing or targeting someone for the purpose of having fun or even by well-planned means

II. BACKGROUND

Researches on Cyberbullying Incidents show that 11.4% of 720 young peoples surveyed in the NCT DELHI were victims of cyberbullying in a 2018 survey by Child Right and You, an NGO in India, and almost half of them did not even mention it to their teachers, parents or guardians. 22.8% aged 13-18 who used the internet for around 3 hours a day were vulnerable to Cyberbullying while 28% of people who use internet more than 4 hours a day were victims. There are so many other reports suggested us that the impact of Cyberbullying is affecting badly the peoples and children between age of 13 to 20 face so many difficulties in terms of health, mental fitness and their decision making capability in any work. Researchers suggest that every country should have to take this matter seriously and try to find solution. In 2016 an incident called Blue Whale Challenge led to lots of child suicides in Russia and other countries. It was a game that spread over different social networks and it was a relationship between an administrator and a participant. For fifty days certain tasks are given to participants. Initially they are easy like waking up at 4:30 AM or watching a horror movie. But later they escalated to self harm which led to suicides. The administrators were found later to be children between ages 12-14.

III. LITERATURE SURVEY

Lot of research have been done to find possible solutions to detect Cyberbullying on social networking sites. Ting, I-

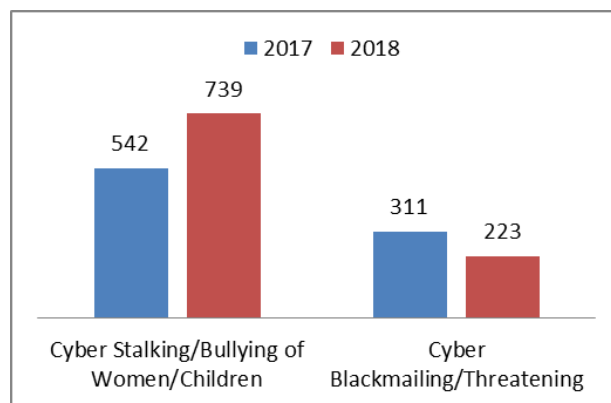


Fig. 1. Cyberbullying cases in India 2017-2018

Hsien[1] used an approach using keyword matching, opinion mining and social network analysis and got a precision of 0.79 and recall of 0.71 from datasets from four websites. Patxi Gal'an-Garc'ia et al.[2] proposed a hypothesis that a troll (one who cyberbullies) on a social networking sites under a fake profile always has a real profile to check how other see the fake profile. They proposed a Machine learning approach to determine such profiles. The identification process studied some profiles which has some kind of close relation to them. The method used was to select profiles for study, acquire information of tweets, select features to be used from profiles and using ML to find the author of tweets. 1900 tweets were used belonging to 19 different profiles. It had an accuracy of 68% for identifying author. Later it was used in a Case Study in a school in Spain where out of some suspected students for Cyberbullying the real owner of a profile had to be found and the method worked in the case. The following method still has some shortcomings. For example a case where trolling account doesn't have a real account to fool such systems or experts who can change writing styles and behaviours so that no patterns are found. For changing writing styles more efficient algorithms will be needed.

Mangaonkar et al. [3] proposed a collaborative detection method where there are multiple detection nodes connected to each other where each node uses either different or same algorithm and data and results were combined to produce results. P. Zhou et al.[4] suggested a B-LSTM technique based on concentration. Banerjee et al.[5]. used KNN with new embeddings to get an precision of 93%.

Kelly Reynolds, April Kontostathis and Lynne Edwards[6] propose a Formpring (A forum for anonymous questions-answers) dataset which gives recall of 78.5% using Machine learning Algorithms and oversampling due to imbalance in cyberbullying posts. Jaideep Yadav, Kumar and Chauhan [7] used a latest language model developed by google called BERT which generates contextual embeddings for classification. The model gave a F1 score of 0.94 on form spring data and 0.81 on Wikipedia data. Maral Dadvar and Kai Eckert[8] trained deep neural networks on Twitter, Wikipedia and Formspring datasets and used the model on Youtube dataset for the same and achieved F1 score of 0.97 using Bidirectional Long Short-Term Memory (BLSTM) model. Sweta Agrawal and Amit Awekar [9] used similar same datasets for training Deep Neural Networks but one of its key focus is swear words and their use as features for the task. They determined how the vocabulary for such models varies across various Social Media Platforms. Yasin N. Silva, Christopher Rich and Deborah Hall[10] built BullyBlocker, a mobile application that informs parents of cyberbullying activities against their child on Facebook which counted warning signs and vulnerability factors to calculate a value to measure probability of being bullied.

IV. PROPOSED METHODOLOGY

Cyberbullying detection is solved in this project as a binary classification problem where we are detecting two majors form of Cyberbullying: hate speech on Twitter and Personal attacks on Wikipedia and classifying them as containing Cyberbullying or not.

Fig. 2 describes the methodology used for solving the problem which is applied on both the datasets.

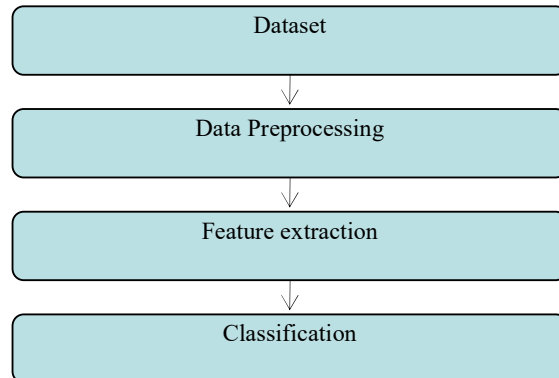


Fig. 2. Methodology

V. DATASET

A. Twitter Dataset

The Twitter Dataset is combined from two datasets containing hate speech :

- Hate Speech Twitter Dataset by Waseem, Zeerak and Hovy, Dirk[11] which contains 17000 tweets labelled for sexism or racism. The tweets are mined using the annotations .5900 tweets are lost due to accounts being deactivated or tweet deleted.
- Hate Speech Language Dataset by Davidson, Thomas and Warmley, Dana and Macy, Michael and Weber, Ingmar[12]. It contained 25000 tweets obtained by crowdsourcing.

This gives total 35787 tweets for the task distribution for which is shown in Fig. 3. For the following dataset, 70 percent (25,050) of this dataset is used as training data and 30 percent as testing data (10,737) .

B. Wikipedia Dataset

The Wikipedia dataset by Wulczyn, Thain and Dixon[13] contains 1M comments labelled for Personal attacks. For the analysis 40000 comments are used from the dataset from which 13000 comments are labelled as Cyberbullying due to personal attack. These comments are extracted from conversations between editors of pages on Wikipedia labelled by 10 annotators via Crowd Flower. For this dataset the same split (70 percent i.e 28000 to training data and 30 percent i.e 12000 to testing data) is used. Fig. 4 shows its distribution

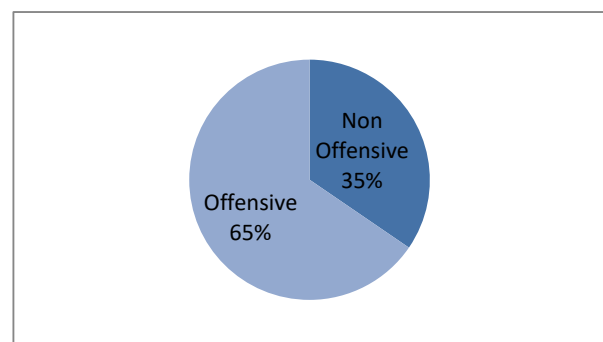


Fig. 3. Distribution of Tweets in Twitter Dataset

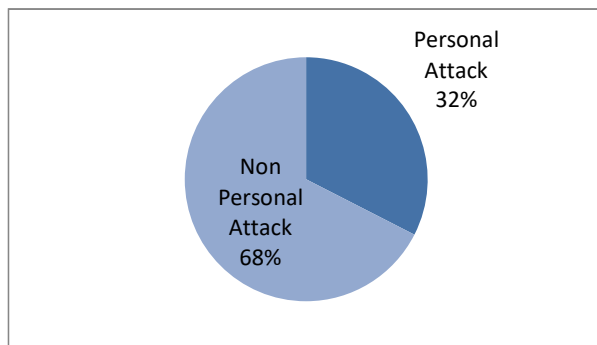


Fig. 4. Distribution of Comments for Wikipedia Dataset

VI. DATA PREPROCESSING

Fig. 5 shows a data processing pipeline used for both the datasets. First all text data are converted to lowercase. Then some words like “what’s” or “can’t” are converted to “what is” or “can not”. Also, all the punctuations are removed using the string library. Then following Natural Language Processing techniques are used using Natural Language Toolkit:

- **Tokenization:** In tokenization we split raw text into meaningful words or tokens. For example, the text “we will do it” can be tokenized into ‘we’, ‘will’, ‘do’, ‘it’. Tokenization can be done into words called word tokenization or sentences called sentence tokenization. Tokenization has many more variants but in the project we use Regex Tokenizer. In regex tokenizer tokens are decided based on rule which in the case is a regular expression. Tokens matching the following regular expression are chosen Eg For the regular expression ‘\w+’ all the alphanumeric tokens are extracted.
- **Stemming:** Stemming is the process of converting a word into a root word or stem. Eg for three words ‘eating’ ‘eats’ ‘eaten’ the stem is ‘eat’. Since all three branch words of root ‘eat’ represent the same thing it should be recognized as similar. NLTK offers 4 types of stemmers: Porter Stemmer, Lancaster Stemmer, Snowball Stemmer and Regexp Stemmer. The following project uses PorterStemmer.
- **Stop word Removal:** Stop words are words that do not add any meaning to a sentence eg. some stop words for english language are: what, is, at, a etc. These words are irrelevant and can be removed. NLTK contains a list of english stop words which can be used to filter out all the tweets. Stop words are often removed from the text data when we train deep learning and Machine learning models since the information they provide is irrelevant to the model and helps in improving performance.

VII. FEATURE EXTRACTION

Feature extraction is important for Natural Language Processing. Text data can not be classified by classifiers therefore they need to be converted to numerical data. Each document (tweet or comment in this case) can be written as a vector and those vectors can be used for classification. The following project studies three Feature extraction methods:

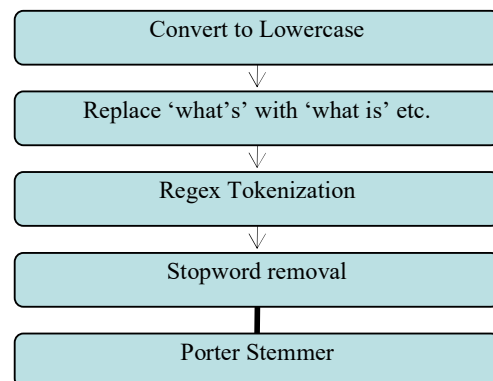


Fig. 5. Data Preprocessing Pipeline

Bag of Words, TF-IDF and Word2Vec.

A. Bag of Words model

The BoW that is bag of words model is a simple method of extracting features from documents that uses occurrence of words within a document. Bag of Words model has two important parts:

- A vocabulary of words(tokens) derived from all documents
- A way of measuring all these words as features in each document

It is referred to as ‘bag’ because the model only concerns with the word rather than its order of occurrence in the document. The intuition for this method is that similar documents have similar words in them.

The Bag of Words model uses the following procedure: A vocabulary is designed from all the documents. The vocabulary may consist of all words (tokens) in all documents or some top frequency tokens e.g. top 10 features with max occurrences in the corpus. Also features can be extracted for vocabulary in multiple forms based on number of words used per feature. e.g. for the sentence ‘This was the best ever’.

- **Unigram model** where single words are used eg ‘this’, ‘was’, ‘the’, ‘best’, ‘ever’ are the features for the corpus.
- **Bigram model** uses two words at a time for a feature e.g. ‘this was’, ‘was the’, ‘the best’, ‘best ever’ are features for the corpus.
- **N-gram model** is the generalised model where n can be 1,2, 3,... or even more than one value of N can be possible eg. extracting all unigram and bigram features

When the vocabulary is designed what is left is to transform all the documents based on the vocabulary using a way of measuring features. Generally, two types are used, first is a binary one where features are 1 or 0 depending on whether they exist in a document or not. But it does might not work on some sentences. e.g There is a difference between ‘very very good’ and ‘just good’. Therefore we can use the second method i.e frequencies of features in a

documents. Bag of words is a simple but quite effective method for sentiment analysis[14] but has certain limitations. It does not consider context or ordering of words which can make a lot of difference in some cases. Also, Vocabulary design becomes difficult in large datasets due to increase in number of features.

e.g 'Is it interesting' has a different meaning than 'It is interesting'.

B. TF-IDF Model

TF-Idf method is similar to the bag of words model since it uses the same way to create a vocabulary to get its features. TF-IDF addresses a problem not seen much in the corpus, but is important for better extraction of features. The value of Tf-Idf increases with the increase in frequency of a word in same document and decreases with decrease in frequency of documents that have the word in the corpus. It has two elements, which are

- Term frequency(Tf) is a calculation of frequency of a word in a document. It is measured as chance of finding a text word inside a document. It is measured as the frequency of a word W_i appearing in a document R_j , divided by total words in document R_j

$$tf(W_i, R_j) = \frac{\text{No. of times } W_i \text{ appears in } R_j}{\text{Total no. of documents in } R_j} \quad (1)$$

- Inverse document frequency (Idf) shows how frequent or rare a word is throughout the corpus. It is used to identify rare words in corpus. Idf value is higher for rarer words. IDF is getting by dividing the complete number of Words in document D in the corpus by the number of Words in files that consist the term t , and then calculating log value of resulting.

$$idf(d, D) = \log \frac{|D|}{\{d \in D : t \in D\}} \quad (2)$$

In above equation, $|D|$ denotes no of documents in the corpus and denominator term denotes number of documents which have the word t . Sometimes 1 is added to denominator to ensure there is no division by zero.

$$TfIdf(t, d, D) = tf(t, d) * idf(d, D) \quad (3)$$

The high TF-IDF means that word is frequent in a document but rare in the corpus making it more useful as a feature. A low or close to 0 TF-IDF means that these words almost occurs in all document making it less useful as a feature. TF-IDF solves some of the major issues in Bag of Words model thus making it more efficient.

C. Word2Vec

Word2Vec[15] is a Feature extraction method that uses word embeddings which was developed in 2013 by Google. It is used to represent word in vector form. This can be used to find similarity between words as two similar words have smaller angle between their vectors or cosine of angle between them is close to 1

$$sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \quad (4)$$

In (4) A and B are word vectors and θ is angle between both vectors. Word2Vec is a neural network method that uses this as an approach to train the model and construct word embeddings. There are two methods for the construction of the word embeddings:

- Common Bag of Words Model(CBOW): Common Bag of Words model takes as input of multiple words and predicts the word based on the context. Input can be one word or multiple words. A soft max is used at output. CBOW uses negative log likelihood and is more probabilistic rather than deterministic.
- Skip Gram Model: The skip gram model is just the reverse of CBOW model in which multiple context words are predicted using a single input word. Here the total number of words represented by X are predicted using the neural network. CBOW model takes a mean of context of input words but two semantics can be clicked for a single word. i.e. two vector of Apple can be predicted. First is for the firm Apple and next is Apple as a fruit.

Both of these methods use forward and back propagation to train the neural networks and find the best parameters. For each document then a feature vector can be created by concatenating and combining all word vectors in that document. Combination of word vectors can be done by summation or by averaging all word vectors. Selection between the both is based on data.

VIII. CLASSIFICATION

After getting feature vector for the training data by fitting it on the Feature extraction methods above, testing data is transformed using the same scheme without fitting it on the vectorizers or training it on the word2vec model. Using the training data following classifiers will be trained and tested on.

A. Support Vector Machine(SVM)

This theorem is basically used to plot a hyperplane that creates a boundary between data points in number of features (N)-dimensional space. To optimize the margin value hinge function is one of best loss function for this. Linear SVM is used in the following case which is optimum for linearly separable data. In case of 0 misclassification, i.e. the class of data point is accurately predicted by our model, we only have to change the gradient from the regularisation arguments.

In case of misclassification, i.e. our model makes a mistake in our data point's class prediction, we add the reduction with the gradient update regularisation.

B. Logistic Regression

It is a classification model and not a regression model. The probabilistic function used to model the output of problem is sigmoid function

$$sig(x) = \frac{1}{\{1 + \exp(-x)\}} \quad (5)$$

$$A=LT+C \quad (6)$$

$$T(x) = \text{sig}(A) \quad (7)$$

In (7) $T(x)$ is hypothesis function for our classifier, L is weights derived by classifier, C is bias derived by classifier and T is feature vector(input). If $h(x) > 0.5$ then class is 1 else class is 0. Since sigmoid lies between 1 and 0 it is ideal for classification.

C. Random Forest

A random forest consists of many individual decision trees which individually predict a class for given query points and the class with maximum votes is the final result. Decision Tree is a building block for random forest which provides a prediction by decision rules learned from feature vectors. An ensemble of these uncorrelated trees provide a more accurate decision for classification or regression.

D. Multi Layered Perceptron

Multi Layered Perceptrons are the Artificial Neural Networks containing at least 3 layers: one input, one output and at least one hidden layer. Each node has a activation value calculated using an activation function in a process called forward propagation and back propagation is used to train the weight used in the neural networks. It is generally used when data is linearly non seperable. Activation functions used can be relu or sigmoid. Sigmoid function is similar to the tanh function which is hyperbolic in nature between -1 and 1. Relu is defined as $f(x) = \max(0, x)$. Multi Layered Perceptrons can be created and trained using Keras Framework.

IX. EXPERIMENTS AND RESULTS

Google colab was used for the experiments. For each classifier the following parameters were evaluated on the test sets

- Accuracy(A): is defined as no of correct predictions divided by total number of predictions.

$$A = \frac{\text{True positives}}{\text{Size of dataset}} \quad (8)$$

- Precision(P): Out of all the positive predictions by the classifier how many are actually positive.

$$P = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \quad (9)$$

- Recall(R): Out of all the positive inputs how many were predicted positive

$$R = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \quad (10)$$

- F-measure(F): Calculates HM (harmonic mean) of precision and helps in comparison of both precision and recall.

$$F = \frac{2 \times P \times R}{P + R} \quad (11)$$

The following configurations are used for the feature selection methods for both datasets used:

- Bag of Words Model: Top 10000 features out of Unigrams, Bigram and Trigram features were selected based on frequency.
- TF-IDF Model: Same as Bag of words model
- For the word2vec model both the skips-gram and Common Bag of Words(CBOW) model were trained. 200 features from both models were combined to get 400 features for each word embedding and for each document summation was used to generate document vector. word2vec was trained on the training sets for 30 epochs with a window of 5 words.

The classifiers were loaded through sklearn library except the Multi Layer Perceptrons which were made in Keras. Two MLPs were used: one for Bag of Words and tfidf for 10000 feature input and other for 400 feature input of Word2vec. The architectures of both neural networks are shown in Fig. 6 and Fig. 7. The Classifiers used are Linear SVM (SVC), Random Forest Classifier (RF), Logistic Regression (LR) and Multi Layered Perceptron (MLP).

Tables 1 and 2 show results for Twitter and Wikipedia dataset respectively.

The Twitter dataset which contained tweets related to Hate speech show F- measures above 0.9 for all three feature selection methods. The values for Word2Vec model are a bit less but are ideal considering it used 400 features instead of other methods using 10000. TF-IDF method combined with Linear SVM gives best recall and F-measure.

For the Wikipedia dataset which contained comments with Personal attacks it shows F-measures only around 0.8 for all models. The TF-IDF with Linear SVM still get the best F-measure but Word2Vec with Multi Layered Perceptron gives better recall.

X. CONCLUSION

Cyber bullying across internet is dangerous and leads to mishappenings like suicides, depression etc and therefore there is a need to control its spread. Therefore cyber bullying detection is vital on social media platforms. With availability of more data and better classified user information

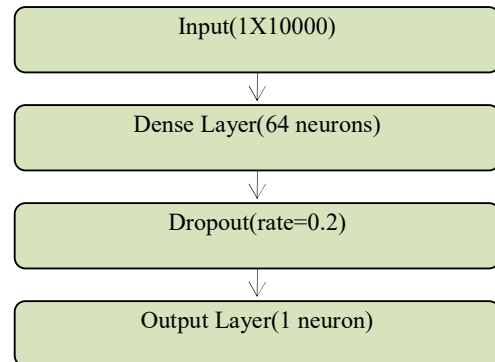


Fig. 6. Multi Layered Perceptron used for Bag of Words and TF-IDF inputs

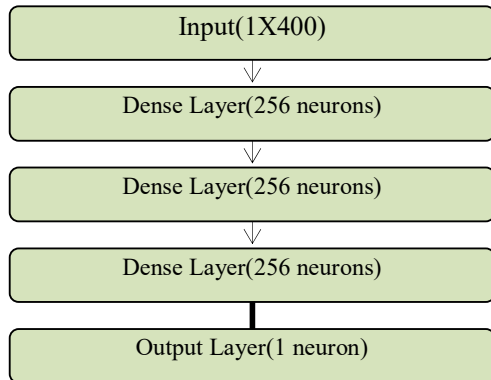


Fig. 7. Multi Layered Perceptron used for Word2Vec input

for various other forms of cyber attacks Cyberbullying detection can be used on social media websites to ban users trying to take part in such activity In this paper we proposed an architecture for detection of cyber bullying to combat the situation. We discussed the architecture for two types of data: Hate speech Data on Twitter and Personal attacks on Wikipedia. For Hate speech Natural Language Processing techniques proved effective with accuracies of over 90 percent using basic Machine learning algorithms because tweets containing Hate speech consisted of profanity which made it easily detectable. Due to this it gives better results with BoW and Tf-Idf models rather than Word2Vec models. However, Personal attacks were difficult to detect through the same model because the comments generally did not use any common sentiment that could be learned however the three feature selection methods performed similarly. Word2Vec models that use context of features proved effective in both datasets giving similar results in comparatively less features when combined with Multi Layered Perceptrons. As seen by changing nature of

REFERENCES

- [1] I. H. Ting, W. S. Liou, D. Liberona, S. L. Wang, and G. M. T. Bermudez, "Towards the detection of cyberbullying based on social network mining techniques," in *Proceedings of 4th International Conference on Behavioral, Economic, and Socio-*

- Cultural Computing, BESC 2017*, 2017, vol. 2018-January, doi: 10.1109/BESC.2017.8256403.
- [2] P. Galán-García, J. G. de la Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," 2014, doi: 10.1007/978-3-319-01854-6_43.
- [3] A. Mangaonkar, A. Hayrapetian, and R. Rajc, "Collaborative detection of cyberbullying behavior in Twitter data," 2015, doi: 10.1109/EIT.2015.7293405.
- [4] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," 2016, doi: 10.1145/2833312.2849567.
- [5] V. Banerjee, J. Telavane, P. Gaikwad, and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network," 2019, doi: 10.1109/ICACCS.2019.8728378.
- [6] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," 2011, doi: 10.1109/ICMLA.2011.152.
- [7] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying Detection using Pre-Trained BERT Model," 2020, doi: 10.1109/ICESC48915.2020.9155700.
- [8] M. Dadvar and K. Eckert, "Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study," *arXiv*. 2018.
- [9] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," *arXiv*. 2018.
- [10] Y. N. Silva, C. Rich, and D. Hall, "BullyBlocker: Towards the identification of cyberbullying in social networking sites," 2016, doi: 10.1109/ASONAM.2016.7752420.
- [11] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," 2016, doi: 10.18653/v1/n16-2013.
- [12] T. Davidson, D. Warmesley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," 2017.
- [13] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," 2017, doi: 10.1145/3038912.3052591.
- [14] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," *Artif. Intell. Rev.*, vol. 53, no. 6, 2020, doi: 10.1007/s10462-019-09794-5.
- [15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

TABLE 1: RESULTS FOR TWITTER DATASET

Measure	Bag of Words				TF-IDF				WORD2VEC			
	SVC	RF	LR	MLP	SVC	RF	LR	MLP	SVC	RF	LR	MLP
Accuracy	0.906	0.914	0.921	0.903	0.920	0.914	0.917	0.911	0.890	0.867	0.894	0.898
Precision	0.935	0.947	0.959	0.930	0.949	0.949	0.952	0.937	0.935	0.886	0.935	0.932
Recall	0.920	0.921	0.920	0.922	0.927	0.918	0.920	0.927	0.894	0.915	0.901	0.918
F-measure	0.928	0.934	0.939	0.927	0.939	0.933	0.936	0.932	0.914	0.901	0.918	0.922

TABLE 2: RESULTS FOR WIKIPEDIA DATASET

Measure	Bag of Words				TF-IDF				WORD2VEC			
	SVC	RF	LR	MLP	SVC	RF	LR	MLP	SVC	RF	LR	MLP
Accuracy	0.871	0.886	0.890	0.876	0.894	0.887	0.892	0.869	0.876	0.857	0.879	0.879
Precision	0.817	0.892	0.879	0.828	0.879	0.906	0.915	0.814	0.833	0.828	0.832	0.814
Recall	0.798	0.755	0.785	0.803	0.798	0.745	0.752	0.795	0.795	0.731	0.808	0.835
F-measure	0.808	0.818	0.829	0.815	0.837	0.818	0.825	0.805	0.813	0.776	0.820	0.825

Detection of Malicious Transactions using Frequent Closed Sequential Pattern Mining and Modified Particle Swarm Optimization Clustering

Rajni Jindal

Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
rajnijindal@dce.ac.in

Indu Singh

Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
indusingh@dtu.ac.in

Abstract— In current times, with data security being recognised as an irrefutable requirement within an organisation, the importance of institution of intrusion detection system has grown manifolds. Identification of outsider attacks as well as misuse of database privileges by authorised entity has been a primary requirement in modern intrusion detection systems. In this paper we present BIDE (BI-Directional Extension) an efficient algorithm for mining frequent closed sequences without candidate maintenance and modified Particle Swarm Optimization clustering based malicious query detection (BPSOMQD), an advanced approach that detects and prevents malicious transactions from disrupting the consistency of the database. This method incorporates frequent closed sequential pattern mining which forms the basis for generation of data dependency rules. Further to recognise anomalous user activity, modified Particle Swarm Optimization algorithm is proposed which is used to generate role profiles associated with the transaction. A combination of Multilevel Rule Similarity Score (MRSS) between data dependency rules with incoming transaction and Cluster Similarity Index (CSI) with generated role profiles is considered to categorise the transaction as malicious or non-malicious. Experimental evaluation of proposed approach shows remarkable results on a characteristic banking database with accuracies over 92.37%.

Keywords—Database security, Database intrusion detection, Sequential pattern mining, Role Profiles, Particle Swarm Optimization clustering

I. INTRODUCTION

With growing dependence of organisations on database systems for management of information, there is huge level of risk associated in cases of security breaches leading to database being compromised. It is estimated that volumes of business data worldwide across all companies, doubles every 1.2 years [1]. Hence Intrusion Detection Systems (IDS) is supposed to meet the three basic criteria of data security which are Confidentiality, Integrity and Availability collectively known as CIA triad [2]. In order to mitigate potential threats to databases, intrusion detection systems employed must be capable to handle both internal and external attacks.

Primarily IDS are classified into two types — signature based and anomaly based. Signature based intrusion detection operates by comparing incoming transaction with known patterns of attacks [3] and thus offers limited functionality and

poor performance over novel attacks on the database. Furthermore, the need to continuously update the signatures base makes it a less favourable choice among organisations. Thus to identify unprecedented attacks, anomaly based IDS [4] is used, which models legitimate user transactions patterns and measures deviation of incoming transaction with those patterns to classify it as normal or malicious.

In this paper, our objective is to detect and prevent all types of malicious attacks on the database. A majority of intrusion detection systems can identify only external attacks, while in this paper our approach BPSOMQD(BIDE and Particle Swarm Optimization clustering based malicious query detection), combining association rules and cluster analysis, manages to prevent privilege abuse along with external attacks. In this approach, data dependency rules are generated by using BIDE algorithm to mine user transactions database containing legitimate user transactions. BIDE algorithm mines efficiently the complete set of frequent closed sequences. In BIDE algorithm [10], we do not need to keep track of any single frequent closed sequence (or candidate) for a new pattern's closure checking, which leads to the proposal of a deep search space pruning method. Cluster analysis is done to obtain role profiles by means of modified Particle Swarm Optimization algorithm to detect anomalous user patterns. All incoming transactions are validated through this detection mechanism and then an alarm is generated if the transaction is found to be malicious, terminating the transaction and hence preventing any sensitive information residing in the database. The main objectives of BPSOMQD are:

1. To develop a novel technique which incorporates legitimate user access sequences to determine permissibility of incoming transaction.
2. To propose a system architecture which prevents unauthorised access, privilege abuse and modification of sensitive attributes.
3. To combine the features of signature and anomaly-based IDS and achieve optimal results for all types of attacks and outperform traditional IDS.

The rest of the paper is organised as follows. Section 2 describes related work in the domain of intrusion detection system and database security. In Section 3, the architecture and

algorithm of proposed approach is illustrated. Section 4 presents the experimental results of our approach. Lastly Section 5 draws the conclusion of the paper.

II. RELATED WORK

The domain of intrusion detection system is widespread with many researchers actively working on development of Network Intrusion Detection Systems but only a limited number of significant researches are carried out in case of Database Intrusion Detection Systems.

Hu et al. [5] devised a sequential pattern mining technique which classifies a transaction to be malicious if it fails to comply with mined data dependencies. It helps to identify those group of malicious transactions which adhere to user behaviour individually. But the major limitation of this approach was that it didn't take attribute sensitivity into consideration. To overcome this limitation, Srivastava et al. [6] proposed an approach which assigned weights to all operations on data attributes and transactions not following the data dependency rules were classified as malicious. But the major disadvantage of this method was the lack of well-defined procedure for assigning weights to the attributes.

Doroudian et al. [7] devised a hybrid approach for identifying malicious transaction at both transaction and inter-transaction level. At transaction level, it uses a set of predefined expected transactions whereas at inter-transaction level, sequential rule mining algorithm was used to identify dependencies among transactions. Sohrabi et al. [8] proposed a novel approach ODADRM which extracted rules for infrequent transactions as well by using leverage as rule value measure to minimise uninteresting data dependencies. Ranao et. al [9] presented a Query Access detection technique using PCA and Random Forest algorithm to perform dimensionality reduction and obtain relevant data. But increase in system performance attributed to reduced dimensionality comes at a cost of increased True Positive rate.

Mohammad Raza et. al [17], used anomaly detection concept to present a novel approach for distinguishing between regular and irregular activities on databases. They proposed a new density-based clustering method called Clustering-based Intrusion Detection (CID) which clustered queries based on the similarity measure and labeled them. As per claimed by the research, there is no study other than CID in the domain of database IDS which considers string based query distance metric. Similarly, in the research work of Asmaa Sallam et. al [20], anomaly detection techniques were used for Database access monitoring by detecting the misuse scenarios. But this paper's techniques captures the normal data access rates from past queries of user activity during a training phase to build profiles that describe the data access patterns of the Database users. An increase in a user's data access rates beyond the normal levels is flagged as anomalous to indicate that the behavior of the user is suspicious and requires further analysis.

III. OUR APPROACH

In this paper, we propose a two-phase intrusion detection system approach that incorporates association rule mining and role profile clustering in sequential phases in the training phase to generate association rules and role profiles respectively. The

main objective behind this approach is to use these two techniques in conjunction to effectively capture the intrusive activities and produce noteworthy results by combining the effectiveness of signature as well as anomaly-based intrusion detection system.

The two phases incorporated in this approach are the learning phase and detection phase. The learning phase commences with extracting transactions from transaction logs and preprocessing them into desired format to apply frequent closed sequential pattern mining algorithm and build association rules for detection phase using BIDE algorithm [10]. BIDE is an efficient algorithm for mining frequent closed sequences. It triumphs the problem of the candidate maintenance-and-test paradigm, prunes the search space more deeply and checks the pattern closure in a more efficient way while consuming much less memory in contrast to the previously developed closed pattern mining algorithms. Association rules satisfying minimum length and confidence values are stored to be later used to detect whether transactions are malicious. In second part of learning phase, database logs are used to generate role profiles pertaining to existing users of the organisations using modified Particle Swarm Optimization algorithm. The PSO algorithm can produce high-quality solutions within shorter calculation time and more stable convergence characteristics than other stochastic methods [21]. It is assumed that only legitimate transaction logs and database logs are provided in this phase.

During the detection phase, all incoming transactions are processed and compared with the rules generated in the learning phase to calculate the Multilevel Rule Similarity Score of the transaction with existing rules. This Multilevel Rule Similarity Score along with Cluster Similarity Index to role profiles is used to conclude whether the transaction is malicious.

A. System Architecture

The main idea behind this approach is to develop an intrusion detection system which is decoupled from the existing database management system so as to function independently in various environments without affecting the internal structure of the system. It aims to identify intrusion attacks and prevent execution of malicious transactions which could compromise the integrity and security of database. The overall architecture of the proposed approach is divided into two phases, the Learning Phase and the Detection Phase.

B. Learning Phase

The overall learning phase of the proposed approach is illustrated in Figure 1. In this phase, transaction logs are utilized to extract data dependency rules which are used to calculate Multilevel Rule Similarity Score in detection phase. Apart from this, Database logs are used to capture role profiles of existing authorised users capable of performing transactions in order to identify deviation from normal user behaviour. Data dependency rules and role profiles are collectively used in detection phase to classify whether the incoming transaction is malicious.

1) Transaction Preprocessor

Transaction preprocessing is an important part of learning phase which transforms all the transactions stored in the audit logs into suitable format which can later be fed into subsequent

modules of learning phase for generating association rules and role profiles.

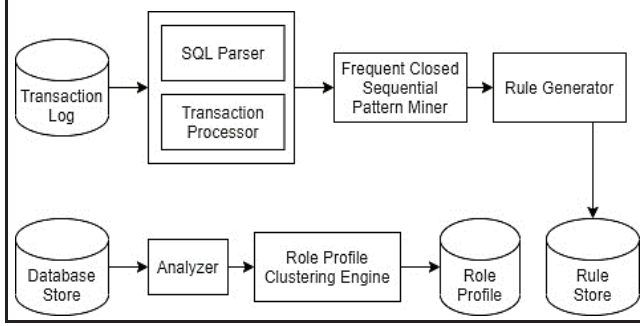


Fig. 1. Learning Phase architecture of proposed approach

Definition 1. Transaction (T): A transaction T is considered as an atomic unit to be executed by the user which comprises of various SQL queries (Select, Insert, Update, Alter, Delete). Each transaction is assigned a unique transaction ID (T_ID) and consists of various queries.

Definition 2. Query (Q): - Each Query Q is interpreted as a sequence of read and write operations on item sets present in the database.

Definition 3. Operation (Op(x)): - An operation is essentially an execution of read or write action on item set x.

Definition 4. Sequence (Seq): - A sequence is defined as a list of read and write operations enlisted in temporal order of execution.

Definition 5. Frequent Sequential Pattern (FSP): - A pattern which appears in at least min sup number of sequences present in the database is called as frequent sequential pattern.

Definition 6. Frequent Closed Sequential Pattern (FCSP): - A frequent sequential pattern is which is not included in another sequential pattern having exactly same support is called as frequent closed sequential pattern.

1) Frequent Closed Sequential Rule Miner

In this module, we have used BIDE algorithm [10] to mine frequent closed sequential patterns from existing transaction logs to generate data dependency rules.

Table 1 describes a sample set of sequences to understand the functioning of frequent closed sequential rule miner module.

TABLE I. SEQUENCES FOR MINING SEQUENTIAL PATTERNS

TID	Sequence
1	R12, W00, W20, W22, W21, R12, R23 W13, W15
2	R03, W14, W25, W24, W14, R03, R04, R04
3	R03, W11, W25, R14, W14, W04, R04
4	R12, W20, W21, R10, R23, W21, W15, R02

5	R12, R24, W20, W14, W21, W01, R23, W12, W15, R02
6	R03, W25, W14, R04, R26
7	R12, W04, W20, W22, W21, W25, R23, W20, W15, R02
8	R03, R20, W25, W14, R26, R04
9	R03, R22, W25, R03, W14, R04, R04
10	R03, R23, W25, W20, W14, W24, R04, R04

Post the generation of frequent closed sequential patterns, those sequences are transformed into read and write rules among which those satisfying the minimum length, minimum support and minimum confidence criteria are added to the final rules set.

TABLE II. MINED DATA DEPENDENCY RULES

Association Rule	Support	Confidence
{R03, W25, W14} → {R04}	0.6	1.000000
{R03, W25, R04} → {W14}	0.6	1.000000
{R03, W14, R04} → {W25}	0.6	1.000000
{R04, W25, W14} → {R03}	0.6	1.000000
{R03, W25} → {R04, W14}	0.6	1.000000
{R03, W14} → {R04, W25}	0.6	1.000000
{R03, R04} → {W25, W14}	0.6	1.000000
{W25, W14} → {R03, R04}	0.6	1.000000
{R04, W25} → {R03, W14}	0.6	1.000000
{R04, W14} → {R03, W25}	0.6	1.000000
{R03} → {R04, W25, W14}	0.6	1.000000
{W25} → {R03, W14, R04}	0.6	0.857143
{W14} → {R03, W25, R04}	0.6	0.857143
{R04} → {R03, W25, W14}	0.6	1.000000

Table 2 displays the association rules generated from sequences listed Table 1.

Algorithm 1 generates data dependency rules using frequent closed sequential patterns obtained from BIDE algorithm. These sets of data dependency rules are later used in detection phase for identification of malicious transactions. Step 1-2 initializes sequential pattern base and rules set to be used. In step 3-5 a subroutine BIDE Frequent Closed Sequen-

tial Pattern Generator (BFCSPG) is called for all the sequences in the database. From steps 6-14, if a sequence is closed in the initial database, satisfies minimum length criteria as well as has a length of its projected database satisfying the minimum support criteria, then it is added to the sequential pattern base. In steps 15-22, the evaluating sequence is projected using the projected database of that sequence to ensure whether a longer super sequence of that sequence is closed in the database. If such a sequence is found, then the current sequence is removed from sequential pattern base and the super sequence is added to it. This method ensures that only frequent closed sequential patterns exists in the database which reduces the number of association rules generated.

Later in steps 23-29, the frequent closed sequential patterns are used to generate the data dependency rules which satisfies the minimum confidence criteria along with minimum support and minimum length criteria satisfied by the patterns. If all the criteria are satisfied, then such a rule is added to the final rule set.

2) Modified PSO Clustering Algorithm

Particle swarm optimization (PSO) is a population-based stochastic search process, modeled after the social behavior of a bird flock [19]. Particle Swarm Optimization is a widely used evolutionary algorithm for performing clustering of data [11]. So, in our approach, we have incorporated a variation of traditional Particle Swarm Optimization algorithm to cluster the data. In this variation, the algorithm primarily focuses on obtaining global best score as the clustering process commence while during termination stages the focuses shifts on the personal best scores. The objective behind this variation is to quickly allow the algorithm to converge by searching for global best score while relying on the personal best scores during the last few iterations to search for solution near the personal best centroids.

Algorithm 1 : Frequent Closed Sequential Rule Miner

Data: DB: Initial Database
 min_sup : minimum support
 min_conf : minimum confidence
 min_len : minimum length of sequential pattern

Result: R : Rule Set

Initialization

1. Sequential Pattern Base : $\pi \leftarrow \Phi$
2. Rule Set : $R \leftarrow \Phi$

Extract Frequent Sequential Patterns

3. **for** sequences in DB from $i \leftarrow 1$ to n **do**
4. BFCSPG ($\langle \rangle, (i, -1)$)
5. **end for**

Procedure BFCSPG ($f, DB|_f$)

6. $sup = len(DB|_f)$
7. **if** length (f) $\geq min_len$ **then**
8. **if** $sup < min_sup$ **then**
9. **return**
10. **end if**
11. **if** isClosed (DB, $f, DB|_f$) **then**
12. $\pi = \pi \cup (f, sup)$

```

13.   end if
14. end if
15. for next_item, rem_DB in nextEntries(DB, DB|_f) do
16.   new_f = f  $\cup$  next_item
17.   if len(DB|_f) == len(rem_DB) and (f, sup) in  $\pi$ 
18.     then
19.        $\pi = \pi - (f, sup)$ 
20.       if cannot_prune(DB, new_f, rem_DB) then
21.         BFCSPG (new_f, rem_DB)
22.       end if
23.     end if
24.   end for

```

Extract Dependency Rules

```

23. for frequent sequential pattern fsp in  $\pi$  do
24.   for each rule  $\in$  Rule_Generator(fsp)
25.     if (confidence(rule)  $\geq min\_conf$ )
26.        $R = R \cup rule$ 
27.     end if
28.   end for
29. end for

```

Algorithm 2 highlights the modified PSO algorithm developed in this study. Step 1-8 initializes various parameters used in modified PSO algorithm. According to line 9, the complete algorithm is run for maximum specified number of transactions. From step 10-20, the fitness of each particle is evaluated and if the fitness is greater than personal best score of current particle then the personal best score and position is updated. The particle position with maximum personal best score is then selected as global best position and global best score is set to be the maximum personal best score. This update in personal best and global best values continues in each iteration. In step 21-29, the velocity and position of each particle is updated according to the steps 25 and 27. After completion of maximum number of iterations, the final cluster centroids G_pos represent the role profiles corresponding to the given database of transactions which are later used in Detection phase to measure similarity of incoming transaction with existing role profiles of database logs.

Algorithm 2 : Modified PSO Clustering

Data: X : database logs , K : number of clusters

Result: C_k : Cluster centroids = G_pos

Initialization

1. $N \leftarrow$ number of particles
2. $P \leftarrow$ list of personal best scores initialized -inf
3. $G \leftarrow$ global best score initialized as -inf
4. $T \leftarrow$ maximum number of iterations
5. $P_pos \leftarrow$ list of personal best positions initialized as Φ
6. $G_pos \leftarrow$ global best position initialized as Φ
7. $x \leftarrow$ randomly initialized list of particles
8. $v \leftarrow$ randomly initialized velocity of particles in every dimension

Learning Phase

```

9.  for t ← 1 to T
10.  for i ← 1 to N
11.    f = fitness_function(xi)
12.    if f > Pi then
13.      Pi = f
14.      P_posi = xi
15.    end if
16.    if Pi > Gi then
17.      G = Pi
18.      G_pos = P_posi
19.    end if
20.  end for
21.  for i ← 1 to N
22.    for k ← 1 to K
23.      G_inc = c * (1 -  $\frac{t}{T}$ ) * rand1 * (xi - G_pos)
24.      P_inc = c * ( $\frac{t}{T}$ ) * rand1 * (xi - P_posi)
25.      vik = w * vik + G_inc + P_inc
26.    end for
27.    xi = xi + vi
28.  end for
29. end for

```

C. Detection Phase

Detection phase utilizes the association rules and role profiles generated in the Learning Phase to identify malicious transactions and prevent such transactions from executing in the database system environment. In this phase, each query of incoming transaction is parsed, pre-processed and transformed to generated sequence of operations. These sequences of operations are then equated against rules generated in previous phase to obtain a Multilevel Rule Similarity Score. Along with that, the current user profile is compared with existing role profiles to calculate a Cluster Similarity Index. Finally, a combination of Multilevel Rule Similarity Score and Cluster Similarity Index is used to classify the transaction as malicious or not. If the transaction is found to be malicious, then an alarm is raised, transaction is aborted and rolled back to restore the original state of database before commencement of the transaction. The complete detection phase architecture of the proposed approach is described in the Figure 2.

1) Rule Matcher

In the first step of detection phase, the read/write sequences from the queries of incoming transaction is compared with data dependency rules generated from the frequent closed sequential pattern miner of the learning phase.

Definition 7. Multilevel Rule Similarity Score (MRSS): - Multilevel Rule Similarity Score for a given sequence measures the degree of similarity between the sequence from query and the data dependency rules generated from transaction logs. It is the maximum similarity score obtained from the sequence against all data dependency rules by using a weighted combination of 4 individual similarity measure namely SS1, SS2, SS3, SS4.

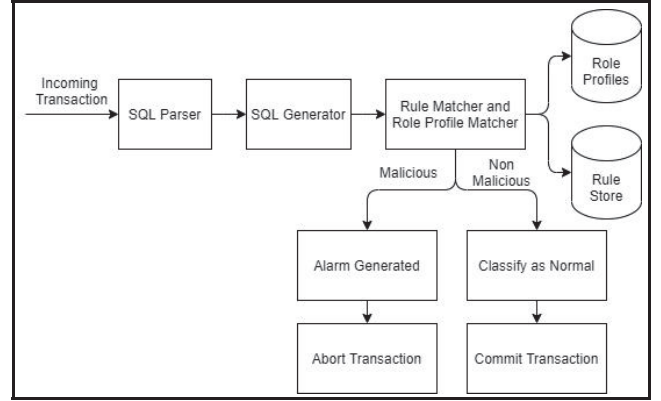


Fig. 2. Detection Phase architecture of proposed approach

Where w_1, w_2, w_3, w_4 are user defined similarity level weights.

$$MRSS = \frac{w_1 * SS1 + w_2 * SS2 + w_3 * SS3 + w_4 * SS4}{w_1 + w_2 + w_3 + w_4}$$

Similarity Score 1 (SS1): It is the ratio of number of common attributes present in data dependency rule and the sequence to that of attributes present in the sequence.

$$SS1 = \frac{\text{Count of common attributes}}{\text{Total number of attributes in sequence}}$$

Similarity Score 2 (SS2): It incorporates edit distance between rule and sequence into account and is defined as

$$SS2 = 1 - \frac{\text{edit distance}}{\max(\text{len}(\text{rule}), \text{len}(\text{sequence}))}$$

where edit distance is defined as the minimum number of insertion/deletion/updating operations required to transform sequence into rule.

Similarity Score 3 (SS3): It incorporates the length of longest common subsequence between rule and sequence into account and is defined as

$$SS3 = \frac{\text{Length of the longest common subsequence}}{\text{Total number of attributes in sequence}}$$

Similarity Score 4 (SS4): It is the ratio of sum of common attributes count present in data dependency rule and the sequence to product of L2 norm of vectors of data dependency rule and sequence.

$$SS4 = \frac{\text{Sum of count of common attributes}}{||\text{Rule}|| \times ||\text{Sequence}||}$$

where $||x||$ is the L2 norm of x .

2) Classification Algorithm

Definition 8. Cluster Similarity Index (CSI):-

Let P be the list of membership probabilities of given transaction X from all the K clusters. Cluster Similarity Index (CSI) is

calculated as the maximum membership probability of given transaction from all the K clusters.

$$CSI = \max(P_k) \forall k \in [1, K]$$

In the second step of detection phase Multilevel Rule Similarity Score of previous step is combined with Cluster Similarity Index to give the overall prediction whether a transaction is malicious.

Algorithm 3 describes the overall procedure of detection phase. In step 1, we use each of the queries present in the given transaction to determine whether the transaction is malicious. Step 2-3 generates parse tree which is later used to produce sequence.

From Step 5-11, the generated sequence is compared with all the data dependency rules generated to compute maximum MRSS using SS1, SS2, SS3 and SS4. From Step 12-14, membership of sequence is checked with role profile clusters generated using modified PSO clustering algorithm.

In step 15-16 Cluster Similarity Index is calculated and combined with MRSS to classify the incoming transaction. If the overall score exceeds the dissimilarity threshold, then the transaction is classified as malicious, otherwise non malicious, otherwise non-malicious as described in Step 17-21. Finally, in Step 22-26 we take action according to the class of transaction. In case of malicious transaction, it is rolled back and the alarm is raised otherwise it is committed successfully in the database.

Algorithm 3 : Detection Phase Algorithm

Data: R : Rule Set

K : number of clusters

C : cluster centroids

δ : dissimilarity threshold

T : transaction

w1,w2,w3,w4 : similarity level weights

Result: C : Class (malicious or non-malicious)

During Transaction

```

1. for each of the Query q in TRN do
2.   parseTree ← SQLParser(q)
3.   Seq ← SequenceGenerator(parseTree)
4.   MRSS = 0
5.   for each rule in R do
6.     SS1 = similarity_score1(rule, Seq)
7.     SS2 = similarity_score2(rule, Seq)
8.     SS3 = similarity_score3(rule, Seq)
9.     SS4 = similarity_score4(rule, Seq)
10.    MRSS = max(0,  $\frac{w1 \cdot SS1 + w2 \cdot SS2 + w3 \cdot SS3 + w4 \cdot SS4}{w1 + w2 + w3 + w4}$ )
11.  end for
12.  for k ← 1 to K
13.    Pk = membership(Ck, Seq)
14.  end for
15.  CSI = max(Pk)
16.  C_score =  $\frac{2 \times MRSS \times CSI}{MRSS + CSI}$ 
17.  if C_score >  $\delta$  then
18.    C ← malicious
19.  else
20.    C ← non-malicious
21.  end if

```

```

22. if (C == malicious) then
23.   Rollback – Raise Alarm
24. else
25.   commit
26. end if
27. end for

```

IV. RESULTS

To evaluate the performance of our proposed approach, various experiments were carried out on a conventional banking dataset adhering to TPC - C benchmark [16]. The entire dataset of transactions is broadly divided into two sets of transactions — one comprising of malicious transactions carried out by unauthorised users while other being normal transactions complying with data dependency rules and performed by authorised users. In total 25000 transactions were generated combining both the sets and used for evaluating the performance of the proposed approach.

To measure the performance of proposed approach, the three metrics that were taken are precision, recall and F1-score. Here precision is defined as the ratio of correctly identified malicious transactions known as True Positives (TP) to the total number of transactions classified as malicious. While recall is calculated as ratio of correctly identified malicious transactions to total number of malicious transactions present in the database logs. Precision and Recall are inversely related to each other where improvement in one is generally achieved at the cost of the other. To balance this trade-off and combine both precision and recall into a single metric, F1-score is used. F1-score is defined as harmonic mean of precision and recall which takes both these metrics into account for performance evaluation.

TABLE III. COMPARISON OF PRECISION, RECALL, F1-SCORE AND ACCURACY WITH NO. OF TRANSACTIONS

Transaction Count	Precision	Recall	F1 - score	Accuracy
1000	0.892	0.892	0.892	0.978
2000	0.877	0.907	0.892	0.977
3000	0.898	0.927	0.912	0.982
4000	0.908	0.930	0.919	0.983
5000	0.910	0.938	0.924	0.985

The comparison of Precision, Recall and F1-score with number of transactions is depicted in Figure 3 and Table 3 to highlight the effect of number of transactions on various evaluation metrics. From the graph it can be inferred that as a general trend the results of evaluation metrics increase as the number of transactions increases which is accomplished due to increase in number of association rules and well-defined user profiles obtained using modified PSO clustering algorithm. It can be seen that the value of precision range goes from 0.892 to 0.91 as the number of transactions increase, while that of recall ranges from 0.892 to 0.938. Since F1-score is harmonic

mean of these two, its value will always lie in between the two and ranges from 0.892 to 0.924.

Figure 4 and Table 3 depict the variation in accuracy with increasing number of transactions in the database. It can be observed that accuracy increases from 0.978 to 0.985 as the number of transactions increase from 5000 to 25000. With increased number of transactions, the ability of the model to correctly distinguish both malicious and normal transactions results accounts for increased accuracy.

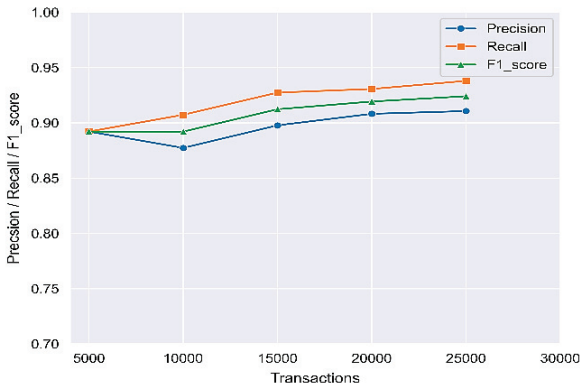


Fig. 3. Variation of Precision, Recall, F1-Score with No. of Transactions

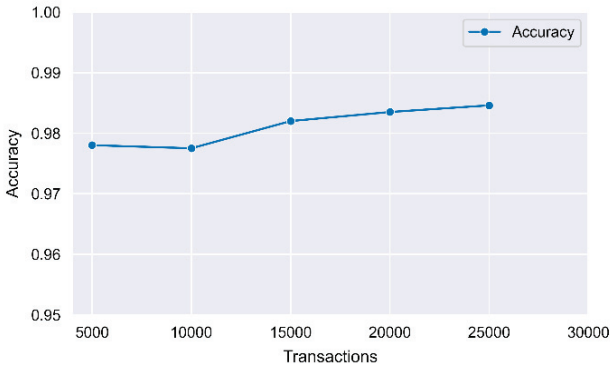


Fig. 4. Variation of Accuracy with No. of Transactions

TABLE IV. PERFORMANCE COMPARISON OF THE PROPOSED SYSTEM

Reference	Technique	Command Syntax	Anomaly Detection	Intrusion Prevention Capabilities	Performance
Hu et al. [5] 2004	Integrated dependency with sequence alignment analysis	✓	✗	Partial	Recall = 0.90 Precision = 0.64 accuracy=74.80%
Srivastava et al. [6] 2008		✓	✗	Partial	Recall = 0.77 Precision = 0.80 accuracy=78.47%
Hashemi et al. [18] 2008	Temporal Mining	✓	✗	Yes	Recall = 0.90 Precision = 0.75 accuracy=81.81%
Doroudian et al. [7] 2014	Mining Dependencies	✓	✗	Yes	Recall = 0.89 Precision = 0.91 accuracy=89.98%
Asmaa Sallam et al[20] 2019	AD technique	✓	✓	Yes	Recall = 0.88 Precision = 0.96 accuracy=91.90%

Keyvanpour et al[17] 2020	Clustering based intrusion detection	✓	✓	Yes	Recall = 0.95 Precision = 0.87 Accuracy=90.80%
Our Approach	BPSOMQD	✓	✓	Yes	Recall =0.910 Precision=0.938 accuracy=92.37%

Table 4 provides a detailed comparison of our approach with various state of art approaches based on evaluation metrics like precision, recall and F1-measure (accuracy). F1-score is the harmonic mean of the precision and recall.

$$F_1 - Measure = \frac{2 \times recall \times precision}{recall + precision}$$

This table highlights that our approach clearly outperformed the existing approaches in terms of performance which can be attributed to the large number of association rules as well as high quality clusters generated representing user profiles using modified PSO clustering algorithm.

V. CONCLUSION AND FUTURE WORK

This paper presents a novel intrusion detection and prevention technique that accommodates the behaviour of user at individual as well role level. Along with that, this approach relies on extraction of data dependency rules from transaction database and combines it with user profiles generated by means of modified PSO clustering algorithm to classify the transaction as malicious or non-malicious. Combining both dependency rules as well as role profiles helps to identify both previously seen attacks as well as novel attacks that deviate from authorized user profile. Our future prospects will focus on a further detailed and comprehensive formulation of sensitivity of operations, which will intern enhance the performance and efficiency of the system.

REFERENCES

- [1] A.J. Fernández-García, L. Iribarne, A. Coral, J. Criado and J.Z. Wang, 2018. A flexible data acquisition system for storing the interactions on mashup user interfaces. *Computer Standards & Interfaces*, 59, pp.10-34.
- [2] E. Bertino and R. Sandhu, 2005. Database security-concepts, approaches, and challenges. *IEEE Transactions on Dependable and secure computing*, 2(1), pp.2-19.
- [3] D.E. Denning, 1987. An intrusion-detection model. *IEEE Transactions on software engineering*, (2), pp.222-232.
- [4] H. S. Vaccaro and G. E. Liepins, "Detection of anomalous computer session activity," *Proceedings. 1989 IEEE Symposium on Security and Privacy*, Oakland, CA, USA, 1989, pp. 280-289, doi: 10.1109/SECPRI.1989.36302.
- [5] Y. Hu and B. Panda, "A data mining approach for database intrusion detection," in *Proceedings of the ACM Symposium on Applied Computing*, pp. 711-716, 2004.
- [6] A. Srivastav, S. Sural, A. K. Majumdar, "Weighted intratransactional rule mining for database intrusion detection", *Proceedings of the 10th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining*, Singapore, 2006.
- [7] M. Doroudian and H.R. Shahriari, "A Hybrid Approach for Database Intrusion Detection at Transaction and Inter-Transaction Levels", *6th Conference on Information and Knowledge Technology (IKT)*, pp. 1-6, 2014.
- [8] M. Sohrabi, M. M. Javidi, S. Hashemi, "Detecting intrusion transactions in database systems: a novel approach", *Journal of Intelligent Info Systems* 42:619-644 DOI 10.1007 Springer 2014.

- [9] C. A. Ranao and S. Chao, "Anomalous query access detection in RBAC-administered databases with random forest and PCA", *Journal Information Sciences*, Volume 369, Issue C, Pages 238-250, 2016.
- [10] J. Wang and J. Han, 2004, April. BIDE: Efficient mining of frequent closed sequences. In *Proceedings. 20th international conference on data engineering* (pp. 79-90). IEEE.
- [11] D. V. Merwe, A. P. Engelbrecht, Data clustering using particle swarm optimization, in: *The 2003 Congress on Evolutionary Computation*, 2003. CEC'03., Vol. 1, IEEE, 2003, pp. 215–220.
- [12] A. Kamra, E. Terzi and E. Bertino, 2008. Detecting anomalous access patterns in relational databases. *The VLDB Journal*, 17(5), pp.1063-1077.
- [13] S. Panigrahi, S. Sural and A.K. Majumdar, 2013. Two-stage database intrusion detection by combining multiple evidence and belief update. *Information Systems Frontiers*, 15(1), pp.35-53.
- [14] V. C. S. Lee, J. A. Stankovic, and S.H. Son, "Intrusion Detection in Real-time Database Systems Via Time Signatures", in *Proceedings of the 6th IEEE Real Time Technology and Application Symposium (RTAS)*, pp. 124-133, 2000.
- [15] X. Gao, C. Shan, C. Hu, Z. Niu and Z. Liu, 2019. An adaptive ensemble machine learning model for intrusion detection. *IEEE Access*, 7, pp.82512-82521.
- [16] TPC-C benchmark: <http://www.tpc.org/tpccdefault.asp>
- [17] M.R Keyvanpour, M.B Shirzad and S. Mehmandoost, "CID: a novel clustering-based database intrusion detection algorithm". *Journal of Ambient Intelligence and Humanized Computing*, pp.1-12, 2020.
- [18] Hashemi, S., Yang, Y., Zabihzadeh, D. and Kangavari, M., 2008. Detecting intrusion transactions in databases using data item dependencies and anomaly analysis. *Expert Systems*, 25(5), pp.460-473.
- [19] J. Kennedy and R. Eberhart, "Particle swarm optimization," *Proceedings of ICNN'95 - International Conference on Neural Networks*, Perth, WA, Australia, 1995, pp. 1942-1948
- [20] A.Sallam, and E. Bertino, 2019, Result-based detection of insider threats to relational databases. In *Proceedings of the Ninth ACM Conference on Data and Application Security and Privacy* (pp. 133-143).
- [21] T.Niknam and B.Amiri, "An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis" *Applied Soft computing*, 2010, pp. 183-197.



DETERMINANTS OF JOB OPPORTUNITIES IN SKILL DEVELOPMENT INSTITUTIONS: INDIAN PERSPECTIVE

Manoj Kumar, Suresh Kumar Garg and Shraddha Mishra

In a fast-growing economy like India and having a comparatively young population, education, especially skill-based education, plays an important role. In the last two decades' emphasis has been placed on this through education policy interventions at all levels of governance. Despite this, the impact is not sufficient. There is a need for the industry to associate with skill development institutions for need-based effective management programmes. This paper attempts to study the students' perceptions of the skill development programmes and their efficiency in providing better job opportunities. The study shows a significant role of trainers, industry connections and institutes infrastructure in giving better jobs opportunities to the trainees of skill development institutes.

KEYWORDS: Skill Development, Industry Support, Institution Management

INTRODUCTION

Growth of India will pick up the pace in economic sense only if the adolescence of our country will get vocational education and acquire relevant skills. The impending need of employability skills have been advocated disputed for increasing work outcomes and helping citizens in adapting with changes and improving upon their career opportunities in the workplace (Yusof, Mustapha,

Manoj Kumar

Research Scholar, Delhi School of Management, Delhi Technological University, Delhi, India.

Email: manoj1960@gmail.com

Suresh Kumar Garg

Professor, Delhi School of Management, Delhi Technological University, Delhi, India.

Email: skgarg63@yahoo.co.in

Shraddha Mishra ✉

Assistant Professor, IILM University, Gurugram, India.

Email: shraddhamishra29@gmail.com



This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Mohamad, & Bunian, 2012). In Indian economy the literacy level is increasing day by day, however, the increase in unemployable literate is as an area of concern for the country. Ministry of education has made a lot of efforts to turn down the situation and encouraged vocational education in order to increase skills among the youth. Indian educational institutions have been facing several challenges in getting world-class status or global rankings (Banker & Bhal, 2020).

Today's employers demand different types of skills from their employees and workers than they were in the earlier period as a result of technological advancement and globalization process (Cinaret et al., 2009). In different words, now a day's workplaces require workers and employees with high technical skills attached with developed employability skills (G. K. G. Singh & Singh, 2008). Krishnan et al., (2019) stated that automation along with industrial transformation will impact working skills and other skills set in India.

The key issues that need to be addressed are as follow: what kinds of skills will be needed for future of workplaces and how can we craft our youth resilient as well as adaptable to these kind of change by incorporating lifelong learning, up-skilling and re-skilling. For creating capable global skilled workforce, skilling has to be an amalgamation of knowledge, aptitude, attitude and the appropriate competencies needed to perform various job roles. Education for all religion come in through the process of vocational education and it poses the various global education challenges (Marshall, 2010).

At this juncture when world is looking at Indian Skill Development Programmes and when Indian human resources are needed all across the globe, it is important to maintain quality assurance of these programmes. Industrial collaborations with institutes of trainings have proven successful model in developed countries like Germany, Finland, Brazil, etc. Though Skill development has become the buzz word in India and all stakeholders understand importance of skill development, it is important that efforts for the same may be synchronized with clear understanding of roles of each stakeholder.

The study tries to examine the students' perceptions towards the skill development programmes and its efficiency for providing better job opportunities to them. The study proposes the theoretical background mentioned in Figure 1. The paper is sub-divided into six sections. First section is about the introduction and basic background of study. Second section presents the Literature review, third section is about the research methodology and subsequent two sections are about the findings and discussion. Last section of the paper tries to bring out the conclusion and its implication with future research avenues.

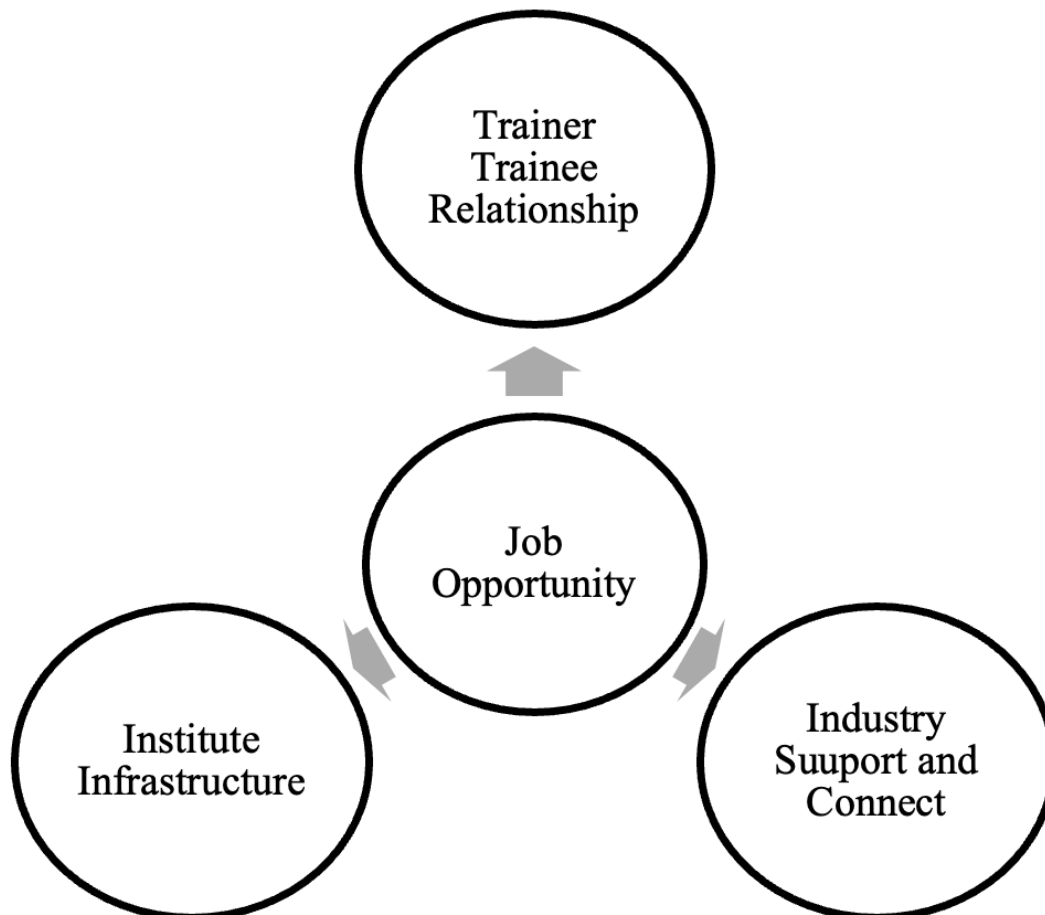


Figure 1. Theoretical Framework

REVIEW OF LITERATURE

The section of literature review examines the available studies on vocational education and skill development. It provides a basic background for a larger research, which is sub divided into three segments, i.e., 'Education and Infrastructure', 'Industry Support and Employability' and 'Trainer - Trainee Relation and Employability'. The literature is classified on the basis of the diverse nature of parameter to be considered for study. The reviews done in the study are not restricted to India but we have also tried to explore the condition of vocational education in other economies.

Education and Infrastructure

As per [Agarwal \(2007\)](#) for building of fully integrated education system in India, there is elongated path ahead because the confrontation is not limited to the regulatory framework used in Higher Education but other dimensions are also there. The major challenge is that vocational education is normally

perceived as an inferior education as compared to formal education and resultant in forcing the individual students to land up within the formal education system.

Erasmus and Breier (2015) found that almost all sectors of the economy usually suffer from scarcity of professionals and artisans; organizations, business units and government have the same opinion that there are huge shortages of technicians, artisans and engineers (Sheppard and Ntenga, 2014). [Stahl et al. \(2012\)](#) encompass that most of the companies have already established a world class training centres and huge learning campuses; they have already started working with the best institutes and universities in the world, where they are using the latest development being done in leadership for crafting and utilizing tools and technologies for making the best talents.

As per contemporary studies on skill development in technical education, Authors, [Greenan, Humphreys, and McIlveen \(1997\)](#) explored that programmes ought to concentrate on students' needs and be more intuitive in outline; curriculum advancement and on using proper educational methods which upgrade learning and create authority and relational aptitudes.

Industry Support and Employability

Nerdrum and Erikson (2001) have emphasized the importance of human capital and stated that the source of growth of any organization or nation has always been human capital and mental and physical and physical abilities and human resources are considered as prolific economic agents. Murphy et al., (1999) have advocated that on- job training and industry mentoring are the suitable and well accepted methods for the development of staff. [P Pfeffer and Jeffrey \(1998\)](#) suggested a list of human resource (HR) practices that are being adopted by the effective firms. One of the prominent practices is to make high investment in training and skill development.

[J. P. Robinson \(2000\)](#) has defined employability skills as the basic skills essential for getting job done and doing well on the job or work. Morrison and Hall (2002) found that employability skills are likely to smooth the progress of jobs within and among the organizations. [McArdle, Waters, Briscoe, and Hall \(2007\)](#) had proposed that employable individuals or trainee always obtain themselves into a proactive approach to fit into place in the domestic market and labour market. Better employability skill could also support employees to fine tune themselves as per the various changes obligatory to augment working skills or abilities which is as per the environmental needs of the workplace and demand of market ([Kazilan, Hamzah, & Bakar, 2009](#)).

[Cranmer \(2006\)](#) had noticed that there is high mismatch between the skills taught at school level and the skills actually desirable in job employment.

Therefore, youngsters or students entering in the skilled labour market segments are new or prone to take up anything available in the market than to decide or choosing the jobs meant for them (Hopper, 1977). Equally, there is occurrence of a mismatch between them and their jobs to high extent (Takase, Nakayoshi, & Teraoka, 2012).

Explorative activities and developmental activities are not linked with previous work experience (Savickas et al., 2009). It has been demonstrated in industry that skills of employability are significantly helping adults in adapting changes and improving upon the skills with better career opportunities in their respective workplace (Rasu et al., 2010). Results postulated that employability skills are significantly and positively correlated with adaptability of good career opportunities.

Trainer and Trainee Relation and Employability

Faculty or Trainer's commitment and their friendly behaviour with the trainees or students are always considered as an important factor by many scholars in the recent studies. (Hue, 2010; Tabbodi, 2009; Tiwari, 2019). Impact of vocational education training and its reforms on the educational practices of any economy has been noticed by numeral research studies (Hedberg & Harper, 1996; Mulcahy, 1996; V. M. Robinson & Robinson, 1993; Sanguinetti, 1994; Smith, 1997).

Billett (1999) has noticed that the execution of a unified or singular curriculum based framework have imposed throughout the years of training and the related reform has destabilized the educators' autonomy. Administrations of education in economy have conventionally had immense expectations of their training institute which provides vocational education and strengthens the training systems (Maurer, 2012). The augmentations of the entire global economy, with its associated social process are now permitting the smooth flows of information across national capital boundaries (Lingard, Knight, & Porter, 1994; Seddon, 1999). The importance of curriculum and the delivery of the content by trainers or faculty are also registered as an important factor in the education sector (Bhadwal & Panda, 1991; Hue, 2010).

The growth of 'new competitive state' has also been promoted from intervention done by governments organizations in favour of existing market forces this also acts as the 'primary steering mechanism' for the nation (Lingard et al., 1994). Education always serves for global or national interests and the development of human capital. Influence of globalizations as well as technology on the character of work along with the structure of the existing workforce has been registered by many researchers (Attwell, 1997; Waterhouse et al. 1999; Young & Guile 1997). Based on literature review collected from sociology and history

of education sector, (Benavot, 1983) portrayed few viewpoints that exist on the rise of vocational education in the economy during early twentieth century. Modern culture and its characteristics will always powerfully influence the human values and the choices; these choices have the huge range of issues that exist within the economy. These values are highly pushing the trainees to consider innovative ways of thinking and inducing new skills gained by the vocational educations ([M. Singh, 2013](#)).

The review suggests that any linkage between these important factors of skill-based education or vocational education is still not much explored. Hence, this study will make an attempt to establish relationship among the following three segments, i.e., Education and Infrastructure, Industry Support and Employability and Trainer - Trainee Relation and Employability.

RESEARCH METHODOLOGY

In the existing literature, to examine educational efficiency, three main research approaches were employed: a literature review; individual interviews or questionnaire and focus group discussions. The approach to gather the data in this study is deliberately focused on interviewing through questionnaire, since an interview scale on the changing role of vocational education training staff development is given by Harris et al. 2005. In view of the fact that general group discussion only generates the contextual information, the researchers firmly believe that there is the need of information to be collected from the individuals to enable a basic understanding of the perception, preference and personal impact of these changes. In this study, the conviction was to test that what trainee say in general group discussion or in friends may be differently reflected with what is happening personally around them. This is empirical research done on the basis of review of literature; the study has formed a self-administered questionnaire, which is developed for passed out, existing and potential trainees of skill development institutes and vocational educational institutes. The questionnaire includes 37 items to elucidate the factors affecting the employability of individual trainees or students of skill development institution. The collected data is analysed with the help of various statistical techniques like exploratory with confirmatory factor analysis and generalized linear model, these were performed with the help of computer software R Jamovi package 1.0.0. The sample of 515 trainees or students has been collected from the population size of 5155 students in NCT of Delhi, India. The sample size was considered by using the formula for sample size as mentioned in [Hulley, Newman, and Cummings \(2007\)](#).

The sample of 515 represents the total population size of 5155 students undergoing through training, as on 31st December 2019 in the respondent insti-

tutions. As suggested by Cochran (1977), we have considered almost ten percent of the population size. We had circulated the survey to 650 respondents but only 521 responses were received back. The response rate was 80.15 percent. However, only 515 were considered for the final analysis. To calculate the sample size (515) and oversampling of the sample size has been done to solve the issue of non-respondents in sampling suggested by Donald (1967), Hagbert (1968) and Johnson (1959). For the final research and analysis, the sample of 515 samples has been found suitable and further used in the study.

DEMOGRAPHIC CHARACTERISTICS

The respondents' profile has been sub-categorized on the basis of education, income, area of residence, salary expectation and expected sector of employment (Table 1).

Table 1

Demographic Characteristics.

Components	Choice	<i>f</i>	%
I joined course after my education up to	8 th class	28	5.44
	10 th class	129	25.05
	12 th class	241	46.80
	Graduation	107	20.78
	Any Other	10	1.94
I joined course because I wanted to get a good employment. My choice from the given sectors	Govt. Sector	265	51.46
	Public Sector	157	30.49
	Private Sector	30	5.83
	To be a part of my family business	28	5.44
	To start my own business	35	6.80
After completing this programme, I expect to get pay package	Minimum wages fixed by govt.	31	6.02
	Less than 15,000	52	10.10
	15,000 to 20,000	138	26.80
	20000 – 25000	178	34.56
	More than 25,000	116	22.52
After completing this programme, I expect to enter into a job profile of	A shop floor employee	29	5.63
	Executive	165	32.04
	Marketing Executive	103	20.00
	Service Executive	117	22.72

Continued on next page

Table 1 continued

Components	Choice	<i>f</i>	%
After school education what is your first preference	Liaison officer	91	17.67
	Any Other	10	1.94
	Graduation	41	7.96
	Job	116	22.52
	Short skill training	136	26.41
	Diploma	106	20.58
	ITI	116	22.52
Area	Rural	220	42.72
	Urban	295	57.28
Family Annual Income	Less than 3 lakh	71	13.79
	3.1 lakh - 5 lakh	90	17.48
	5.1 lakh - 7.25 lakh	209	40.58
	7.26 lakh - 10 lakh	115	22.33
	More than 10 lakh	30	5.83

f - Frequency Source: Authors Compilation

Study has considered the students of only those institutes where the skill-based education is being carried out. In the study 42.72 percent of the respondents were from the rural areas and rests are from the urban areas. The family income suggest that 40.58 percent of the respondents are from the income group of five to seven lakhs of income group and the second highest frequency of 115 respondents are from the second income group, which suggest that the higher the income group the lower the chances of the person to opt for the skill-based education. The reason behind this is that high income group prefers to send their wards for professional education as compared to the skill-based education.

Approximately 32.04 percent of the trainees are aspiring to become the executives, just after the completion of their skill-based education and 51.46 percent of them are aspiring for the government jobs after the completion of the course as they believe that the course has been initiated by the government of India. Hence, Government should accept their candidature in the public sector jobs or the government department jobs. Out of total respondents, 46.8 percent of the student joined the institute just after their senior secondary level and 25.05 percent has joined the course after their secondary level education. Most of the institute's courses are basically the skill-based course hence, it loses its importance after the graduation. Table 1 reveals that Soft skill training is the major preference of the students to learn just after the school education. However, second preference has been given to job and training. The subsequent preference is gaining diploma and pursuing other graduation degree.

FACTOR ANALYSIS

Factor analysis is a popular approach that has been extensively used in management and social sciences. Factor analysis is a statistical methodology that takes an exploratory and confirmatory approach to data analysis for inferential purposes (Byrne, 2001). Essentially, as the study is trying to develop a new framework, the entire analysis may be viewed as a combination of exploratory factor analysis and confirmatory analysis with mediation and moderation effects among the factors (Ullman, 2001). Hence, the study has taken the support of factor analysis in order to satisfy the objective.

Cronbach alpha test was performed in the study to check the reliability of questions taken in the questionnaire (Cronbach, 1951). Further, in order to assess the suitability of the data for principal component analysis, the uniqueness derived from the factor analysis were also assessed through KMO and Bartlett test. The results of different factors for reliability and sample adequacy are given in Table 2. The Cronbach's alpha test resulted in 94.5 percent of scale reliability. It indicates that the internal consistency of the selected scale is good (Bohrnstedt & Knoke, 1994). KMO and Bartlett's test (Table 2) which measures the sampling adequacy was done to test the eligibility of the data. The value of KMO is $0.945 > 0.5$, this value was observed and it indicates multivariate normality amongst variables. As the significance value observed in the research is less than 0.05, hence, factor analysis was performed consequently.

Table 2

KMO and Bartlett's Test .

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.945
Bartlett's Test	Chi-square	13384
	Degree of Freedom	595
	Sig.	.001

The items in the respective construct were individually subject to principal component analysis (PCA) with varimax rotation and it is based on Eigen value. 'Maximum likelihood' extraction method was used in combination with a 'varimax' Rotation. Literature suggests that items having factor loadings less than 0.5 can be eliminated (Hair et al., 2005). However, we have eliminated only those items which shows the loading less than 0.35 and the items showing the cross loadings. All having Eigen values of unity and above were removed (Hair et al., 2005). The results of factor loading are shown in Table 3.

Table 3**Principal Component Analysis.**

Constructs	Items/Variables	1	2	3	4
Institute Infrastructure	S5	0.888			
	S24	0.828			
	S25	0.792			
	S8	0.759			
	S7	0.754			
	S10	0.747			
	S6	0.735			
	S23	0.682			
	S17	0.594			
	S22	0.555			
	S26	0.554			
	S14	0.505			
Trainer Trainee Relation	S35		0.861		
	S33		0.797		
	S31		0.785		
	S36		0.772		
	S30		0.746		
	S9		0.718		
	S32		0.617		
	S29		0.588		
	S18		0.507		
	S20		0.496		
Industry Support	S1			0.809	
	S2			0.763	
	S13			0.727	
	S4A			0.716	
	S12			0.695	
	S4			0.638	
	S11			0.610	
	S3			0.608	
	S16			0.603	
	S37			0.449	
Job Opportunity	S27				0.555
	S21				0.474

Based upon the analysis shown in Table 3, we may suggest that as per the student's perspective, the four major constructs or factors that are involved in attaining effective skill-based education. As per the statements asked into

questionnaire, the study proposes to frame four constructs, namely, job opportunity, industry support, institute infrastructure and trainer-trainee relationship. The uniqueness score is fairly large for all the items as shown in Table 3; this suggests the appropriateness of data set (Stewart, 1981). Scree plot and factor analysis summary suggests that only four factors explain the 60.1 percent of cumulative variance. SS loadings suggest the sum of the squared loadings. This is used to determine the value of the particular factor. Results also determines the variance explained by individual factor and cumulative factors. Total cumulative variance is received as 60.10 percent for the scale.

FINDINGS OF THE STUDY

Through the factor analysis, we have got four constructs or factors for our study. The varimax rotation suggests that these four constructs include several variables as mentioned in Table 3. To substantiate our results, the confirmatory factor analysis has been employed and results are shown in Table 4. Analysis shows that the items drawn after the exploratory factor analysis are statistically fit for their corresponding factors. The factor loadings show the variance among the items.

Table 4

Factor Loadings

Factor	Indicator	Estimate	SE	Z	P
Institute Infrastructure	S5	1.123	0.0473	23.71	< .001
	S6	0.791	0.0441	17.93	< .001
	S7	0.829	0.0434	19.11	< .001
	S10	1.034	0.0506	20.44	< .001
	S14	0.571	0.0448	12.74	< .001
	S17	0.651	0.0354	18.39	< .001
	S22	0.581	0.0350	16.58	< .001
	S23	0.806	0.0436	18.47	< .001
	S24	1.232	0.0512	24.05	< .001
	S25	1.132	0.0482	23.46	< .001
	S26	0.853	0.0510	16.71	< .001
	S27	0.853	0.0510	16.71	< .001
Trainer Trainee Relation	S1	0.802	0.0377	21.25	< .001
	S2	0.642	0.0322	19.94	< .001
	S3	0.607	0.0357	17.00	< .001
	S4	0.642	0.0362	17.73	< .001
	S4A	0.592	0.0329	18.01	< .001
	S11	0.586	0.0402	14.57	< .001

Continued on next page

Table 4 continued

Factor	Indicator	Estimate	SE	Z	P
Industry Support	S12	0.664	0.0351	18.89	< .001
	S13	0.736	0.0363	20.25	< .001
	S28	0.351	0.0372	9.43	< .001
	S37	0.457	0.0382	11.94	< .001
	S9	0.757	0.0376	20.13	< .001
	S18	0.572	0.0353	16.20	< .001
	S20	0.583	0.0372	15.65	< .001
	S28	0.300	0.0363	8.24	< .001
	S29	0.637	0.0384	16.59	< .001
	S30	0.861	0.0408	21.12	< .001
	S31	0.823	0.0414	19.91	< .001
	S32	0.613	0.0359	17.09	< .001
	S33	0.869	0.0373	23.33	< .001
	S35	0.903	0.0403	22.42	< .001
	S36	0.858	0.0384	22.32	< .001
Job Opportunity	S27	0.575	0.0371	15.50	< .001
	S21	0.729	0.0384	18.97	< .001

The results reveal that there are four constructs involved in attaining effective skill-based education for better jobs in industry such as: job opportunity, industry support, institute infrastructure and trainer-trainee relationship. Job opportunity as a latent variable includes two statements, industry support and institute infrastructure as another latent variable includes 11 statements each and trainer-trainee relation include 10 statements. The detail of the estimates, Z statistic and P value is also mentioned in Table 4.

The estimates and related significance value with covariance suggest that almost all the variables considered in the study are significant as compared to the construct like demographics which are not significant for confirmatory factor analysis. Hence, in path analysis also bring the co-varying impact of one variable on another. Factor estimates can be determined through the factor covariance as given in Table 5 and the path analysis. The path analysis reveal the co-variation effect of one factor on another and further Figure 2 shows the direct and indirect effect of one factor on another. Results suggests that there is a good correlating relationship among all the variables.

Table 4 and 5 reveals the significant factors of the study. Hence in the study four constructs have been formed, i.e., institute infrastructure, trainer and trainee relationship, industry support and job opportunities available in the industry. Somehow the statistical data in Table 5 supports the consideration of all four constructs in our study. Covariance talks about the co-

Table 5**Factor Covariances.**

Construct	Covariance	Estimate	SE	Z	P
Institute Infrastructure	Trainer Trainee Relation	0.37	0.04	8.62	<.001
	Industry Support	0.44	0.03	11.50	<.001
	Job Opportunity	0.59	0.03	15.23	<.001
Trainer-Trainee Relationship	Industry Support	0.42	0.04	10.64	<.001
	Job Opportunity	0.48	0.04	10.75	<.001
Industry Support	Job Opportunity	0.74	0.03	23.02	<.001

movement of one factor with another factor which represents that one factor cannot work efficiently without the contribution of another factor in the study. Table 5 and 6 supports the path diagram and shows the co-variance relationship among the variables. The robustness of the model (shown in Table 6) can be judged by the goodness fit indices and badness fit index and it can be obtained through the path analysis as well.

Table 6**Model Fit Data.**

Test for Exact Fit	
χ^2	3317
Df	488
P	<.001
Fit Measures	
CFI	0.765
TLI	0.746
SRMR	0.110
Lower (RMSEA 90% CI)	0.103
Upper (RMSEA 90% CI)	0.110
AIC	39526
BIC	39976

The chi-square test measures the exact fit and the value revealed in the model is significant at 5 percent confidence level. Goodness-of-Fit can be measured by Comparative Fit Index (CFI) and Tucker Lewis Index (TLI). Comparative Fit Index (CFI) measures the incremental fit, which is 0.765, with the prescribed range of 0 to 1 being acceptable and the higher values indicating a better fit which is good for CFI. Tucker Lewis Index (TLI) also indicates the Goodness-of-fit index, the value of TLI as 0.746 was found in our study and the said values are within the prescribed limit of 0 to 1. Root Mean Square Error of Approximation (RMSEA) measures the Badness-of-Fit index and the value of RMSEA is 0.103 was found to be acceptable, as 0.10 is the well recommended limit for the acceptance of models, with lower RMSEA values is 0.110 which indicates a better fit model.

Cudeck and Browne (1983) established AIC and BIC to perform a cross-validation, with Akaike information criteria (AIC) being slightly liberal and Bayesian information criterion (BIC) being more conservative than AIC. Homburg (1991) applied AIC and BIC measures to SEM (structural equation modelling) and suggested that both AIC and BIC are performing well at identified level in the data-generating model. The higher the values of AIC and BIC the better the model is and here the results suggest that the value of BIC is better than AIC as shown in Table 6. All the latent variables considered in the study are interrelated and the structural model was revealed between the entire latent variable and are statistically significant. The estimates of all the variables considered in the study are shown in Table 5 with their respective p-values. Through model fit, we can say that these constructs are validated and can be used for any linear predictive modelling. To understand the dependencies of job opportunities in skill development institute we have applied general linear model. For models without interactions, Table 7 shows the indirect effects (mediated), the direct effects, and the total effects. The main predictive equations can be proposed in the given way.

Full Model of the study is given below:

Job opportunity ~ Industry Support + Institute Infra + Trainer Trainee Relation (1)

After taking Industry Support and Institute Infra as Mediators two indirect effect models are:

Trainer Trainee Relation \Rightarrow Industry Support \Rightarrow Job opportunity (2)

Trainer Trainee Relation \Rightarrow Institute Infra \Rightarrow Job opportunity (3)

The direct effects are the effects computed keeping the mediator's constant, thus the un-mediated effects. The total effects are the effects computed without the mediators, or, equivalently, the sum of the indirect and the direct effects. Mediators are variables expected to mediate the indirect effects. In this study

no demographic variable has been found significant to put mediating impacts on the model. The basic framework of the study suggests the model mentioned in Figure 2.

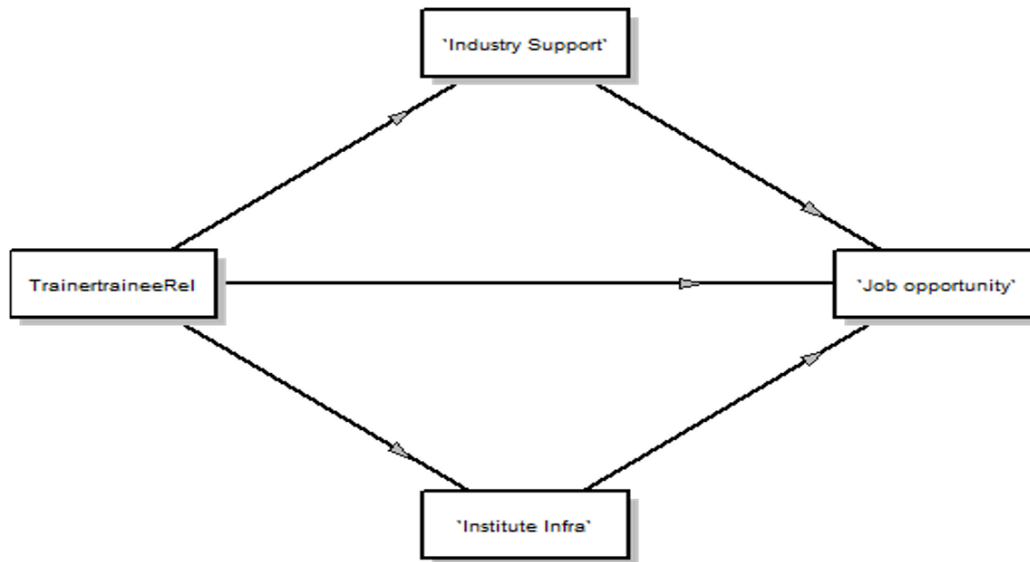


Figure 2. Model Diagram

The study shows the direct effect of Trainer-trainee relation lies on Job opportunity of a trainee or students. Supplementary to this finding one finding suggests that job opportunity of a student depends on the institutes infrastructure, industry support and trainer trainee relation as shown in full model of Table 7. The mediating effects suggest that a mediator construct intervenes the relation of two constructs. Here the finding suggests that institutes infrastructure, industry support is two mediators that influence the job opportunity of students and its direct effect with trainer trainee relation. In other words, we may conclude that trainer trainee relations will only the help to seek better job opportunity when there is better institutes infrastructure as well as industry support.

The findings suggest that the job opportunities of the students depend upon the kind of relationship that exists between the trainer and trainee, the better the relation will lead to better learning environment and thus to better understanding of the skills and the concepts. Table 8 reveals the model information and it includes a two-mediator model, a full model and two models with indirect effect of trainer trainee relationship with the job opportunity. This result may also suggest that the trainer trainee relationship works out well in the situation of better industry support and the institutes' infrastructure. The same has been show through mediation analysis shown in Table 7.

Table 7**Mediation Analysis.**

Indirect and Total Effects						
Type	Effect	Estimate	SE	β	Z	P
Indirect	Trainer-Trainee Relation-Industry Support-Job opportunity	0.05	0.005	0.24	9.18	<.001
	Trainer-Trainee Relation-Institute Infrastructure-Job opportunity	0.01	0.003	0.09	5.52	<.001
Component	Trainer-Trainee Relation-Industry Support	0.54	0.041	0.50	13.15	<.001
	Industry Support-Job opportunity	0.09	0.007	0.49	12.82	<.001
	Trainer-Trainee Relation-Institute Infrastructure	0.59	0.062	0.38	9.56	<.001
	Institute Infrastructure- Job opportunity	0.03	0.004	0.24	6.75	<.001
Direct	Trainer-Trainee Relation-Job Opportunity	0.02	0.0	0.10	2.68	0.007
Total	Trainer-Trainee Relation-Job Opportunity	0.09	0.008	0.43	11.04	<.001

DISCUSSION AND CONCLUSIONS

The paper focuses on the Indian education system and also considers it as an essential precondition for the employability and productivity of youth. Through this study, an effort has been made to develop guidelines that may prove milestone in the direction of industrial collaboration in educational institutions particularly for the purpose of skill development. Various aspects of skill development institutions have been studied through this empirical analysis with a special focus on role of industry.

The study suggests that Trainer Trainee relationship is really important for better job opportunity. Faculty or Trainer's commitment and their friendly behaviour with the trainees or students are always considered as an important factor by many scholars in the recent studies for getting better jobs in industry. (Hue, 2010; Tabbodi, 2009; Tiwari, 2019). The employability skills in marketplace refers to general as well as nontechnical competencies mandatory for performing almost all jobs in the industry, regardless of levels of jobs or its type (Ju, Zhang, & Pacha, 2012). There are many studies which have shown that the relevant job role of industry is dependent upon the industry connect or industrial training given to the students (Bynner, 2001; Gutman & Schoon, 2012). The study has proposed the linkage between industry connect, institutes infrastructure and job opportunities for trainees (Sherer & Eadie, 1987).

Nurturing skill-oriented training in India could be considered as a significant channel for improvising the working conditions of youth or individuals, as well as it can boost the employability of those trainees who are quite vulnerable in terms of skill set. Expansion of the skill development institutions is essential because of industry requirements of skilled manpower (Ahmad & Buchanan, 2016). These institutions mainly prepare the trainees for employment in the formal sector of industry. In this study, student perceptions for the skill development programme were collected and analysed by conducting exploratory factor analysis and other tests with the help of software R Jamovi 1.0.0 package. Through this study, an effort has been made to develop a model for improving the skills and employability of the students with industrial collaboration. The model suggests that the trainer trainee relationship works out well in the situation of better industry support and the institutes' infrastructure. The finding is similar to the research of Tooley (2005).

Therefore, the suggestion has been given to the policy makers that industry support with better industry connect in terms of industrial training and placement training is the need of the hour. It is also noticed that the institutes' infrastructure is not very well equipped as per students' perception. The improvement in infrastructure will lead to better training and skill development opportunities. Indian skills and innovations are not new to the world, glimpse of ancient Indian sculptures; carpentry, weaving, foundry and other crafts are quite evident from the archaeological remains of ancient Indian history. Therefore, it is recommended that skill development institutes and vocational education institutions should enhance the infrastructure and various industrial training lecturers, the similar findings were presented by Rahman et al., 2015. The future scope of the study suggests that as the students or trainee's perspective has been registered in the paper, similarly the trainer's point of view will enhance the importance of this study and will also provide a broader view to bridge the gap between employability raised by skill development institutes in India. The results also suggest that better relationship with

the trainer improves the employability skill of the trainers and also promotes industry connect.

REFERENCES

- Agarwal, P. (2007). Higher education in India: Growth, concerns and change agenda. *Higher Education Quarterly*, 61(2), 197-207.
- Ahmad, S. Z., & Buchanan, F. R. (2016). Choices of destination for transnational higher education: "pull" factors in an Asia Pacific market. *Educational Studies*, 42(2), 163-180. <https://doi.org/10.1080/03055698.2016.1152171>
- Banker, D. V., & Bhal, K. T. (2020). Creating world class universities: Roles and responsibilities for academic leaders in India. *Educational Management Administration & Leadership*, 48(3), 570-590. <https://doi.org/10.1177/1741143218822776>
- Bhadwal, S. C., & Panda, P. K. (1991). The Effect of a Package of Some Curricular Strategies on the Study Habits of Rural Primary School Students: a year long study. *Educational Studies*, 17(3), 261-271. <https://doi.org/10.1080/0305569910170304>
- Bynner, J. (2001). Childhood risks and protective factors in social exclusion. *Children & Society*, 15(5), 285-301. <https://doi.org/10.1002/chi.681>
- Cochran, W. G. (1977). The estimation of sample size. In *Sampling techniques* (3rd ed., p. 72-90). John Wiley & Sons.
- Cranmer, S. (2006). Enhancing graduate employability: Best intentions and mixed outcomes. *Studies in Higher Education*, 31(2), 169-184.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334. <https://doi.org/10.1007/bf02310555>
- Cudeck, R., & Browne, M. W. (1983). Cross-Validation Of Covariance Structures. *Multivariate Behavioral Research*, 18(2), 147-167. https://doi.org/10.1207/s15327906mbr1802_2
- Greenan, K., Humphreys, P., & McIlveen, H. (1997). Developing Work-based Transferable Skills for Mature Students. *Journal of Further and Higher Education*, 21(2), 193-204. <https://doi.org/10.1080/0309877970210205>
- Gutman, L. M., & Schoon, I. (2012). Correlates and consequences of uncertainty in career aspirations: Gender differences among adolescents in England. *Journal of Vocational Behavior*, 80(3), 608-618. <https://doi.org/10.1016/j.jvb.2012.02.002>
- Hedberg, J., & Harper, B. (1996). Interactive educational technologies: Effective design and application in the classroom. *3rd International Interactive Multimedia Symposium*, 160-168.
- Hue, M. T. (2010). The challenges of making school guidance culturally responsive: Narratives of pastoral needs of ethnic minority students

- in Hong Kong secondary schools. *Educational Studies*, 36(4), 357-369.
- Hulley, S. B., Newman, T. B., & Cummings, S. R. (2007). Choosing the study subjects: specification, sampling, and recruitment. *Designing Clinical Research*, 3, 27-36.
- Johnson, P. O. (1959). Development of the Sample Survey as a Scientific Methodology. *The Journal of Experimental Education*, 27(3), 167-176. <https://doi.org/10.1080/00220973.1959.11010620>
- Ju, S., Zhang, D., & Pacha, J. (2012). *Employability Skills Valued by Employers as Important for Entry-Level Employees With and Without Disabilities* (Vol. 35). SAGE Publications. Retrieved from <https://dx.doi.org/10.1177/0885728811419167>
- Kazilan, F., Hamzah, R., & Bakar, A. R. (2009). Employability skills among the students of technical and vocational training centers in Malaysia. *European Journal of Social Sciences*, 9(1), 147-160.
- Lingard, B. V., Knight, J., & Porter, P. (1994). Restructuring Australian schooling: Changing conceptions of top-down and bottom-up reforms. *School and community relations: Participation, policy, practices*, 81-99.
- Marshall, K. (2010). Education for all: where does religion come in. *Comparative Education*, 46(3), 273-287.
- Maurer, M. (2012). Structural elaboration of technical and vocational education and training systems in developing countries: the cases of Sri Lanka and Bangladesh. *Comparative education*, 48(4), 487-503.
- McArdle, S., Waters, L., Briscoe, J. P., & Hall, D. T. T. (2007). Employability during unemployment: Adaptability, career identity and human and social capital. *Journal of Vocational Behavior*, 71(2), 247-264. <https://doi.org/10.1016/j.jvb.2007.06.003>
- Mulcahy, D. (1996). Performing competencies of training protocols and vocational education practices. *Australian and New Zealand Journal of Vocational Education Research*, 4(1), 35.
- Pfeffer, J., & Jeffrey, P. (1998). *The human equation: Building profits by putting people first*. Harvard Business Press.
- Robinson, J. P. (2000). What are employability skills. *The workplace*, 1(3), 1-3.
- Robinson, V. M., & Robinson, V. M. (1993). *Problem-based methodology: Research for the improvement of practice*. Oxford: Pergamon Press.
- Sanguinetti, J. (1994). Exploring the discourses of our own practice: a case study. *Australian Journal for Adult Literacy Research and Practice*, 5(1), 31-31.
- Savickas, M. L., Nota, L., Rossier, J., Dauwalder, J.-P., Duarte, M. E., Guichard, J., ... van Vianen, A. E. (2009). Life designing: A paradigm for career construction in the 21st century. *Journal of Vocational Behavior*, 75(3), 239-250. <https://doi.org/10.1016/j.jvb.2009.04.004>

- Seddon, T. (1999). A self-managing teaching profession for the learning society. *Unicom*, 25(1), 15-29.
- Sherer, M., & Eadie, R. (1987). Employability skills: Key to success. *Thrust*, 17(2), 16-17.
- Singh, G. K. G., & Singh, S. K. G. (2008). Malaysian graduates' employability skills. *UNITAR e-Journal*, 4(1), 15-45.
- Singh, M. (2013). Educational practice in India and its foundations in Indian heritage: A synthesis of the East and West. *Comparative Education*, 49(1), 88-106.
- Smith, L. A. (1997). Open education revisited: Promise and problems in American educational reform (1967-1976). *Teachers College Record*, 99(2), 371-415.
- Stahl, G., Björkman, I., Farndale, E., Morris, S. S., Paauwe, J., Stiles, P., ... P (2012). Six principles of effective global talent management. *Sloan Management Review*, 53(2), 25-42.
- Stewart, D. W. (1981). The application and misapplication of factor analysis in marketing research. *Journal of Marketing Research*, 18(1), 51-62.
- Tabbodi, M. L. (2009). Effects of leadership behaviour on the faculty commitment of humanities departments in the University of Mysore, India: regarding factors of age group, educational qualifications and gender. *Educational Studies*, 35(1), 21-26. <https://doi.org/10.1080/03055690802288510>
- Takase, M., Nakayoshi, Y., & Teraoka, S. (2012). Graduate nurses' perceptions of mismatches between themselves and their jobs and association with intent to leave employment: a longitudinal survey. *International Journal of Nursing Studies*, 49(12), 1521-1530. <https://doi.org/10.1016/j.ijnurstu.2012.08.003>
- Tiwari, A. (2019). The corporal punishment ban in schools: Teachers' attitudes and classroom practices. *Educational Studies*, 45(3), 271-284.
- Tooley, J. (2005). Management of Private-aided Higher Education in Karnataka, India. *Educational Management Administration & Leadership*, 33(4), 465-486. <https://doi.org/10.1177/1741143205056214>
- Yusof, H. M., Mustapha, R., Mohamad, S. A. M. S., & Bunian, M. S. (2012). Measurement Model of Employability Skills using Confirmatory Factor Analysis. *Procedia - Social and Behavioral Sciences*, 56, 348-356. <https://doi.org/10.1016/j.sbspro.2012.09.663>

Development of Efficient Antimicrobial Zinc Oxide Modified Montmorillonite Incorporated Polyacrylonitrile Nanofibers for Particulate Matter Filtration

Priya Bansal and Roli Purwar*

*Discipline of Polymer Science and Chemical Technology, Department of Applied Chemistry,
Delhi Technological University, Shahbad Daultpur, Delhi 110042, India*

(Received August 4, 2020; Revised November 10, 2020; Accepted November 23, 2020)

Abstract: Ultrafine particulate matter and airborne microorganisms present in atmosphere are responsible for affecting the human health and the global climate. The development of bifunctional membranes which can simultaneously filter the particulate matter (PM) and inhibit the growth of microorganisms is the need of the hour. In this study, electrospun polyacrylonitrile (PAN)/zinc oxide modified montmorillonite (ZnO-Mt) nanofibrous nanocomposites with varying concentrations of ZnO-Mt ranging from 0.25 % to 1.00 % (w/w) have been prepared to be used as filtration membranes. The addition of ZnO-Mt in PAN dope solution affects its viscosity and ionic conductivity. The surface morphology of the nanofibrous membranes was studied using field emission scanning electron microscopy. The average diameter of PAN nanofibers and its nanocomposites was found to be between 247 nm to 468 nm. An increase in porosity, air permeability and water vapor transmission rate of the nanofibrous membranes was observed with an increase in concentration of ZnO-Mt in PAN nanofibers upto 0.75 %. The addition of ZnO-Mt enhanced the thermal stability of PAN nanofibrous membranes from 188 °C to 310 °C. The filtration efficiency of the nanofibrous membranes was evaluated using environment particle air monitor instrument. PAN/ZnO-Mt nanofibrous membranes having 0.75 % w/w ZnO-Mt exhibited filtration efficiency of 99.6 %. The antimicrobial property of PAN/ZnO-Mt nanofibrous membranes was studied against *S. aureus* and *E. coli* bacterial strains showing 98.85 % and 96.23 % antibacterial activity respectively.

Keywords: Zinc oxide modified montmorillonite, Polyacrylonitrile, Nanofibrous nanocomposite, Particulate matter, Thermal properties

Introduction

The presence of particulate matter (PM) in the atmosphere is amongst one of the serious environmental issues being faced by majority of the population in the world which in turn is having adverse effects on the human health and the global climate further disturbing the ecological balance [1]. The emerging industrialization and increased use of automobiles has led to an increase in the concentration of toxic pollutants in the atmosphere. PM is generally a mixture of water droplets present in air, dust, smaller organic and inorganic particles generated through vehicular emissions and incomplete burning of fossil fuels [2]. Based on the size of particles, PM has been categorized into PM_{2.5} (aerodynamic diameter of the particles $\leq 2.5 \mu\text{m}$) and PM₁₀ (aerodynamic diameters of particles between 2.5 and 10 μm) [3]. PM has become a hazard for the human health because of its smaller size which is responsible for its easier penetration into the respiratory tract and blood vessels leading to respiratory and cardiovascular problems [4,5].

In addition to PM, the presence of airborne microorganisms such as bacteria, viruses, fungi in the atmosphere are also responsible for adversely affecting the human health. These can result in contagious infectious diseases, allergies and respiratory problems and so on [6,7]. The most extensively used technology for the eradication of these airborne microorganisms is the utilization of antimicrobial air filters

[6,8-10].

An ideally efficient multifunctional fiber filter needs to be thinner so as to have maximum filtration efficiency, suggesting that the nanofibers are prospective to be used for fabrication of filters resulting in better filtration efficiency and lower air resistance [11]. The nanofibrous membranes are more advantageous for effective capture of PM_{2.5} and microorganisms due to larger aspect ratio, interconnected pore structure, uniform diameters and ease of incorporating multifunctional nanoparticles in the fibers [6,8,9,11-20]. Recent research on electrospun polyacrylonitrile nanofibrous membranes suggest it to be one of the attractive materials for filtration application because of their ability to form fibers easily, abrasion resistance, chemical stability and anticorrosive properties [5,12,15,17,21,22]. Jing and coworkers developed interlinked polyacrylonitrile/diethylammonium dihydrogen phosphate nanofibrous membranes having a desirable window size and adsorption of PM_{2.5} moisture bound particles [5]. Polyacrylonitrile/polysulfone composite nanofibrous membranes with binary structures have been prepared for effective PM_{2.5} filtration having 99.99 % filtration efficiency [15]. Polyacrylonitrile/poly(acrylic acid) nanofibrous membranes have been reported with a filtration efficiency >99.92 % against NaCl aerosol particles and lower pressure drop of 310 Pa [17]. Polyacrylonitrile/attapulgite composite nanofibers have been developed by acid activation of attapulgite and its modification with 3-aminopropyltriethoxysilane which were capable of effective for capture of PM and heavy metals [21].

*Corresponding author: roli.purwar@dtu.ac.in

Zinc oxide (ZnO) nanoparticles having varying morphologies show unique properties such as photocatalytic [23], photoluminescence [24], optical [25], antimicrobial activity [26] etc. The metal oxide nanoparticles especially ZnO supported montmorillonite provide reactive absorbance with high surface area, improved chemical reactivity and display added physicochemical functions of both the metal oxide nanoparticles and montmorillonite [27,28].

The aim of this study is to develop bifunctional PAN/ZnO-Mt nanofibrous nanocomposites which can be effective against bacterial strains and also capture PM_{2.5}. In this work, montmorillonite has been modified with zinc oxide nanoparticles via adsorption method. The addition of montmorillonite to ZnO enhances its antibacterial properties. A series of thermally stable PAN/ZnO-Mt nanofibrous nanocomposites have been developed using electrospinning technique. The properties of dope solution, morphological, structural, physical, and thermal properties of nanofibrous nanocomposites have been evaluated. The particulate matter filtration efficiency and antimicrobial activity of the nanocomposites have been studied. The developed nanofibrous nanocomposites on addition of ZnO-Mt are thermally stable upto 288 °C, thus making them suitable to be used at higher temperatures.

Experimental

Materials

Clay nanopowder (Mt) (~80-150 nm) and zinc oxide nanoparticles (~30 nm) were purchased from SRL, India. Zinc oxide nanoparticle modified montmorillonite (ZnO-

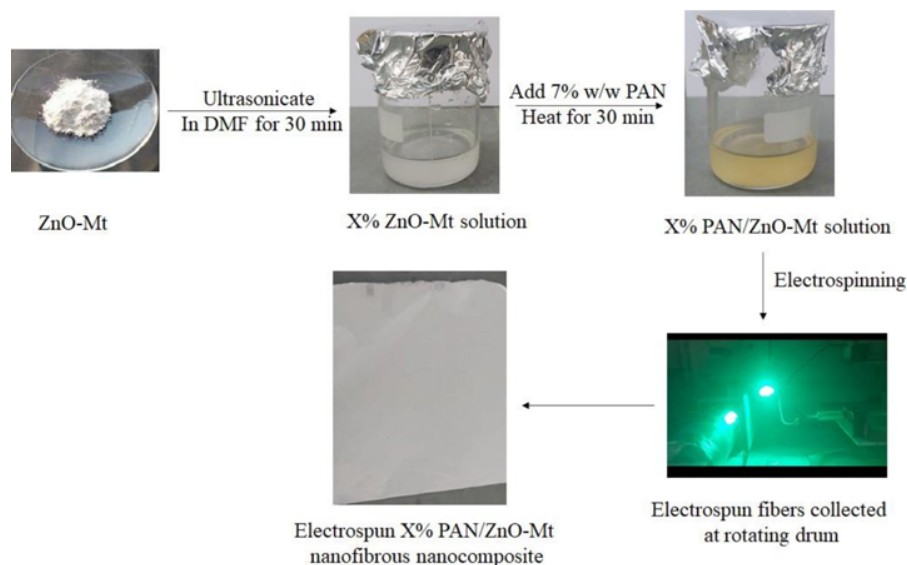
Mt) has been prepared by constantly stirring 3 g solution of Mt in 30 ml of 1 mM ZnO nanoparticle dispersion in water for a week at room temperature. The suspension was centrifuged to get ZnO-Mt. The obtained clay was dried and grinded using mortar pestel. ZnO-Mt was found to have an average particle size of 17 nm (supplementary data). Polyacrylonitrile (PAN) (average molecular weight=150,000) was purchased from Sigma Aldrich. N,N-Dimethylformamide (DMF) was purchased from Merck. Distilled water was used to carry out the experiment.

Preparation of Dope Solution

PAN dope solution (7 % w/v) was prepared by dissolving PAN powder in DMF at 70 °C with continuous stirring for 30 minutes. ZnO-Mt (0.25 % w/w w.r.t PAN) was ultrasonicated in 20 ml of DMF for 30 minutes and further PAN was added to the solution. The solution was heated at 70 °C for 30 minutes till a homogenous solution was obtained. This solution was denoted as 0.25 % PAN/ZnO-Mt solution. Similarly, 0.50 %, 0.75 % and 1.00 % PAN/ZnO-Mt solutions were prepared by varying the concentration of ZnO-Mt from 0.50 %, 0.75 % to 1.00 % respectively. The solutions were cooled down to room temperature and used for electrospinning of nanofibrous nanocomposites.

Characterization of Dope Solution

The viscosity of the solution was measured via Anton Paar Modular compact rheometer 302 (MCR) of cone plate geometry (40-2 °) of 0.21 mm gap at room temperature. The tests were carried out at a constant shear rate of 0.5 s⁻¹. The ionic conductivity of the solutions was measured using CON



Here X%= 0.25%, 0.50%, 0.75% and 1.00%

Figure 1. Schematic representation of preparation of nanofibrous nanocomposite.

700 conductivity meter at room temperature conditions.

Electrospinning of Nanofibrous Nanocomposites Membrane

Electrospun PAN nanofibrous membranes were prepared by electrospinning machine (Royal Enterprises, India). A 12 ml syringe containing 7 % (w/v) PAN solution was placed on the pump. The process of electrospinning was carried through the needle tip to the grounded drum roller covered with aluminium foil. The solution was electrospun for 4 h at a flow rate of 6 ml/h, providing a high voltage of 18 kV to form stable jets and the nanofibers were collected at a tip to collector distance of 12 cm at room temperature conditions. PAN/ZnO-Mt nanofibrous nanocomposites were electrospun under similar conditions using the prepared 0.25 %, 0.50 %, 0.75 % and 1.00 % PAN/ZnO-Mt solutions. The nanofibrous nanocomposites containing ZnO-Mt in varying concentrations have been denoted as 0.25 % PAN/ZnO-Mt, 0.50 % PAN/ZnO-Mt, 0.75 % PAN/ZnO-Mt and 1.00 % PAN/ZnO-Mt nanofibrous nanocomposites.

Characterization of Nanofibrous Nanocomposites

The surface morphology of the nanofibrous nanocomposites was analyzed using field emission scanning electron microscope with environmental SEM FEI Quanta 200 F with oxford EDS IE 250X Max 80, Netherlands. The average diameter of the nanofibrous membranes was calculated by measuring diameters of 20 nanofibers using ImageJ software. The XRD analysis was carried out using X-ray diffraction, D8 Advance, Bruker by varying 2θ from 10° to 35° . The degree of crystallinity of the nanofibrous membranes was calculated using the equation given below.

$$\text{Degree of crystallinity} = \frac{A_c}{A_c + A_a} \times 100$$

where A_c and A_a are area of crystalline and amorphous peaks respectively.

The thermogravimetric analysis of the nanofibrous membranes was studied on Perkin Elmer TGA instrument. For the analysis approximately 6 mg of the sample was weighed and the measurements were carried out under nitrogen atmosphere at a heating rate of $10^\circ\text{C}/\text{min}$ from 30°C to 800°C . Air permeability of the nanofibers was measured using air permeability tester WIRA (ASTM D 737, ISO 9237). Burst strength of the membranes was measured (Diaphragm Bursting, Materiau IngenierieI, France, ASTM D 3787-07).

Liquid displacement method [29] was used to determine the porosity of the nanofibrous nanocomposites. Hexane was used as the displacement liquid because of its easy permeation through the pores of the fibers. A rectangular piece of dimensions 20×20 mm was immersed in 10 ml of hexane (V_1) in a graduated measuring cylinder for an interval of 10 minutes. The volume of hexane after

immersing the samples was recorded (V_2). The residual hexane volume (V_3) in cylinder after removal of hexane impregnated sample was also recorded. The porosity was calculated by the equation:

$$\text{Porosity (\%)} = \frac{V_1 - V_3}{V_2 - V_3} \times 100$$

Standard ASTM D 1653 was used to study the WVTR of the nanofibrous nanocomposites. In this method, the samples measuring 9.5 cm^2 were sealed on the beaker containing distilled water at room temperature conditions and weighed. After 24 hours the samples were weighed again. Water vapor transmission rate ($\text{g}/\text{m}^2/\text{day}$) was calculated by the equation given below:

$$\text{WVTR} = \frac{W_1 - W_2}{A \times 24}$$

where W_1 and W_2 represent the weight of assembly before and after 24 h of water evaporation, respectively, and A represents the transmission area of the sealing samples.

Filtration Efficiency

The filtration efficiency of the nanofibrous nanocomposites was evaluated using Environmental Particle Air Monitor (EPAM-5000, HAZ-DUST, USA) having flow rate of 4 l/min and sample rate of 1 second for 6 hours [30]. The monitor of the equipment is highly sensitive and is based on the scattering of light for measurement of concentrations of particle in mg/m^3 . The levels of PM10, PM2.5 and PM1.0 can be monitored using the interchangeable size selective impactors. The frequency of certain volume of collected air is termed as the sample rate of respirable suspended particulate matter (RSPM). In this test, the filters of $1 \mu\text{m}$, $2.5 \mu\text{m}$ and $10 \mu\text{m}$ are selected to filter the particle size of $1 \mu\text{m}$, $2.5 \mu\text{m}$ and $10 \mu\text{m}$ respectively, present in the volume ambient air are to be passed through the filter. The particle concentration of the sample was determined by calculating difference between the weight of the sample before and after the test. This is expressed as the concentration of particulate matter collected in mg/m^3 . Filtration efficiency of the



Figure 2. HAZ DUST EPAM 5000 Instrument used for RSPM test.

nanofibrous membranes have been calculated using the following equation

$$\text{Filtration efficiency (\%)} = \frac{Y-X}{X} \times 100$$

where X and Y are the particle concentration of PAN nanofibers (control) and PAN/ZnO-Mt nanofibers.

The performance of the nanofibrous nanocomposites as a filter was performed by use of sodium chloride aerosols of average diameter of 300-500 nm. The experiment was carried under a constant velocity flow of 4 l/min.

Antimicrobial Activity of Nanofibrous Nanocomposite

The antibacterial activity of the nanofibrous mats was studied by modified disc diffusion test (AATCC 30) against Gram positive *S. aureus* and Gram negative *E. coli* as described by Purwar *et al.* [31,32]. Circular discs (diameter= 6 mm) of the nanofibrous membranes were sterilized for 24 hours under UV light in laminar air flow which were further placed on cultured agar plates and incubated at 37 °C for 24 hours. The incubated samples were shaken in 10 ml sterilized water at 120 rpm for 30 minutes for the release of bacteria. The sample was further serially diluted upto 10^{-5} times. 20 μ l of serially diluted sample was spread over agar plates and incubated at 37 °C for 24 hours. The experiment was performed in triplicates. The bacterial colonies were counted and % antibacterial activity was determined using the following equation.

$$\text{Antibacterial activity} = \frac{X-Y}{X} \times 100$$

where X and Y denote the number of colonies in control and treated sample.

Results and Discussion

Characterization of Dope Solution

The variation of viscosity of the solution at a constant shear rate of 0.5 s^{-1} has been studied (Figure 3a). It was found that addition of ZnO-Mt to PAN affects the viscosity of the solution. The viscosity of the solution increases on the addition of ZnO-Mt into PAN solution upto 0.50 % (w/w) concentration of ZnO-Mt while decreases on further increase in the concentration of ZnO-Mt (Figure 3a). The viscosity of the PAN/ZnO-Mt solution upto 0.75 % (w/w) was found to be higher than PAN solution. The increase in viscosity of the solution upto 0.50 % PAN/ZnO-Mt solution might be due to stronger interaction between PAN and ZnO-Mt in the solution. The stronger interaction could limit the polymeric chain mobility in the suspension resulting in higher entanglement of polymeric chains which facilitates extrusion of the solution during the process of electrospinning [33]. Whereas on increasing the concentration of ZnO-Mt to 0.75 % and 1.00 % in PAN solution, the interaction between

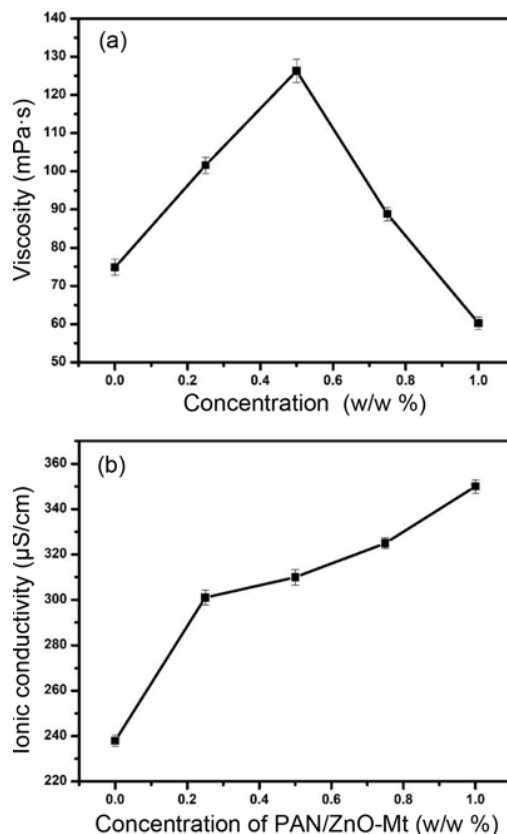


Figure 3. (a) Viscosity and (b) ionic conductivity of PAN/ZnO-Mt nanofibrous nanocomposites.

PAN and ZnO-Mt weakens due to accumulation of greater amount of ZnO-Mt resulting in a decrease in viscosity of the solution. Researchers have reported enhancement as well as reduction in viscosity of polymer solution on incorporation of clay [34,35]. The dispersion and exfoliation of platelets during mixing reduces the shear viscosity of the nanocomposites in comparison to pristine polymer [36].

The ionic conductivity of pure PAN solution was found to be 238 $\mu\text{S/cm}$ solution and further increased on addition of ZnO-Mt into PAN solution upto 350 $\mu\text{S/cm}$. The ionic conductivity of the solution is governed by two factors, the mobility of the ions and the amount of carrier ions present in a solution. The amount of carrier ions in a solution is affected by the concentration of ions present in the solution. Mt possess negative charge and OH group linked to Al or Si. This charge is responsible for the interaction between Mt and PAN. The ionic conductivity of PAN solution increases on addition of ZnO-Mt as the number of charged ions increases in the solution (Figure 3b). Sikkantar *et al.* [37] reported that the ionic conductivity of polyacrylonitrile-ammonium bromide polymer electrolyte increased from $1.3 \times 10^{-4} \text{ S/cm}$ in 95:5 composition to $2.5 \times 10^{-3} \text{ S/cm}$ in 70:30 composition due to the increase in mobility of the charge carriers.

Morphology of PAN/ZnO-Mt Nanofibrous Nanocomposites

For the effective capture of the particulate matter, the nanofibrous nanocomposites must have smaller pores with thinner diameter of the nanofibrous membranes, high porosity and a uniformity in structure. The generation of nanofibers of finer diameters result in enhanced performance of the filter [38]. The nanofibrous membranes have been developed by varying the concentrations of ZnO-Mt in PAN solution from 0.25 % to 1.00 %. The field emission scanning electron micrographs of the PAN/ZnO-Mt nanofibrous

nanocomposites are shown in Figure 4.

The surface morphology of the nanofibrous nanocomposites suggested that on addition of ZnO-Mt into the PAN matrix, the roughness on the surface of the nanofibers increased. PAN nanofibrous membranes were smooth (Figure 4a) in comparison to PAN/ZnO-Mt nanofibrous nanocomposites. As the concentration of ZnO-Mt increases in PAN/ZnO-Mt nanofibrous nanocomposites respectively, the fibers changed their orientation from beaded to beadless uniformly densely packed nanofibers. But as the concentration increased to

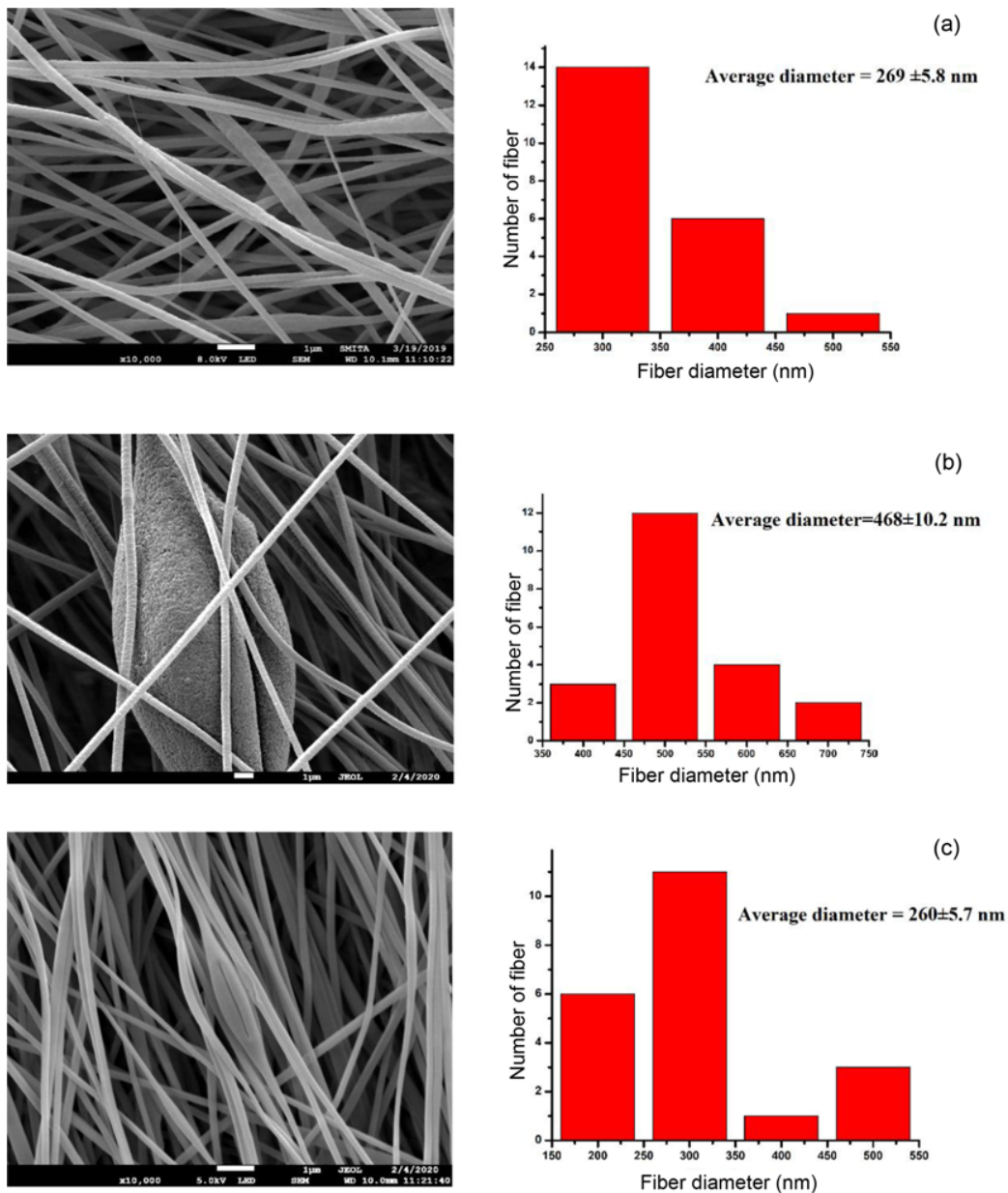


Figure 4. FESEM images of (a) PAN nanofibers, (b) 0.25 %, (c) 0.50 %, (d) 0.75 %, and (e) 1.00 % PAN/ZnO-Mt nanofibrous nanocomposites.

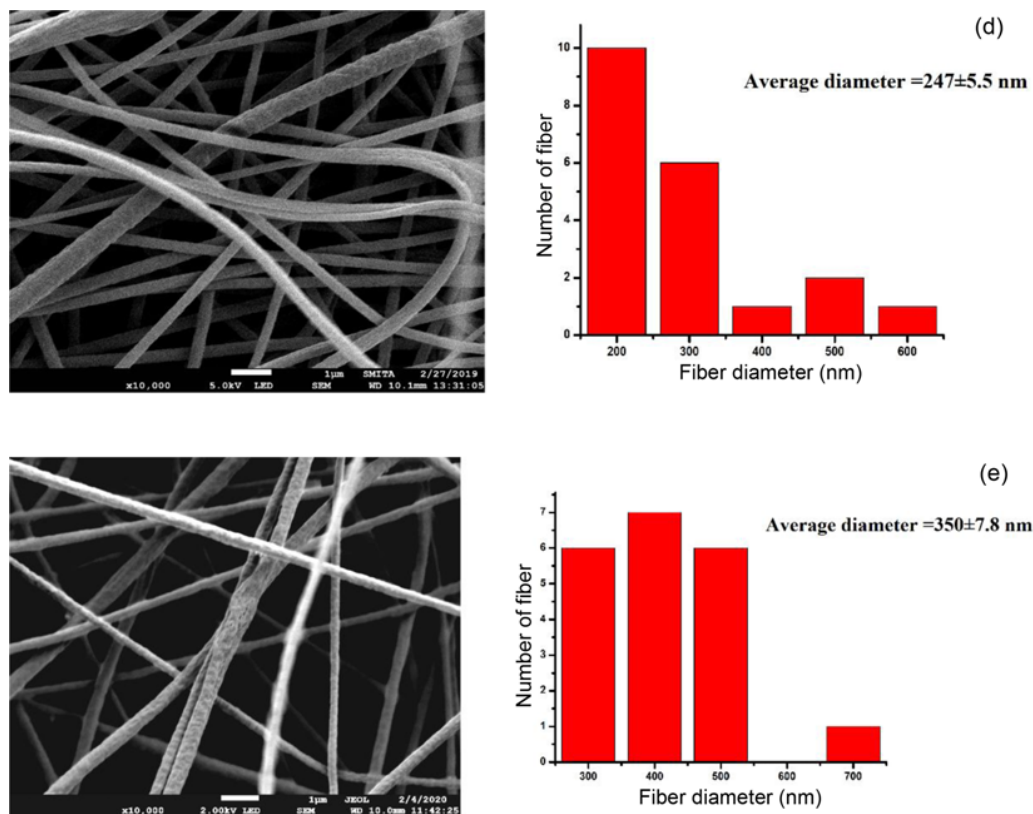


Figure 4. Continued.

1.00 %, beaded and less dense nanofibers were formed. An increase in the concentration results in instability of the solution at the tip of spinneret during electrospinning resulting in formation of beaded fibers with a thicker diameter [39]. The diameter of the fibers increased at higher concentration due to viscosity changes in the solution which is due to the entanglement of the molecules of the polymeric chain [33].

The viscosity and electrical conductivity of the dope solution has a major effect on the diameter of the nanofibers. The size of the fiber and its uniformity is dependent on the viscosity of the polymer solution [40]. Non-uniform beaded fibers are formed at lower viscosities whereas at a higher viscosity the electrospinning jet experiences a difficulty in ejection. The increase in conductivity is also responsible for a decrease in the diameter of the fibers as the electrospinning spinneret jet carries higher amount of electric charges which results in application of much higher forces of elongation on the surface of the fiber and finer diameter fibers are formed. Thus, a solution of an optimum viscosity and conductivity is required for electrospinning [39]. The fiber morphology affects the capture mechanism of the particles and also the performance of the filter [40]. In this study, a regular trend in average diameter of nanofibers was not observed due to the

combined effect of viscosity and conductivity of the solution resulting in formation of nanofibers. 0.75 % PAN/ZnO-Mt nanofibrous nanocomposites had the smallest average diameter of 247 nm amongst the series of nanofibers developed. The addition of clay increases the fiber diameter as an increase in clay content increases the viscosity as well as electrical conductivity of the solution [41]. Liu *et al.* [17] reported that the average diameter of polyacrylonitrile/polyacrylic acid nanofibers is dependent both on the viscosity and conductivity of the dope solution. Similar results have been obtained by the addition of sodium montmorillonite into PAN matrix which resulted in an increase in the surface roughness of the produced composite nanofibers and decrease in average diameter of the nanofibers with respect to their pristine counterpart [42].

Structural Properties of PAN/ZnO-Mt Nanofibrous Nanocomposites

The presence of ZnO-Mt in the nanofibrous membranes was analyzed using XRD (Figure 5). XRD pattern shows a peak at 16.9° for PAN nanofibers whereas for PAN/ZnO-Mt nanocomposite nanofiber two peaks were observed at 16.9° and 31.8° . The appearance of peaks at 30.9° to 33.1° suggested the presence of ZnO-Mt. Strong anatase peaks of

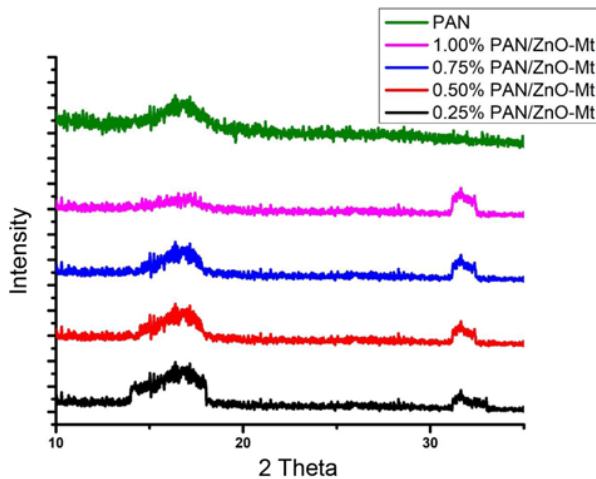


Figure 5. XRD patterns of PAN nanofibers and PAN/ ZnO-Mt nanofibrous nanocomposites.

ZnO are observed at 31 °, 34 ° and 36 ° [43], the modification of Mt with ZnO resulted in broadening of peak. It was observed that with addition of ZnO-Mt to PAN nanofibers the peak appeared at 16.5 ° for PAN was broadening, suggesting localization of ZnO-Mt into the nanofibrous membranes. The fiber diffraction pattern for PAN having hexagonal crystal system shows two equatorial peaks at $2\theta=29.5^\circ$ and $2\theta=17^\circ$ with degree of crystallinity of PAN fibers in the range 8.8 % to 11.27 %. The crystallization slows down during electrospinning because of rapid solidification of stretched chains of polymer at higher elongation rates hindering crystal formation [44-46]. The degree of crystallinity of PAN was found to be 13.33 % whereas on addition of ZnO-Mt to the PAN nanofibrous membranes, the degree of crystallinity increased in the order 0.25 % PAN/ZnO-Mt (38.10 %) > 0.50 % PAN/ZnO-Mt (28.03 %) > 0.75 % PAN/ZnO-Mt (26.82 %) > 1.00 % PAN/ZnO-Mt (17.65 %). The zinc oxide modified clay may act as nucleating site and enhanced the crystallinity of nanocomposite fibers. However, higher amount of clay in nanofibrous matrix affects the crystallization behavior of polyacrylonitrile polymer chain and reduces the crystallinity.

Thermal Properties of PAN/ZnO-Mt Nanofibrous Nanocomposites

The performance of the nanofibrous nanocomposites under thermal conditions was studied by thermogravimetric analysis. The thermal degradation of the nanofibrous nanocomposites was found to be a two-step degradation process whereas for PAN nanofibers it was a three-step degradation process (Figure 6). On addition of ZnO-Mt into PAN nanofibrous membranes the onset temperature of first step of degradation was found to increase from 92 °C in PAN nanofibers to 288 °C in nanofibrous nanocomposites. The second step of thermal decomposition starts at 188 °C

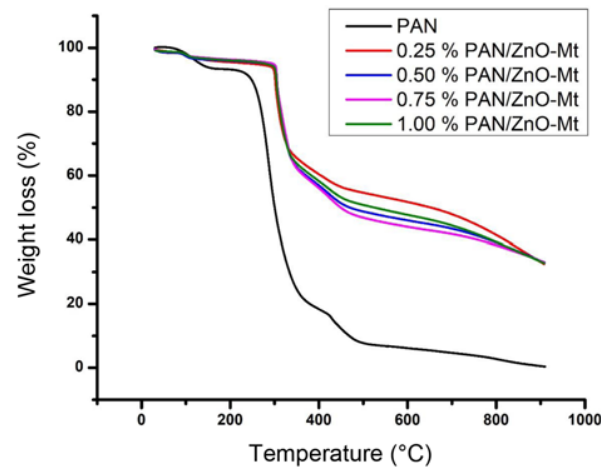


Figure 6. Thermogram of PAN nanofibers and PAN/ZnO-Mt nanofibrous nanocomposites.

for PAN nanofibers and increases to 381 °C in nanofibrous nanocomposites. The nanocomposite fibers show only 40 % weight loss at 900 °C higher in comparison to PAN nanofibers due to stability of ZnO-Mt at higher temperatures. Addition of ZnO-Mt enhances the thermal stability of PAN nanofibers due to the synergistic effect of ZnO-Mt. The tremendous improvement in thermal properties of nanofibers suggests uniform dispersion of ZnO-Mt in nanofibrous matrix. The addition of ZnO-Mt into PAN nanofibers resulted in much more thermally stable nanofibrous nanocomposites in comparison to PAN nanofibers as the presence of ZnO-Mt prevents degradation of PAN. ZnO nanoparticles have also been reported to enhance the thermal stability of PAN nanofibers [43]. Similar results have been reported by Almuhammed *et al.* [42] for electrospun PAN/Na-MMT hybrid nanofibers comprising 5, 10 and 19 wt % Na-MMT, an increase in thermal stability of the nanofibers was observed with incorporation of Na-MMT in PAN.

Physical Properties of PAN/ZnO-Mt Nanofibrous Nanocomposites

One of the basic requirements of a nanofibrous nanocomposite is its porosity as the porous membrane offers a large surface area for binding. The porosity of the nanofibrous nanocomposites was studied using liquid displacement method. The average porosities of the nanofibrous nanocomposites are tabulated in Table 1. 0.75 % PAN/ZnO-Mt nanofibrous nanocomposites were found to have the highest porosity of 85.70 % respectively. The porosity of the nanofibrous nanocomposites were found to increase with an increase in the concentration of ZnO-Mt upto 0.75 % w/w. whereas 1.00 % PAN/ZnO-Mt nanofibrous nanocomposite had porosity equal to PAN nanofibers. It was observed that the porosity of the nanofibrous nanocomposites increased with decrease in fiber diameter suggesting that

uniformly developed interconnected network of fibers was accountable for better porosities of the nanofibrous nanocomposites.

The transportation characteristics of nanofibrous membranes is essential to determine the suitability of the membrane as a filter which were studied by evaluation of water vapor transmission rate of the nanofibrous nanocomposites. The water vapor transmission rate was found to increase with an increase in the concentration of ZnO-Mt in PAN/ZnO-Mt nanofibrous nanocomposites. The increase in water vapor transmission rate in case of PAN/ZnO-Mt in comparison to PAN is due to enhancement of porosity. PAN nanofibers were more densely packed compared to PAN/ZnO-Mt and had lower pore size which resulted in lower water vapor transmission rate of the nanofibrous nanocomposites. As indicated from FESEM images, the addition of ZnO-Mt to PAN nanofibers resulted in an increased pore size and lower nanofiber diameters which increased its water vapor transmission rate compared to PAN.

The measurement of passage of air through a specified area of a fiber is termed as air permeability which determines its thermal comfort [47]. The air permeability is also one of the important measure to study the permittivity of the membrane for filters. The air permeability of a fiber is majorly dependent on the porosity, pore size and diameter of the fiber influencing its openness [48]. PAN nanofibers were found to have an air permeability of $5.4 \text{ l/m}^2/\text{s}$ and was found to increase with addition of ZnO-Mt to PAN nanofibers. The air permeability of the nanofibrous membranes was directly related to the porosity of the nanofibrous membranes. The air permeability increased by 4.7 times in 0.75 % PAN/ZnO-Mt in comparison to PAN nanofibrous membranes. Roche *et al.* [49] developed laminated PAN nanofibers for air filtration having air permeability of $4 \text{ l/m}^2/\text{s}$ which was comparatively lower due to adhesion method reducing the porosity of the membrane. Wang *et al.* [50] fabricated waterproof and breathable electrospun PAN nanofibers modified with waterborne fluorinated polyurethane having an air permeability of $5.9 \text{ l/m}^2/\text{s}$ suitable for protective clothing.

Burst Strength of PAN/ZnO-Mt Nanofibrous Nanocomposites

The ability of a material to maintain its continuity on

application of pressure is defined as the burst strength of the material. The burst strength of the nanofibrous nanocomposites was found to increase from 8.9 to 27.8 N/mm^2 . As the concentration of ZnO-Mt increases to 0.75 % in PAN nanofibers, the burst strength increases as the modified clay present within the nanofibers act as a reinforcing filler whereas a slight decrease was observed for 1.00 % PAN/ZnO-Mt nanofibrous nanocomposites which might be due to a less dense structure of the membrane as observed from FESEM.

Filtration Efficiency of PAN/ZnO-Mt Nanofibrous Nanocomposites

The filtration performance of the nanofibrous nanocomposites was evaluated by environment particle air monitor test. The capture of PM_{2.5} was performed in environmental conditions (the machine containing the filter samples was kept at busy roadside to capture the PM_{2.5} for 6 hours). The performance of the nanofibrous nanocomposites as a filter is shown in Figure 7. It was observed that as the concentration of ZnO-Mt in nanofibrous nanocomposites increased from 0.25 % to 0.75 % the filtration efficiency of the membrane increased and further decreased as the concentration increases to 1.00 %. The RSPM test analysis showed that 0.75 % PAN/ZnO-Mt nanofibrous nanocomposites adsorbed the highest amount of PM_{2.5} (0.966

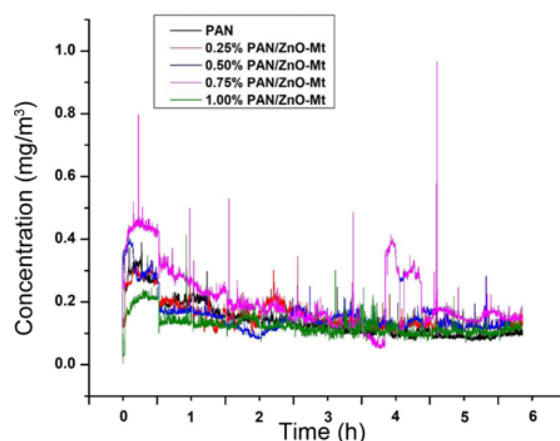


Figure 7. RSPM test of PAN and PAN/ZnO-Mt nanofibrous nanocomposites.

Table 1. Various parameters of the nanofibrous nanocomposites

	PAN	PAN/ZnO-Mt			
		0.25 %	0.50 %	0.75 %	1.00 %
Water vapor transmission rate ($\text{g/m}^2/\text{day}$)	0.7876	0.9767	1.0417	7.8125	2.8646
Porosity (%)	50.00	75.00	80.00	85.70	50.00
Air permeability ($\text{l/m}^2/\text{s}$)	5.4	7.9	14.2	25.3	6.6
Burst strength (N/mm^2)	8.9	15.7	21.6	27.8	25.4

mg/m³) while PAN nanofibers were comparatively less efficient in filtering PM2.5. This suggests that the addition of ZnO-Mt enhances PM2.5 capture over the nanofibrous membranes. The filtration efficiency of the nanofibrous membranes increased from 56.5 % to 99.6 % as concentration increased from 0.25 % to 0.75 % whereas a rapid decrease to 26.5 % was observed in case of 1.00 % PAN/ZnO-Mt (Table 2). The particle concentration of PM2.5 increases from 31.67 mg/m³ in PAN nanofibers to 79.58 mg/m³ in 0.75 % PAN/ZnO-Mt nanofibrous nanocomposite. Further a decrease in particle concentration to 50.43 mg/m³ was observed for 1.00 % PAN/ZnO-Mt nanofibrous nanocomposite. This trend in filtration efficiency of the nanofibrous membranes is due to the increasing amount of ZnO-Mt which decreases the diameter of the nanofibers which is further responsible for an efficient capture of PM2.5. As the surface roughness of the fiber increases, it provides larger number of sites for adsorption of PM2.5, thereby increasing the filtration efficiency which is evident from the RSPM test of the nanofibrous membranes. From the above studies it was found that 0.75 % PAN/ZnO-Mt nanofibrous nanocomposites had the highest filtration efficiency. So, their surface morphology was studied after the RSPM test (Figure 8). The surface of the nanofibrous membranes changed after the capture of PM2.5. The PM2.5 particles were found to agglomerate at the surface of the nanofibrous membranes. It

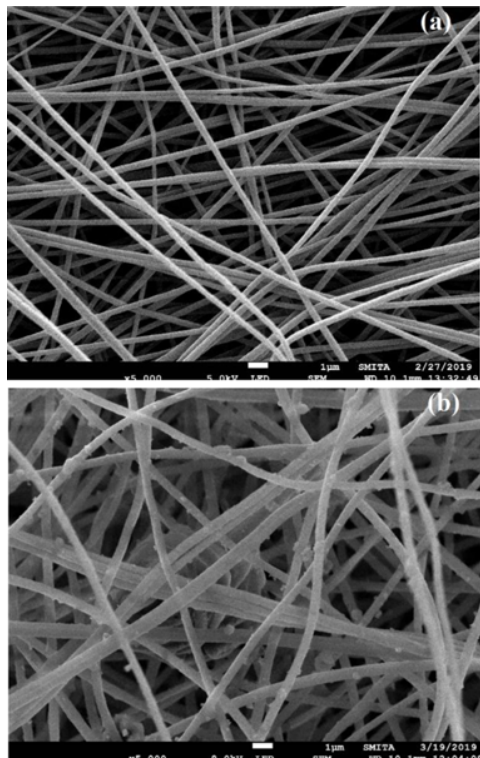


Figure 8. FESEM images (a) before RSPM test and (b) after RSPM test 0.75 % PAN/ZnO-Mt nanofibrous nanocomposites.

Table 2. Filtration efficiency of PAN/ZnO-Mt nanofibrous nanocomposites

S. no.	Nanofibrous nanocomposites	Filtration efficiency (%)	Particle concentration (mg/m ³)
1	PAN		31.67
2	0.25 % PAN/ZnO-Mt	56.5	62.39
3	0.50 % PAN/ZnO-Mt	94.6	77.59
4	0.75 % PAN/ZnO-Mt	99.6	79.58
5	1.00 % PAN/ZnO-Mt	26.5	50.43

was found that the particles got attached one above other on the nanofibrous membranes forming agglomeration or bigger sized particles which can also be due to the affinity of these particles to get attached on the surface of these nanofibers. The dust particles of PM2.5 accumulated on the nanofibrous membranes via van der Waals interaction [51]. The particles of PM2.5 move and merge as larger particles along the nanofibrous membranes, leaving surface for absorption of more particles the particles attach over one another forming agglomeration of larger sizes [52]. Similar results of agglomeration of PM2.5 wrapped and deposited over nanofibrous network has been observed for PAN/DEAP nanofibers [5]. The EDX analysis (Figure S4) of 0.75 % PAN/ZnO-Mt nanofibrous nanocomposite before and after RSPM test was studied (Table 3). It was observed that the particles of elemental composition carbon (C), nitrogen (N), sulphur (S) and oxygen (O) were mainly captured on the nanofibrous nanocomposites suggesting capture of nitrides, sulphates and carbon emissions.

Jing *et al.* [5] enhanced capture of PM2.5 generated by cigarette burning using electrospun PAN/ionic liquid nanofibers in comparison to PAN nanofibers mainly due to improvement in surface roughness, dipole moment and hydrophilicity of PAN on modification. PAN/PAA nanofibrous membranes with weight ratio 6:4 have been reported with filtration efficiency of 99.994 % for sodium chloride aerosol particles (300-500 nm) [17]. The RSPM test for 0.75 % sericin/PVA/clay nanofibers having burst strength 10 N/mm² filtered 0.725 mg/m³/s of particulate matter compared to other concentrations [32].

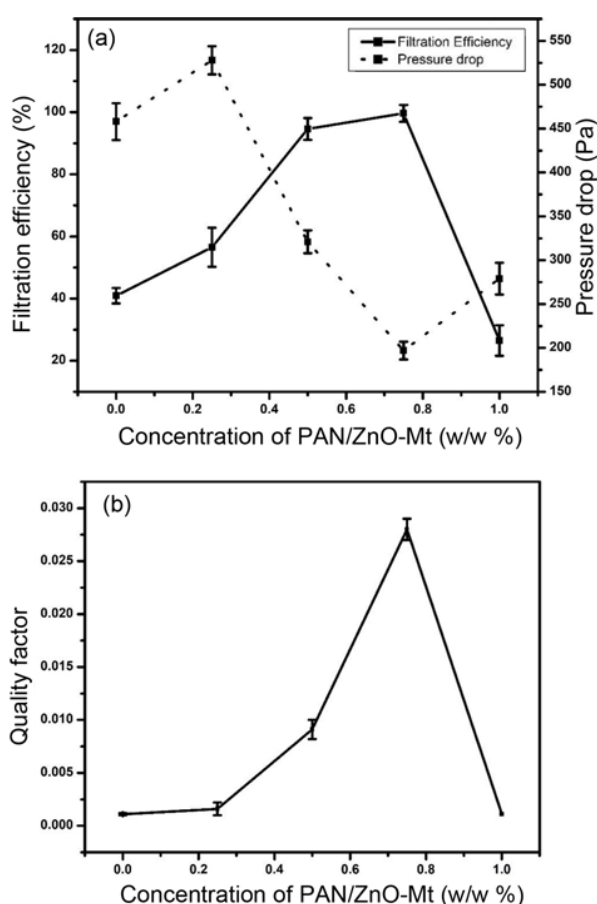
To develop a filter for capture of PM2.5 we need to

Table 3. EDX analysis of 0.75 % PAN/ZnO-Mt nanofibrous nanocomposite before and after RSPM test

Elements	Before RSPM (wt %)	After RSPM (wt %)
C	85.12	65.07
N	11.98	27.89
O	1.21	5.99
S	-	0.70
Zn	1.09	0.34

Table 4. Filtration efficiency, pressure drop and quality factor of nanofibrous nanocomposites

S. no.	Nanofibrous nanocomposites	Filtration efficiency (%)	Pressure drop (Pa)	Quality factor
1	PAN	41.6	458	0.0012
2	0.25 % PAN/ZnO-Mt	53.6	528	0.0014
3	0.50 % PAN/ZnO-Mt	92.8	321	0.0082
4	0.75 % PAN/ZnO-Mt	99.4	197	0.0259
5	1.00 % PAN/ZnO-Mt	21.5	279	0.0009

**Figure 9.** (a) Filtration efficiency and pressure drop of nanofibrous nanocomposites and (b) quality factor of nanofibrous nanocomposites.

evaluate its pressure drop (ΔP) and quality factor ($QF = \ln(1 - \eta) / \Delta P$) (Table 4). The lower the pressure drop and higher the quality factor, better is the filtration efficiency of the membrane [33]. The performance of the nanofibrous nanocomposites as a filter is shown in Figure 9. The pressure drop and quality factor of PAN nanofibrous membrane was found to be 458 Pa and 0.0011 Pa^{-1} . It was observed that as the concentration of ZnO-Mt in nanofibrous nanocomposites increased from 0.25 % to 0.75 % the pressure drop of the membrane decreased from 528 Pa to 197 Pa and further

increased to 279 Pa as the concentration increases to 1.00 %. As the surface roughness of the fiber increases, it provides larger number of sites for adsorption of PM_{2.5}, thereby increasing the filtration efficiency and decreasing its pressure drop (Figure 9a). The quality factor of the nanofibrous nanocomposites was found to increase with increase in concentration of ZnO-Mt from 0.0014 Pa^{-1} to 0.0259 Pa^{-1} and further decreased with 1.00 % PAN/ZnO-Mt nanofibrous nanocomposites (Figure 9b). The increase in surface roughness of the nanofibrous nanocomposites with increase in concentration results in an increase in filtration efficiency of the nanofibrous nanocomposites with a lower pressure drop and higher quality factor [33]. The smaller size of the nanofibrous membranes helps in collision of aerosol particles with the surface of the nanofibrous membrane which results in deposition of the particles over the surface of the nanofibrous membrane and in turn increasing the filtration efficiency [53]. The filtration efficiency increases from 41.6 % for PAN nanofibers to 99.4 % for 0.75 % PAN/ZnO-Mt nanofibrous nanocomposites. The increase in concentration to 1.00 % decreases the filtration efficiency to 21.5 %. A saturation point is achieved and higher concentration does not enhance the filtration performance [54]. On the basis of pressure drop, quality factor and filtration efficiency, 0.75 % PAN/ZnO-Mt nanofibrous nanocomposites were found to be suitable to be used as a filter membrane.

Antimicrobial Study of PAN/ZnO-Mt Nanofibrous Nanocomposites

The antibacterial activity of the nanofibrous nanocomposites was studied against Gram positive *S. aureus* and Gram negative *E. coli* bacterial strains. The zone of inhibition was absent around the nanofibrous mats and no growth of bacteria was observed beneath PAN/ZnO-Mt nanofibrous nanocomposites suggesting antibacterial activity in the nanofibrous mats having non-leaching or barrier mechanism (i.e., negligible migration of ZnO-Mt from the nanofibrous nanocomposites). 0.75 % PAN/ZnO-Mt nanofibrous nanocomposites showed excellent antibacterial activity of 96.23 % and 98.85 % against *E. coli* and *S. aureus* respectively (Table 5).

ZnO has been used at micro and nanoscale as an antibacterial agent. When the size of the particle reduces to nanoscale, it comes in interaction with the surface and/or core of the bacteria, entering the cell and exhibits significant bactericidal activity [55]. The antibacterial activity of ZnO nanoparticles depends on several factors such as size of the particle, concentration of solution, morphology, surface modification and defects [56,57]. Various mechanisms have been reported for the antibacterial activity of ZnO nanoparticles namely, destruction of bacterial cell integrity when ZnO nanoparticles comes in direct contact with the bacterial cell wall, release of antimicrobial Zn^{2+} ions and formation of reactive oxygen species [58,59].

The clay modified with ZnO can easily penetrate inside

Table 5. Antimicrobial activity of PAN/ZnO-Mt nanofibrous nanocomposites

S. no.	PAN/ZnO-Mt nanofibrous nanocomposites	<i>Escherichia coli</i>		<i>Staphylococcus aureus</i>	
		CFU/ml $\times 10^5$	Antibacterial activity (%)	CFU/ml $\times 10^5$	Antibacterial activity (%)
1	Control	504.00 \pm 5.3	-	438.00 \pm 7.9	-
2	0.25 %	180.38 \pm 2.9	64.21 \pm 3.6	67.23 \pm 3.8	84.65 \pm 2.8
3	0.50 %	138.30 \pm 3.1	72.56 \pm 2.1	45.55 \pm 4.6	89.60 \pm 1.5
4	0.75 %	19.00 \pm 3.6	96.23 \pm 0.7	5.00 \pm 2.6	98.85 \pm 0.6
5	1.00 %	46.82 \pm 2.7	90.71 \pm 0.8	35.61 \pm 2.5	91.87 \pm 0.9

the bacterial cell wall due to its small size, destructing the cell integrity. When ZnO-Mt is incorporated in the polyacrylonitrile nanofibrous mat, stacked structure of the clay further breakdowns providing more surface to ZnO to interact with bacterial cell wall. In this case ZnO has not leached out/released in the agar medium from polymeric matrix, the bacterial cell which comes in contact of the fibrous matrix gets killed. The non-leaching phenomenon of ZnO nanoparticle from the nanofibrous surface may be due to formation of weak complex acrylonitrile with Zn cation or $\text{CH}_2\text{CH}(\text{CN})^- \text{Zn}$. It was observed that ZnO nanoparticles incorporated alginate/polyvinyl alcohol nanofibrous nanocomposite diffuse out from the matrix and show zone of inhibition. It was also mentioned that the release of higher concentration of zinc oxide (2 %) highly toxic to the cells. However, the bacterial cells when comes in contact with the surface of the fibrous matrix, their growth gets inhibited [60].

Conclusion

Electrospun PAN/ZnO-Mt nanofibrous nanocomposite membranes have been successfully developed in this study. The addition of ZnO-Mt into PAN nanofibrous membranes increased its ability to capture PM_{2.5} from the atmosphere. The nanofibrous nanocomposites were found to be thermally stable and the thermal stability increases on addition of ZnO-Mt to PAN nanofibrous membranes. PAN/ZnO-Mt nanofibrous nanocomposites were found to be effective against Gram positive and Gram negative bacterial strains. The antimicrobial efficiency increased with an increase in concentration of ZnO-Mt. 0.75 % PAN/ZnO-Mt nanofibrous nanocomposite was found to have 96.23 % and 98.25 % antibacterial activity against *E. coli* and *S. aureus* bacterial strains. Thus, 0.75 % PAN/ZnO-Mt nanofibrous nanocomposites can be a promising filter membrane for PM_{2.5} and microorganisms. Thereupon the nanofibrous nanocomposites serves as a bifunctional membrane which can simultaneously capture PM_{2.5} and inhibit the growth of microorganisms and can also provide future approach for its development to be an effective antimicrobial air filter.

Acknowledgements

The authors thankfully acknowledge the financial support

received from Science and Engineering Research Board (SERB), Department of Science and Technology (DST), Govt. of India (EMR/2017/002833) and Prof. S K Singh and Dr. Rajeev Mishra from Department of Environment Engineering, DTU for particulate matter filtration efficiency test.

Electronic Supplementary Material (ESM) The online version of this article (doi: 10.1007/s12221-021-0914-0) contains supplementary material, which is available to authorized users.

References

1. J. S. Apte, J. D. Marshall, A. J. Cohen, and M. Brauer, *Environ. Sci. Technol.*, **49**, 8057 (2015).
2. M. Xie, M. P. Hannigan, and K. C. Barsanti, *Environ. Sci. Technol.*, **48**, 9053 (2014).
3. L. P. Naeher, K. R. Smith, B. P. Leaderer, L. Neufeld, and D. T. Mage, *Environ. Sci. Technol.*, **35**, 575 (2001).
4. M. C. Turner, D. Krewski, C. A. Pope, Y. Chen, S. M. Gapstur, and M. J. Thun, *Am. J. Respir. Crit. Care Med.*, **184**, 1374 (2011).
5. L. Jing, K. Shim, C. Y. Toe, T. Fang, C. Zhao, R. Amal, K. N. Sun, J. H. Kim, and Y. H. Ng, *ACS Appl. Mater. Interfaces*, **8**, 7030 (2016).
6. M. H. Mohraz, F. Golbabaie, I. J. Yu, M. A. Mansournia, A. S. Zadeh, and S. F. Dehghan, *Int. J. Environ. Sci. Technol.*, **16**, 681 (2019).
7. J. Douwes, P. Thorne, N. Pearce, and D. Heederik, *Ann. Occup. Hyg.*, **47**, 187 (2003).
8. K. Desai, K. Kit, J. Li, P. Michael Davidson, S. Zivanovic, and H. Meyer, *Polymer (Guildf)*, **50**, 3661 (2009).
9. L. Chen, Ph.D. Thesis, MIT, Cambridge MA, 2009.
10. H. Liu, C. Cao, J. Huang, Z. Chen, G. Chen, and Y. Lai, *Nanoscale*, **12**, 437 (2020).
11. S. Yan, Y. Yu, R. Ma, and J. Fang, *Polym. Adv. Technol.*, **30**, 1635 (2019).
12. M. Cao, F. Gu, C. Rao, J. Fu, and P. Zhao, *Sci. Total Environ.*, **666**, 1011 (2019).
13. S. A. Hosseini and H. V. Tafreshi, *Powder Technol.*, **201**, 153 (2010).
14. X. Mao, Y. Si, Y. Chen, L. Yang, F. Zhao, B. Ding, and J. Yu, *RSC Adv.*, **2**, 12216 (2012).

15. S. Zhang, H. Liu, X. Yin, J. Yu, and B. Ding, *ACS Appl. Mater. Interfaces*, **8**, 8086 (2016).
16. J. Choi, B. J. Yang, G. N. Bae, and J. H. Jung, *ACS Appl. Mater. Interfaces*, **7**, 25313 (2015).
17. Y. Liu, M. Park, B. Ding, J. Kim, M. El-Newehy, S. S. Al-Deyab, and H. Y. Kim, *Fiber. Polym.*, **16**, 629 (2015).
18. C. Liu, P. C. Hsu, H. W. Lee, M. Ye, G. Zheng, N. Liu, W. Li, and Y. Cui, *Nat. Commun.*, **6**, 6205, (2015).
19. S. Lee, A. R. Cho, D. Park, J. K. Kim, K. S. Han, I. J. Yoon, M. H. Lee, and J. Nah, *ACS Appl. Mater. Interfaces*, **11**, 2750 (2019).
20. R. Balgis, C. W. Kartikowati, T. Ogi, L. Gradon, L. Bao, K. Seki, and K. Okuyama, *Chem. Eng. Sci.*, **137**, 947 (2015).
21. B. Wang, Z. Sun, Q. Sun, J. Wang, Z. Du, C. Li, and X. Li, *Environ. Pollut.*, **249**, 851 (2019).
22. M. Hashmi, S. Ullah, and I. S. Kim, *J. CRBIOT*, **1**, 1 (2019).
23. Y. Wang, X. Zhao, L. Duan, F. Wang, H. Niu, W. Guo, and A. Ali, *Mater. Sci. Semicond. Processing*, **29**, 372 (2015).
24. A. B. Djurišić, Y. H. Leung, W. C. H. Choy, K. W. Cheah, and W. K. Chan, *Appl. Phys. Lett.*, **84**, 2635 (2004).
25. Z. Fan and J. G. Lu, *J. Nanosci. Nanotechnol.*, **5**, 1561 (2005).
26. A. Sirelkhatim, S. Mahmud, A. Seenii, N. H. M. Kaus, L. C. Ann, S. K. M. Bakhori, H. Hasan, and D. Mohamad, *Nano-Micro Lett.*, **7**, 219 (2015).
27. J. P. Kumar, P. V. R. K. Ramacharyulu, G. K. Prasad, and B. Singh, *Appl. Clay Sci.*, **116-117**, 263 (2015).
28. A. Zyoud, W. Jondi, N. AlDaqqah, S. Asaad, N. Qamhie, A. R. Hajamohideen, M. H. S. Helal, H. Kwon, and H. S. Hilal, *Solid State Sci.*, **74**, 131 (2017).
29. C. M. Srivastava and R. Purwar, *Mater. Sci. Eng. C*, **68**, 276 (2016).
30. A. Sampling, "EPAM-5000", <https://www.skinc.com/catalog/pdf/instructions/1516.pdf> (Accessed April 15, 2021).
31. R. Purwar, P. Mishra, and M. Joshi, *AATCC Rev.*, **8**, 36 (2008).
32. R. Purwar, K. S. Goutham, and C. M. Srivastava, *Fiber. Polym.*, **17**, 1206 (2016).
33. R. Al-Attabi, Y. Morsi, W. Kujawski, L. Kong, J. A. Schütz, and L. F. Dumée, *Sep. Purif. Technol.*, **215**, 500 (2018).
34. S. Fotiadou, C. Karageorgaki, K. Chrissopoulou, K. Karatasos, I. Tanis, D. Tragoudaras, B. Frick, and S. H. Anastasiadis, *Macromolecules*, **46**, 2842 (2013).
35. C. Philippe and B. Claire in "Rubber Nanocomposites: Preparations, Properties and Applications" (S. Thomas and R. Stephen Eds.), pp.353-390, Wiley, Chichester, 2010.
36. S. Ahmadzadeh, A. Nasirpour, J. Keramat, and S. Desobry, *Cellulose*, **22**, 1829 (2015).
37. S. Sikkantkar, S. Karthikeyan, S. Selvasekarapandian, D. V. Pandi, S. Nithya, and C. Sanjeeviraja, *J. Solid State Electrochem.*, **19**, 987 (2015).
38. R. Al-Attabi, L. F. Dumée, J. A. Schütz, and Y. Morsi, *Sci. Total Environ.*, **625**, 706 (2018).
39. N. Bhardwaj and S. C. Kundu, *Biotechnol. Adv.*, **28**, 325 (2010).
40. R. Al-Attabi, L. F. Dumée, L. Kong, J. A. Schütz, and Y. Morsi, *Adv. Eng. Mater.*, **20**, 1700572 (2018).
41. H. J. Haroosh, D. S. Chaudhary, and Y. Dong, *J. Appl. Polym. Sci.*, **124**, 3930 (2012).
42. S. Almuhammed, M. Bonne, N. Khenoussi, J. Brendle, L. Schacher, B. Lebeau, and D. C. Adolphe, *J. Ind. Eng. Chem.*, **35**, 146 (2015).
43. M. Y. Haddad and H. F. Alharbi, *J. Appl. Polym. Sci.*, **136**, 47209 (2019).
44. X. Hou, X. Yang, L. Zhang, E. Waclawik, and S. Wu, *Mater. Des.*, **31**, 1726 (2010).
45. R. Jalili, M. Morshed, and S. A. H. Ravandi, *J. Appl. Polym. Sci.*, **101**, 4350 (2006).
46. S. F. Fennessey and R. J. Farris, *Polymer*, **45**, 4217 (2004).
47. Z. M. Wang in "Nanotechnology in Textiles Theory and applications", 1st ed. (R. Mishra and J. Militky), pp.311-351, Elsevier Ltd., 2010.
48. K. Stevens and M. Fuller in "Textile-led Design for the Active Ageing Population", 1st ed. (J. McCann and D. Bryson), pp.117-138, Elsevier Ltd., 2015.
49. R. Roche and F. Yalcinkaya, *ChemistryOpen*, **8**, 97 (2019).
50. J. Wang, Y. Li, H. Tian, J. Sheng, J. Yu, and B. Ding, *RSC Adv.*, **4**, 61068 (2014).
51. R. Zhang, B. Liu, A. Yang, Y. Zhu, C. Liu, G. Zhou, J. Sun, Po-Chun Hsu, W. Zhao, D. L. Y. Liu, A. Pei, J. Xie, W. Chen, J. Xu, Y. Jin, T. Wu, X. Huang, and Y. Cui, *Nano Letters*, **18**, 1130 (2018).
52. H. Liu, J. Huang, J. Mao, Z. Chen, G. Chen, and Y. Lai, *iScience*, **19**, 214 (2019).
53. R. Al-Attabi, Y. Morsi, J. A. Schütz, and L. F. Dumée, *Sci. Total Environ.*, **647**, 725 (2019).
54. J. Mao, Y. Tang, Y. Wang, J. Huang, X. Dong, Z. Chen, and Y. Lai, *iScience*, **16**, 133 (2019).
55. J. T. Seil and T. J. Webster, *Int. J. Nanomedicine*, **7**, 2767 (2012).
56. L. C. Ann, S. Mahmud, S. K. M. Bakhori, A. Sirelkhatim, D. Mohamad, H. Hasan, A. Seenii, and R. A. Rahman, *Appl. Surf. Sci.*, **292**, 405 (2014).
57. N. Jones, B. Ray, K. T. Ranjit, and A. C. Manna, *FEMS Microbiol. Lett.*, **279**, 71 (2008).
58. M. Li, L. Zhu, and D. Lin, *Environ. Sci. Technol.*, **45**, 1977 (2011).
59. R. Jalal, E. K. Goharshadi, M. Abareishi, M. Moosavi, A. Yousefi, and P. Nancarrow, *Mater. Chem. Phys.*, **121**, 198 (2010).
60. K. T. Shalumon, K. H. Anulekha, S. V. Nair, S. V. Nair, K. P. Chennazhi, and R. Jayakumar, *Int. J. Biol. Macromol.*, **49**, 247 (2011).



Development of Predictive Model for Surface Roughness Using Artificial Neural Networks

Nikhil Rai, Ms Niranjan, Prateek Verma and Prince Tyagi

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

April 28, 2021

DEVELOPMENT OF PREDICTIVE MODEL FOR SURFACE ROUGHNESS USING ARTIFICIAL NEURAL NETWORKS

Nikhil Rai¹, M. S. Niranjana², Prateek Verma³, Prince Tyagi⁴

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

² Department of Mechanical Engineering ,Delhi Technological University,
Delhi,India

³Department of Mechanical Engineering ,Delhi Technological University,
Delhi,India

⁴Department of Mechanical Engineering ,Delhi Technological University,
Delhi,India

Abstract. The need for quality products has been a constant driving force for manufacturing industries. The surface properties are the determinant of the product quality. Surface roughness prediction is now an area of interest in the machining industry. Feed rate, speed of cutting and depth of cut are some of the parameters that influence the prediction of surface roughness. The combined effect of all the three parameters influences the surface roughness to a much significant extent. Data driven prediction is the way ahead. In this study, an Artificial Neural Network is developed fusing the speed of cutting, feed rate and depth of cut. The ANN model is trained using the experimental data already present in the research papers for the prediction as well as optimisation of parameters in CNC lathe for the least possible value of surface roughness of mild steel using statistical techniques and regression models. Further, the ANN model is validated on the basis of two other unseen sets of experimental data on mild steel. From the validation it has been found that prediction of surface roughness by the ANN has higher accuracy as compared to other existing methods.

Keywords: Mild Steel, Surface Roughness, Artificial Neural Networks.

1 Introduction

Engineering application utility of metallic components depends heavily on their surface roughness values. In the present scenario of heavy industrialization and emerging automation techniques, the importance of surface properties have increased manifold making it a determining factor of product quality. Good surface finish is appreciable for improved tribological properties and enhanced resistance to corrosion. Machining and finishing operations on the automated CNC lathes render components variable surface roughness values depending on the process parameters like speed of cutting, feed rate and depth of cut. Researchers have created various prediction models for proper planning and control of cutting conditions and figuring out the optimal parameters for machining.

The consideration of machining parameters becomes crucial so as to perform economical machining with the desired characteristics in the product. The surface integrity produced after machining is recognised to have a great impact on the lifecycle of the product. It represents the nature of the surface condition of the workpiece after machining. In today's dynamically changing world, manufacturing industries are relying more and more on application of optimisation methods in the metal cutting process so that production units can perform optimally

under the rigorous competition pressure in the market and produce products of superior quality.

This research paper is focussed on figuring out the optimal combination of speed of cutting, feed rate and depth of cut to minimise the surface roughness in a CNC lathe turning operation using the ANN technique.

In this work, we have tried to compute the influence of speed of cutting, feed rate and depth of cut on surface roughness and an optimisation model has been created using the artificial neural network technique. Data has been collected from various published papers to train the model and validate the results.

2 Literature Review

A lot of work has been done to optimise the input variable parameters of machining. Residual stress developed during machining impacts the life time and quality of machined components. Neural network based prediction models are used to predict the accuracy of Residual stress development. ANN-FPA models prediction had the accuracy of 99.8% and 99.7% respectively[1]. Surface Roughness prediction is done using Convolution Neural Networks directly from the digital image of surface texture of the machined component instead of doing feature extraction and image segmentation by the virtue of image segmentation[2]. Good Material Removal Rate enhances productivity of machining. Tool chatter degrades MRR[3]. Optimal cutting parameters are predicted using ANN for the stable machining operation in turn increasing the productivity. Accurate prediction of tool life prevents the catastrophic stoppage of machining processes due to tool wear. ANN models are used for the prediction of tool life and cutting edge wear to make it industry ready[4]. Experiments were performed that are designed on the basis of Taguchi's methodology for optimal result. The study demonstrates that the surface roughness increases when the feed rate is increased, the influence of cutting speed was found to be less than that of feed followed by the effect of change in depth of cut[5]. Another researcher has studied the influence of tool overhang along with other parameters on residual stress and surface roughness developed during the turning of aluminium alloy by designing the experiments based on Taguchi's technique. The results obtained reveal that the most optimal result for surface roughness could be obtained by using tool overhang in the medium or lower range[6]. Regression models were developed for predicting the surface roughness and an artificial neural network to account the combined influences of the tool vibration amplitudes and cutting force which provides a model with higher accuracies of prediction[7]. A comparative study has been established with Aluminium alloys and Brass machining on Computer Numerical Control machine and analysed by the help of prediction techniques. It was found that surface properties are dependent on cutting force which is ultimately decided by speed of cutting, feed rate and depth of cut[8]. Some other researchers have also used the taguchi technique. Experiments were conducted by taking Feed Rate, Speed of Cutting & Depth of the cut as cutting process parameters. Experiments are designed on the basis of Taguchi's technique for optimization using orthogonal arrays. In hindsight it was inferred that speed of cutting highly influences the Surface roughness than feed and in case of MRR, depth of cut is the

primary parameter and then the speed of cutting[9].Artificial neural networks are being developed for the prediction of tool life , failure-mode on the basis data obtained by recording different experiments based on multiple values of speed of cutting and feed rate and constant depth of cut. The neural network best predicted the failure-mode prediction. the network training could be improved using the real time data sets[10].

An artificial neural network is developed for predicting so as to control the surface roughness in a computer numerically controlled lathe. Experiments were conducted and the cutting parameters were the speed of cutting, , feed rate and depth of cut. It is found that we can predetermine optimised parameters of cutting for surface roughness of machining operation using the control algorithm and artificial neural network[11].An on-line fuzzy neural network (FNN) model[12] to estimate the flank and crater wear on the basis of modified least square backpropagation[13].It has been found that an on-line FNN model has great accuracy for the estimation of progressive flank and crater wear with very less time for computation[14]. A linear model was generated for three responses i.e, Material Removal Rate , Surface Roughness, Chip Thickness Ratio (CTR) and experiments were conducted based upon the Taguchi's technique of optimised response using Orthogonal array. ANOVA was used to find out the main influences of S/N ratio and graphs are plotted. The resulting optimized value for depth of cut, time and the speed of cutting are best fit for the optimised metal cutting to extract the competitive results from commercial mild steel[15].Mathematical models were also developed for predicting surface roughness[16] on the basis of parameters for cutting and tool vibrations. Tool vibrations were measured using an FFT analyzer. It is inferred that tool vibrations and cutting parameters based prediction models are more accurate[17][18]. Multiple Attribute Decision Making methods[19] had been used for investigating multiple parameters and their impacts on surface roughness.The investigation is done to devise an optimised procedure for selection of tool insert for improved surface finish in turning operation while working on different materials..After analyzing the previous research works, it was observed that enormous work had been done on optimising cutting parameters using statistical tools and techniques for better surface finish but this is felt that a lot better predictions can be done by the virtue of neural networks. Here, an attempt is made to train an artificial neural network for the prediction of surface roughness of the mild steel while being machined on CNC lathe. In order to train the model well, experimental readings from other prediction and optimisation based research work are used. In the end the neural network was validated with two unseen data sets to gauge the efficiency of the model.

3 Methodology

Artificial Neural Network (ANN) algorithms are modern information processing models used to make approximations from real objective functions. The algorithm has taken the inspiration from working of neural cells in the human brain. Artificial neural networks have the scope of modelling linear and non-linear systems. A trained neural network depicts a quick mapping of the given input with the expected output quantities. We have incorporated this modern technique to

quantify the effect of process parameters on surface properties of the material during turning operations on CNC lathe machine tools.

An artificial neural network is represented as an acyclic graph. Different sets of nodes comprise different layers.

Mainly there are three categories of layers

Input layer

Different formats of inputs provided by the programmer.

Hidden layer

It is responsible for the calculations to figure out the hidden features and patterns in the data.

Output layer

Sequential transformation is done on the input received in accordance with the functions of hidden layers and the finally obtained result is conveyed by the virtue of the output layer.

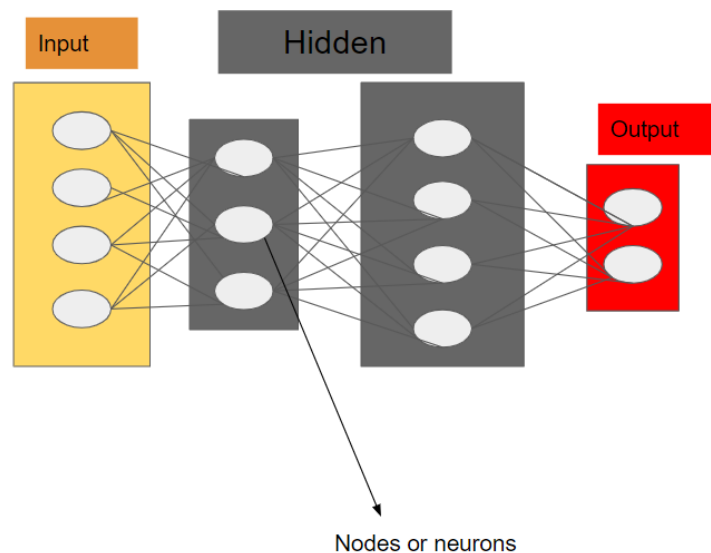


Fig. 1. A perceptron in the form of acyclic graph

ANN calculates the weighted sum of the inputs and adds the bias effect. Following is the representation of the transfer function.

$$\sum W_i * X_i + b(1)$$

The determined weighted sum is passed to an activation function. Activation functions decide whether a node should fire or not. Only those who are fired reach the output layer. Ample activation functions are available that can be applied to the tasks that we perform.

ANN is significantly powerful computer modelling techniques which is being used these days in multiple engineering fields for modelling of

complicated relationships which are difficult to optimise using traditional techniques. Neural networks gain information by detection of patterns in the data and are trained for futuristic predictions. The proposed system is based on the ANN training technology to optimize the machining parameters.

4 Data Set

Table 1 : Experimental data from for Depth of Cut (DOC), feed rate (FR) , Cutting Speed (CS) , and resulting surface roughness on machining of mild steel on CNC lathe.[5]

S. No	CS (mm/min)	FR (mm/rev)	DOC (mm)	SR (μm)	S/N Ratio(dB)
1	60	0.25	0.2	5.6	-14.96
2	60	0.25	0.3	7.1	-17.02
3	60	0.25	0.4	7.4	-17.38
4	60	0.35	0.2	7.1	-17.02
5	60	0.35	0.3	6.03	-15.6
6	60	0.35	0.4	6.98	-16.87
7	60	0.45	0.2	4.85	-13.71
8	60	0.45	0.3	5.55	-14.88
9	60	0.45	0.4	6.31	-16
10	80	0.25	0.2	4.23	-12.52
11	80	0.25	0.3	4.44	-12.94
12	80	0.25	0.4	5.14	-14.21
13	80	0.35	0.2	3.84	-11.68
14	80	0.35	0.3	5.57	-14.91
15	80	0.35	0.4	5.73	-15.16
16	80	0.45	0.2	4.06	-12.17
17	80	0.45	0.3	4.85	-13.71
18	80	0.45	0.4	6.28	-15.95
19	100	0.25	0.2	4.12	-12.29
20	100	0.25	0.3	3.57	-11.05
21	100	0.25	0.4	3.3	-10.37
22	100	0.35	0.2	3.41	-10.65
23	100	0.35	0.3	3.12	-9.88
24	100	0.35	0.4	3.42	-10.68
25	100	0.45	0.2	2.63	-8.39
26	100	0.45	0.3	4.33	-12.72
27	100	0.45	0.4	4.1	-12.25

Table-2 : Experimental data from for Depth of Cut (DOC) ,, feed rate (FR) , Cutting Speed (CS) and resulting surface roughness on machining of mild steel on CNC lathe[9].

EXP.	FR (mm/rev)	DOC (mm)	CS (mm/min)	SR (μm)
1	0.1	0.5	75	1.464
2	0.1	0.75	125	2.062
3	0.1	1	175	2.972
4	0.2	0.5	125	3.284
5	0.2	0.75	175	4.264
6	0.2	1	75	2.22
7	0.3	0.5	175	3.662
8	0.3	0.75	75	2.549
9	0.3	1	125	3.586

Table-3: Experimental data from for depth of Cut (DOC) , Speed and resulting surface roughness on machining of mild steel on CNC lathe [15] .

SPEED (revolutions/min)	TIME	DOC (mm)	SR (μm)	CHIP THICKNESS RATIO	MRR
2000	8	1	0.86	1.08	30
2000	8.2	1.3	0.02	1.12	38.05
2000	8.3	1.5	1.5	1.15	43.37
1500	8	1	1.6	1.09	30
1500	8.2	1.3	0.42	1.1	34.67
1500	8.3	1.5	0.47	1.14	45
900	8	1	0.38	1.09	21.81
900	8.2	1.3	6.18	1.13	28.36
900	8.3	1.5	2.72	1.15	30

5 Results and Discussions

Table 1 is used for training the neural network while Table 2 and 3 are the unseen data sets for validating the neural networks. The neural network architecture is made of three dense layers each containing 250,100,100 neurons and the finally an output layer. Relu activation[20] was applied in each layer. The network was trained using SGD and was validated on 10% of this data during training time. The feed rate , speed of cutting , depth of cut was provided as input to the neural network while Surface roughness was predicted on the basis of these inputs. The network was evaluated on the basis of MSE, MAE, MAPE error and

was trained on 2000 epochs during which the model finally converged. Final training errors are listed in Table 4 as shown below.

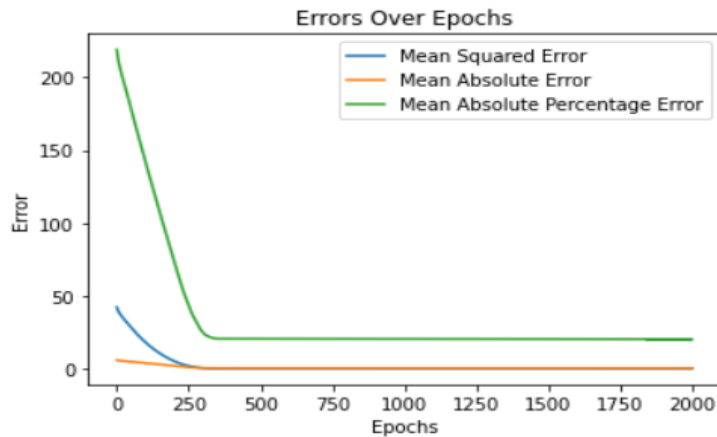


Fig. 2. Error Vs Epochs Curve During training time

The network was tested on two unseen datasets each containing 10 data points. The errors during testing training and validation are quite close to each other hence the model is able to learn patterns in the data really well during training period.

The loss for network is defined as squared difference summation of predicted and truth values of surface roughness over all data points divided by total number of data points.

During training period we have received mse , mae ,mape loss as 0.3415, 0.5293, 20.2629 respectively while during testing itis 0.5654, 0.6178, 38.5695 and 1.0916, 1.1089, 97.7567 respectively for unseen dataset 1 and 2.From these results we can clearly see that neural network is capable of predicting surface roughness on the basis of feed rate , speed of cutting , depth of cut.

6 Conclusion

In this study, we have used the ANN model for CNC turning. The artificial neural network was trained upon 27 data points with parameters feed rate , speed of cutting , depth of cut for the corresponding surface roughness. Further the data set is validated upon 18 unseen data points. It's been found that unseen data values of surface roughness and predicted values of surface roughness are significantly close. The predicted values gives us a mean absolute error of 0.86. In conclusion there is very close agreement between predicted and actual surface roughness value.The prediction of surface roughness as done using the ANN algorithm has shown comparatively better results than the existing models and hence can be relied upon for further prediction for industrial standards application.

Table 4 : Errors During Training ,Validation and Testing Period

	MSE	MAE	MAPE
Training	0.3415	0.5293	20.2629
Validation	0.7040	0.8391	23.3986
Test On Dataset-1 On Dataset-2	0.5654	0.6178	38.5695
	1.0916	1.1089	97.7567

We can know about the surface roughness on a mild steel upon selecting feed rate , speed of cutting , depth of cut which in turn will help tremendously in the decision making for getting a high standard surface finish.Further, the study can be integrated with optimization algorithms like genetic modelling for optimisation of the multiple turning parameters to enable better parameter selection on the CNC lathe to produce quality products in the competitive market landscape.

References

1. Khoshaim, Ahmed B., et al. "Prediction of Residual Stresses in Turning of Pure Iron using Artificial Intelligence-based Methods." *Journal of Materials Research and Technology* (2021).
2. Rifai, A. P., Aoyama, H., Tho, N. H., Dawal, S. Z. M., & Masruroh, N. A. (2020). Evaluation of turned and milled surfaces roughness using convolutional neural network. *Measurement*, 161, 107860.
3. Gupta, Pankaj, and Bhagat Singh. "Local mean decomposition and artificial neural network approach to mitigate tool chatter and improve material removal rate in turning operation." *Applied Soft Computing* 96 (2020): 106714.
4. Mikołajczyk, T., Nowicki, K., Bustillo, A., & Pimenov, D. Y. (2018). Predicting tool life in turning operations using neural networks and image processing. *Mechanical systems and signal processing*, 104, 503-513..
5. Sharma, Sushil Kumar, and Er Sandeep Kumar. "Optimization of Surface Roughness in CNC Turning of Mild Steel (1018) using Taguchi method." *Carbon* 100 (2014): 0-26.
6. El-Axir, M. H., M. M. Elkhabeery, and M. M. Okasha. "Modeling and parameter optimization for surface roughness and residual stress in the dry turning process." *Engineering, Technology & Applied Science Research* 7.5 (2017): 2047-2055.
7. Vasanth, X. Ajay, P. Sam Paul, and A. S. Varadarajan. "A neural network model to predict surface roughness during turning of hardened SS410 steel." *International Journal of System Assurance Engineering and Management* 11.3 (2020): 704-715.
8. Bharilya, R. K., Malgaya, R., Patidar, L., Gurjar, R. K., & Jha, A. K. (2015). Study of optimised process parameters in turning operation through force dynamometer on CNC machine. *Materials Today: Proceedings*, 2(4-5), 2300-2305.
9. Ezugwu, E. O., S. J. Arthur, and E. L. Hines. "Tool-wear prediction using artificial neural networks." *Journal of materials Processing technology* 49.3-4 (1995): 255-26

10. Goyal, Shivam, Varanpal Singh Kandra, and Prakhar Yadav. "Experimental study of turning operation and optimization of MRR and surface roughness using taguchi method." *Int. J. Innov. Res. Adv. Eng* (2016).
-
11. Karayel, Durmus. "Prediction and control of surface roughness in CNC lathe using artificial neural network." *Journal of materials processing technology* 209.7 (2009): 3125-3137.
12. Wang, Ning, Meng Joo Er, and Xian Yao Meng. "A fast and accurate online self-organizing scheme for parsimonious fuzzy neural networks." *Neurocomputing* 72.16-18 (2009): 3818-3829.
13. LeCun, Y., Touresky, D., Hinton, G., & Sejnowski, T. (1988, June). A theoretical framework for back-propagation. In *Proceedings of the 1988 connectionist models summer school* (Vol. 1, pp. 21-28).
14. Chungchoo, C., and D. Saini. "On-line tool wear estimation in CNC turning operations using fuzzy neural network model." *International Journal of Machine Tools and Manufacture* 42.1 (2002): 29-40.
15. Jaiganesh, V., Yokesh Kumar, B., Sevvell, P., & Balaji, A. J. (2018). Optimization of process parameters on commercial mild steel using Taguchi technique. *International Journal of Engineering & Technology*, 7(11), 138-142.
16. Tse, R., and D. M. Cruden. "Estimating joint roughness coefficients." *International journal of rock mechanics and mining sciences & geomechanics abstracts*. Vol. 16. No. 5. Pergamon, 1979.
17. Abouelatta, O. B., and J. Madl. "Surface roughness prediction based on cutting parameters and tool vibrations in turning operations." *Journal of materials processing technology* 118.1-3 (2001): 269-277.
18. Rahman, M. Z., Das, A. K., Chattopadhyaya, S., Reyaz, M., Raza, M. T., & Farzeen, S. (2020). Regression modeling and comparative analysis on CNC wet-turning of AISI-1055 & AISI-4340 steels. *Materials Today: Proceedings*, 24, 841-850..
19. Tzeng, Gwo-Hshiung, and Jih-Jeng Huang. *Multiple attribute decision making: methods and applications*. CRC press, 2011
20. Li, Yuanzhi, and Yang Yuan. "Convergence analysis of two-layer neural networks with relu activation." *arXiv preprint arXiv:1705.09886* (2017).
21. Taka, M., Raygor, S. P., Purohit, R., & Parashar, V. (2017). Selection of tool and work piece combination using Multiple Attribute Decision Making Methods for Computer Numerical Control turning operation. *Materials Today: Proceedings*, 4(2), 1199-1208.
22. Kumar, M. Vijay, BJ Kiran Kumar, and N. Rudresha. "Optimization of machining parameters in CNC turning of stainless steel (EN19) by Taguchi's orthogonal array experiments." *Materials Today: Proceedings* 5.5 (2018): 11395-11407.

Dielectric Modulated Junctionless Biotube FET (DM-JL-BT-FET) Bio-Sensor

Anubha Goel, Sonam Rewari, Seema Verma, S.S. Deswal and R.S. Gupta, Life Senior Member, IEEE

Abstract— In this manuscript, an analytical model has been demonstrated for Dielectric Modulated Junctionless Biotube FET (DM-JL-BT-FET) as a sensor. The Junctionless Biotube FET based sensor has been compared with Nanowire FET under similar biomolecule conditions. It has been demonstrated that Dielectric Modulated Junctionless Biotube FET shows much higher efficiency in Bio-sensing and poses superior device performance characteristic in terms of higher sensitivity, higher drift in drain current, transconductance, I_{on}/I_{off} ratio, Subthreshold Slope and Threshold voltage. The hole concentrations have also been investigated under different biomolecule conditions. Two different biomolecule conditions have been considered in our analysis viz., firstly, varying the biomolecule concentrations and secondly, inserting different biomolecules namely DNA, Biotin and Hydroprotein. Improved bio sensing is observed in Junctionless Biotube FET because of superlative gate control over the channel, owing to architecture of Biotube FET. The analytical results have also been modelled for DM-JL-BT-FET by finding a solution to the 2-D Poisson equation in accordance with the boundary conditions. The analytical results are much in coherence with the results obtained from the simulator.

Index Terms— Bio-sensing; Junctionless; Biotube; Nanowire; Sensitivity.

I. INTRODUCTION

In the analytical biological domain, advances in bio-sensors is one of the most encouraging directives for research. The continuously emerging field of organic-bioelectronics, aims to amalgamate the micro-electronic devices and biological elements for the evolution of an extensive range of portable-analytical devices e.g. DNA Sensors, Protein Sensors and LOC (Lab-on-a-chip) devices efficiently. Reliability in sensing/detecting assorted complex biomolecules, proteins and DNA has gained much significance and has become a baseline benchmark in the field of bio-sensors with the evolution of assorted molecular recognition patterns [1]. Various nano-structures namely NanoWires (NW) [2]-[3], Biotubes (BT) [4]

and NanoCantilevers (NC) [5] have been in the limelight owing to their capacity in use for biomolecule detection.

On a similar note, FET (Field Effect Transistor) type bio-sensors also known as ION-Sensitive FET bio-sensors (ISFET) have shown immense potential in detection of biomolecules with superior sensitivity, superior scalability, CMOS Technology and cost-effective mass production [6]-[10] as some of the key advantages. Betwixt the available FET type bio-sensors brought in thus far, Silicon-based NW Bio-sensors have gained much popularity because of excellent sensitivity [11]-[12]. The main reason behind this excellent sensitivity for Si-based NWs has the larger STVR (Surface-to-Volume-Ratio) and smaller dimensions of the NW structures because of the dimensional compatibility of bio-molecules with nanowires [13]-[14]. The conventional MOSFET suffers from abrupt source drain junction formation, high source drain resistance and difficult fabrication process. These issues can be combated by using Junctionless (JLT) MOSFET with homogeneous source, channel and drain regions ($n^+ - n^+ - n^+$) for NMOS [10]. Researchers have already done extensive work related to analog/RF performance of Junctionless transistor and it has been that Junctionless transistor exhibit superior intrinsic RF scaling capability and better SCEs and HCEs immunity than inversion mode transistor. Due to bulk conduction mechanism, it has less surface scattering than the inversion mode transistor, ideal subthreshold slope, and better linearity [15]-[18]. Due to homogeneous doping across the source, channel and drain, the source drain resistance is automatically reduced. Along with doping techniques and thermal budgeting, creation of precipitous source and drain junctions also shines out as a challenging issue for nano-wires [19]. Lately, to overcome all the shortcoming of the NW structures, a core gate has been added to the NW structure, to further improve the control over the gate, known as Bio-Tube (BT) Architecture [20]-[21]. Also, the benefits of Junctionless structure have been very well acknowledged in the literature, with Channel, Source and Drain regions having high doping concentrations. It has been acknowledged in the literature that Junctionless structures have souped-up electrostatic coupling and receded Short Channel Effects (SCEs) [22]-[23]. The dual tube structure in Junctionless Biotube FET, poses much higher noise immunity along with higher linearity (when applied for wireless applications) [24]. An analytical model demonstrated by Rewari et. al [25] shows that, Junctionless Double Surrounding Gate MOSFET has higher drain current and Transconductance in comparison to Nanowire FET under similar conditions. Marconcini et. al, [26] have shown Hierarchical simulation of transport in silicon nanowire transistors and Luisier et al. [27] have shown the Atomistic full-band simulations of silicon nanowire transistors.

Anubha Goel and R.S Gupta are with the Department of Electronics and Communication Engineering, Maharaja Agrasen Institute of Technology, New Delhi 110086, India. (e-mail: anubhagoel15@gmail.com, rsgupta1943@gmail.com).

Sonam Rewari is with the Department of Electronics and Communication Engineering, Delhi Technological University, New Delhi 110042, India. (e-mail: rewarisonam@gmail.com).

Seema Verma is with Department of Electronics and Communication, Banasthali University, Niwai, Banasthali, Rajasthan -304022, India.(email: seemaverma3@yahoo.com).

S.S. Deswal is with the Department of Electrical and Electronics Engineering, Maharaja Agrasen Institute of Technology, New Delhi 110086, India. (e-mail: satvirdeswal@hotmail.com).

Graphene transistors have also proven to mitigate the SCE's to a greater extent [28]-[35]. Structurally, FET Bio-sensors have an ion-sensitive lamina, an electrolyte solution and a reference electrode replacing the metal gate in conventional MOSFET [36]-[37]. This gives considerably optimum performance with eminent sensitivity for charged-biomolecules. To further improve the device, Dielectric-Modulated (DM) FET with nano-gap cavities at the source and drain ends was demonstrated [38]-[39]. Structurally, bio-molecules are immobilized inside the carved out nano-gaps (formed by etching a vertical nano-gap) beneath the gate material. DM FET bio-sensors operate on the principle that introduction of bio-molecules in the nano-gap placed towards the edge of the gate dielectric, causes fluctuations in the dielectric constant (gate capacitance) which further modulates the electrical performance parameters like current and threshold voltage and thus aids in detection of bio-molecules [38], [40]-[41]. In the absence of bio-molecules, the nano-gap cavity is filled with air (dielectric constant of nano-gap cavity becomes unity). In presence of bio-molecules, threshold voltage, V_{th} reduces owing to the fluctuating dielectric-constant of the nano-gap cavity. A detailed comparative analysis of the sensitivity performance of the biosensors with simulations based reports on other potential nanowire based DMFETs have been deeply investigated by Kanungo et. al and Shafi et. al [42]-[43]. The aforesaid DM FET bio-sensor proclaims higher sensitivity for neutral bio-molecules and negatively charged bio-molecules such as DNA, hydro protein and biotin [31]. Thus this paper proposes an analytical model of a dielectrically modulated biosensor architecture for the first time by assimilating the advantages of junction-less and nano-tube bio-sensors.

II. DEVICE DESIGN AND SIMULATION

Fig. 1 (a) illustrates the Three-Dimensional view of DM-JL-BT-FET sensor which clearly depicts GATE1 (inner-core-gate) and GATE2 (outer-shell-gate). Both GATE1 and GATE2 have indistinguishable gate-metal-work-function along with SiO_2 as the gate-dielectric. As both gates are indistinguishable, they are devised of platinum with metal-gate-work-function equivalent to 5.55 eV. The drain (n+), source (n+) and channel (n+) are consistently doped with $N_D = 10^{19} \text{ cm}^{-3}$, t_{si} (silicon film thickness) = 20 nm, t_{ox} (thickness of oxide) = 2 nm, ϵ_{ox} (oxide permittivity) = $3.9\epsilon_0$. Dielectric Modulated (DM) Junctionless Biotube FET (JL-BT-FET) detects biotin, hydroprotein and DNA by tweaking the dielectric of the oxide. Oxide is stacked in such a manner that biomolecule oxide is followed by the silicon oxide. Fig. 1(b) pictures the Two-Dimensional view of DM-JL-BT-FET. Fig. 1 (c) illustrates the sectional-view of DM-JL-BT-FET. Fig. 1(d) shows the structure of the biomolecules inserted and Fig. 1 (e) pictures the calibration of the experimental work [44] with the simulated work.

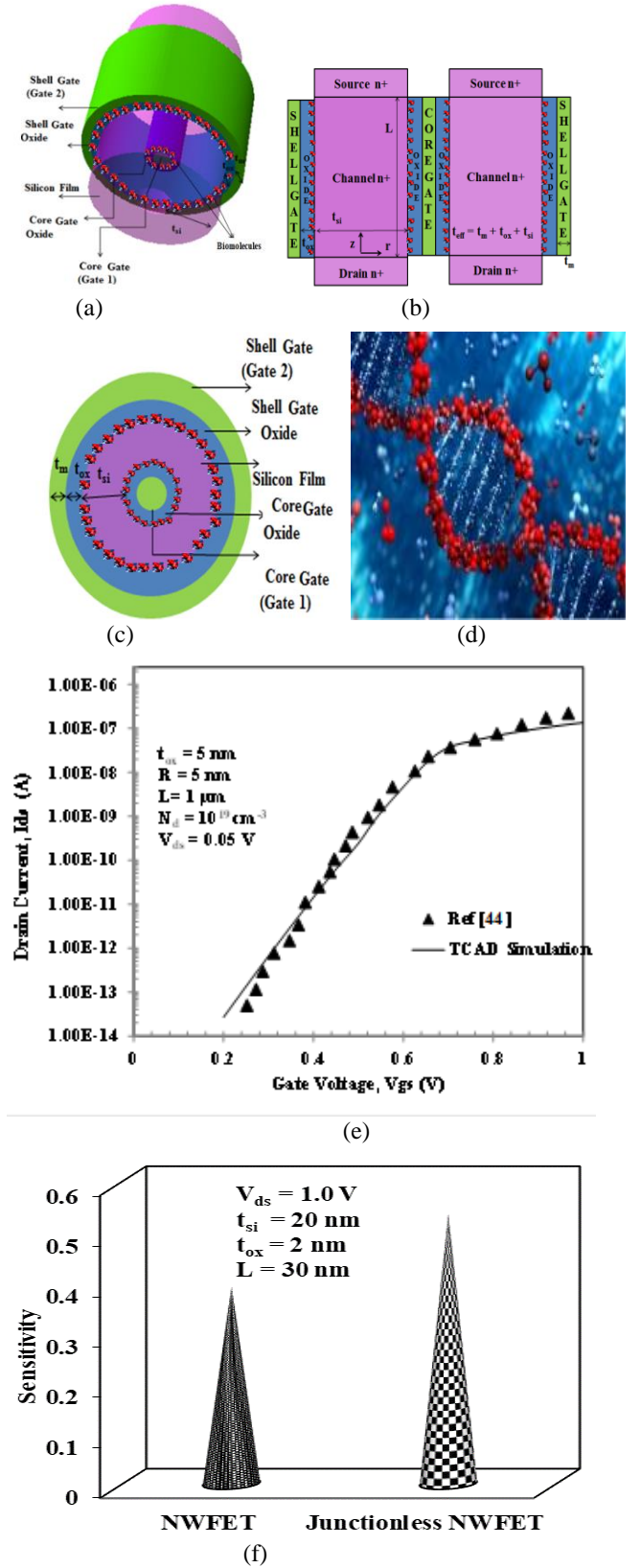


Fig. 1 (a) Three-Dimensional Structure of DM-JL-BT-FET (b) 2-D Correctional Structure of DM-JL-BT-FET (c) Sectional View of DM-JL-BT-FET (d) Biomolecules (e) Calibration with experimental work [44] (f) Sensitivity of NWFET and Junctionless NWFET

Fig 1 (e) shows the Sensitivity Analysis of NWFET and Junctionless NWFET. As it is evident from the figure the sensitivity of Junctionless NWFET is higher than NWFET, it is because of the uniformly doped source, channel and drain regions which reduces the short channel effect of source/drain resistance.

Table 1 tabulates the different device parameters. Silvaco ATLAS 3-D device-simulator [45] has been deployed to realize the device-simulations. The simulation models exploited to realize our simulations along with their descriptions have been tabulated in Table 2.

Table 1. Device Characterization Parameters

Device Specification	DM-JL-BT-FET Sensor	JL-NW-FET
Length of the Channel (nm)	30	30
Doping of the Channel Region (/cm ³)	1x10 ¹⁹	1x10 ¹⁹
Thickness of Silicon (nm)	20	20
Metal-Gate-Work-Function (eV)	5.55	5.13
Source/Drain Length (nm)	15	15
Permittivity of Silicon, ϵ_{si}	3.9 ϵ_0	3.9 ϵ_0
Permittivity of Biotin, ϵ_{Biotin}	5.0 ϵ_0	5.0 ϵ_0
Permittivity of Hydroprotein, $\epsilon_{Hydroprotein}$	2.1 ϵ_0	2.1 ϵ_0
Permittivity of DNA, ϵ_{DNA}	1.0 ϵ_0	1.0 ϵ_0
Thickness of Silicon Oxide, t_{SiO_2} (nm)	1	1
Thickness of Biomolecules, t_{bio} (nm)	1	1
Electron affinity of Si	4.05 eV	4.05 eV
Lattice constant of Si	5.431 Å	5.431 Å
Si-Si Single Bond Length	237×10 ⁻¹² m	237×10 ⁻¹² m
Si-Si Double Bond Length	214×10 ⁻¹² m	214×10 ⁻¹² m
Optical phonon energy of Si	0.063 eV	0.063 eV
Number of atoms in 1 cm ³ of Si	5·10 ²²	5·10 ²²
Auger recombination coefficient of Si, C_n	1.1·10 ⁻³⁰ cm ⁶ s ⁻¹	1.1·10 ⁻³⁰ cm ⁶ s ⁻¹
Auger recombination coefficient of Si, C_p	3·10 ⁻³¹ cm ⁶ s ⁻¹	3·10 ⁻³¹ cm ⁶ s ⁻¹

Table 2. Simulation Models used for Simulations

Simulation Model	Description
Mobility-Model	Lombardi-CVT-Model→Appropriate for non-planar structures along with inversion region modelling.
Recombination Model	SRH (Schottky – Read – Hall) Model → Appropriate for inculcating carrier lifetimes Auger-Model→ Appropriate to inculcate High-Current-Densities with Impact-Ionization.
Concentration Dependent Model	Appropriate to inculcate SRH-Recombination with their lifetimes.
Energy Transport Model	Drift-Diffusion-Model → Appropriate for numerical-techniques.
Statistics	Boltzmann-Model → Appropriate to consider the Carrier-Statistics.

III. ANALYTICAL MODEL

Mathematical interpretations for Subthreshold-Current (I_{sub}) and potential has been realized by solving Two-Dimensional Poisson's equation in cylindrical-coordinates. With appropriate boundary-conditions and deploying the superposition technique [21] the expressions for surface potential as

well as I_{sub} are obtained. Two-Dimensional Poisson's equation can be asserted [19] as:

$$\frac{1}{r} \frac{\partial}{\partial r} \Phi(r, z) + \frac{\partial}{\partial r^2} \Phi(r, z) + \frac{\partial}{\partial z^2} \Phi(r, z) = -\frac{qN_D}{\epsilon_{si}} \quad (1)$$

with, $N_D \rightarrow$ consistent doping concentration and $\phi(r, z) \rightarrow$ potential distribution across the silicon-film. By deploying superposition-technique, the conclusive solution of potential can be disintegrated into: 1. 1-D long-channel solution ($V(r)$) and 2. Two-Dimensional short-channel solution ($U(r, z)$) i.e.

$$\Phi(r, z) = V(z) + U(r, z) \quad (2)$$

Now, equation (1) can be rewritten as 1-D Poisson equation:

$$\frac{1}{r} \frac{\partial}{\partial r} (V(r)) + \frac{\partial^2}{\partial r^2} V(r) = \frac{-q N_D}{\epsilon_{si}} \quad (3)$$

And 2-D Laplace equation as:

$$\frac{1}{r} \frac{\partial}{\partial r} (U(r, z)) + \frac{\partial^2}{\partial r^2} U(r, z) + \frac{\partial^2}{\partial z^2} U(r, z) = 0 \quad (4)$$

Boundary-conditions exploited for solution of 2-D potential $\Phi(r, z)$ are given as follows:

$$(i). \Phi(t, z) = \Phi_{os}(z) \quad (5)$$

$$(ii). \Phi(t - t_{si}, z) = \Phi_{is}(z) \quad (6)$$

$$(iii). \left. \frac{\partial \Phi(r, z)}{\partial r} \right|_{r=t_{eff}} = -\Delta(V_{gs} - \Phi_{os}(z)) \quad (7)$$

$$(iv). \left. \frac{\partial \Phi(r, z)}{\partial r} \right|_{r=t_{eff}-t_{si}} = -\Delta(V_{gs} - \Phi_{is}(z)) \quad (8)$$

$$(v). \Phi(r, L) = V_{bi} \quad (9)$$

$$(vi). \Phi(r, L) = V_{bi} + V_{ds} \quad (10)$$

$$C_{ox} = \frac{2\epsilon_{ox}}{(t_m + t_{oxeff}) \ln(1 + \frac{2t_{oxeff}}{(t_m + t_{oxeff})})} \quad (11)$$

$\Phi_{os}(z) \rightarrow$ consistent outer-surface potential, $\Phi_{is}(z) \rightarrow$ consistent inner-surface potential,

$t_{oxeff} = t_1(SiO_2) + \frac{\epsilon_{ox}}{\epsilon_{bio}} t_2(Bio)$, t_1 (SiO_2) is the thickness of silicon-

oxide layer and t_2 (Bio) depicts the thickness of the biomolecule species, ϵ_{ox} is the dielectric-permittivity of oxide, ϵ_{bio} is the dielectric-permittivity of biomolecules.

The conclusive solution of equation (3) can now be realized as a parabolic approximation:

$$V(r) = P_0 + P_1 r + P_2 r^2 \quad (12)$$

where, $P_0 = V_{gseff} - P_1(t + \frac{1}{\Delta}) - P_2(t^2 + \frac{2t}{\Delta})$, $P_1 = -\delta(t_{eff} - \frac{t_{si}}{2})$

$$P_2 = \delta/2, \delta = \frac{-qN_D}{\epsilon_{si}}, V_{gseff} = V_{gs} - V_{fb} + \frac{qN_f}{C_{ox}}, V_{fb} = \phi_m - (\chi_{Si} - q\phi_f) \rightarrow$$

flat band voltage, $V_{gs} \rightarrow$ applied gate-to-source voltage, $N_f \rightarrow$ biomolecule charged density, $\chi_{Si} \rightarrow$ electron affinity of silicon, $q\phi_f \rightarrow$ channel fermi-potential. The solution of equation (4) can be obtained as:

$$U(r, z) = \sum_{n=1}^{\infty} J_0(\alpha_n r) (M_n \exp(\alpha_n z) + N_n \exp(-\alpha_n z)) \quad (13)$$

with J_0 and $J_1 \rightarrow$ Bessel functions of order 0 and 1 respectively, M_n and N_n are constants calculated using boundary-conditions (9) and (10) and are given in Appendix,

$$\alpha_n \text{ is Eigen values of: } J_1\left(\frac{t_{eff}}{\alpha_n}\right) = \frac{C_{ox}}{\alpha_n \epsilon_{si}} J_0\left(\frac{t_{eff}}{\alpha_n}\right) \quad (14)$$

The 2-D potential can be given as:

$$\Phi(r, z) = V(r) + \sum_{n=1}^{\infty} J_0(\alpha_n r) (M_n \exp(\alpha_n z) + N_n \exp(-\alpha_n z)) \quad (15)$$

$$\text{where } \alpha_n = \frac{1}{\eta_n}.$$

I_{ds} can be formulated [40] as:-

$$I_{ds} = \begin{cases} I_{sub} & \text{for } 0 \leq V_{gs} \leq V_{th} \\ I_{lin} & \text{for } V_{th} \leq V_{gs} \leq V_{dsat} \\ I_{sat} & \text{for } V_{dsat} \leq V_{gs} \leq 1.0 \text{ V} \end{cases} \quad (16)$$

I_{sub} [15] is:-

$$I_{sub} = 2R\pi\mu kT\eta_i \int_0^L \frac{1 - e^{-\frac{qV_{ds}}{kT}}}{\int_0^L e^{\frac{q\Phi(r,z)}{kT}} dr} dz \quad (17)$$

with, Boltzmann's constant, $k = 1.38 \times 10^{-23} \text{ J/K}$, $T = 300\text{K}$, intrinsic-carrier-density, $\eta_i = 1.45 \times 10^{10} \text{ cm}^{-3}$, electron mobility, $\mu = 1300 \text{ cm}^2/\text{Vs}$. I_{lin} [40] is the current in the linear region (with $V_{gs} > V_{th}$ and $V_{ds} < V_{dsat}$).

$$I_{lin} = \frac{\pi t_{si} \mu C_{ox} E_c}{(E_c L + V_{ds})} \left[(V_{gs} - V_{th})^2 V_{ds} - \frac{\theta_{short} V_{ds}^2}{2} \right] \quad (18)$$

Where $\alpha = 1.24$ for $L = 30 \text{ nm}$, E_c is the critical electric field

and θ_{short} can be formulated as [34]:

$$\theta_{short} = \frac{0.1}{\partial \Phi(0, Z_{min})} \text{ at } V_{gs} = V_{th} \quad (19)$$

$$V_{dsat} = V_{gs}(1 - \theta_{short}) \quad (20)$$

Saturation Current, I_{sat} (When, $V_{gs} > V_{th}$ and $V_{ds} \geq V_{gs} - V_{th}$) can be codified as velocity of electrons by substituting V_{ds} as V_{dsat}

($V_{dsat} = V_{gs} - V_{th}$) [46] to get:

$$I_{sat} = \frac{\pi t_{si} \mu C_{ox} E_c^2}{(1 + \frac{V_{dsat}}{E_c L})(L - L_{sat})} \left[\beta (V_{gs} - V_{th})^2 V_{ds} - \frac{\theta_{short} V_{ds}^2}{2} \right] \quad (21)$$

for $V_{sat} = 1.03 \times 10^7 \text{ cm/s}$.

The SS (Subthreshold- Slope) can be expressed as:-

$$SS = \frac{1}{\frac{d}{dV_{gs}} \log(I_{sub})} \quad (22)$$

IV. RESULTS AND DISCUSSION

A. DM-JL-BT-FET as a Bio-sensor

Fig. 2 shows the drift in hole concentration for Nanowire FET and Biotube FET for different biomolecule concentrations. As is evident from Fig. 2, the change in hole concentration for Biotube FET based Bio-sensor is higher than the change in Nanowire FET. This is because of the supplementary inner-gate in DM-JL-BT-FET Sensor. The presence of this inner cylindrical gate exerts an additional gate control over the channel, thereby enhancing the effective field and capacitance. The hole concentrations are inherently dependent on the field and capacitances which increase for DM-JL-BT-FET Sensor in comparison to Nanowire FET Sensor under the influence of different biomolecule concentrations.

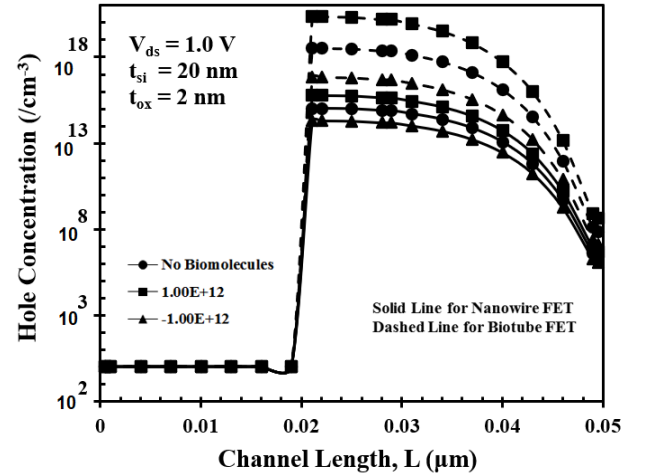


Fig. 2 Hole Concentrations (cm^{-3}) for Nanowire FET $\rightarrow -1e^{12}$, 0, $1e^{12}$ respectively and Biotube FET $\rightarrow -1e^{12}$, 0, $1e^{12}$ respectively.

Fig. 3 shows the drift in electric field for Nanowire FET and Biotube FET for different biomolecule concentrations. As is evident from Fig. 3, the change in electric field, E_z for Biotube FET based Bio-sensor is higher than the change in Nanowire FET. This is because of the supplementary inner-gate in DM-JL-BT-FET Sensor.

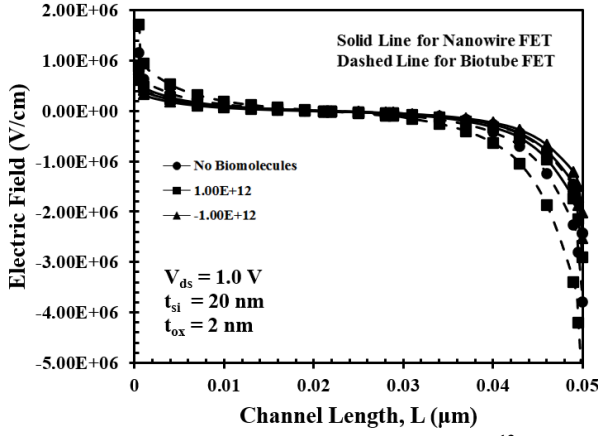
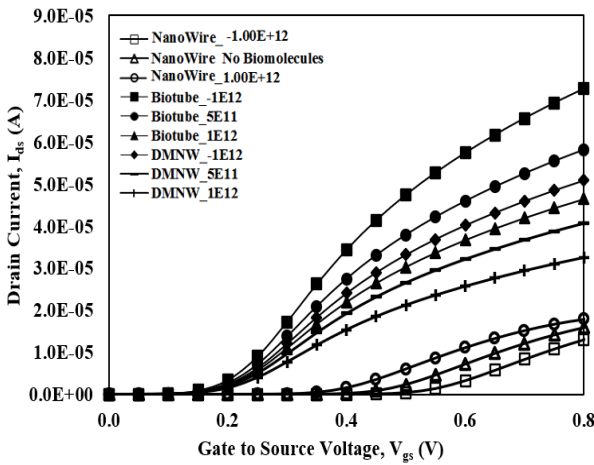
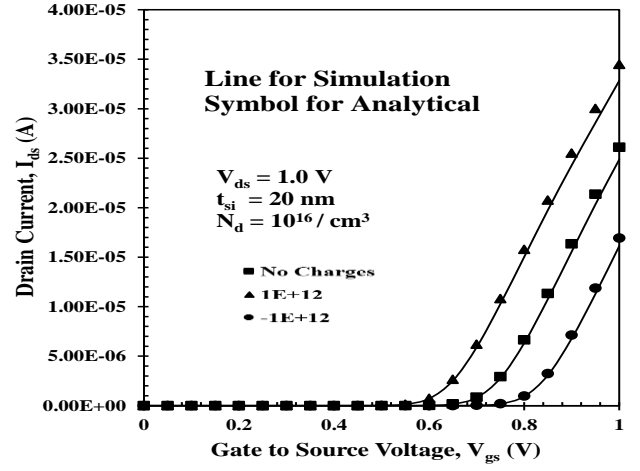


Fig. 3 Electric Field for Nanowire FET $\rightarrow -1e^{12}$, 0, $1e^{12}$ respectively and Biotube FET $\rightarrow -1e^{12}$, 0, $1e^{12}$ respectively.

Fig. 4(a) shows the drift in I_{ds} with V_{gs} for Nanowire FET, Dual Metal Nanowire (DMNW) FET and Biotube FET with different biomolecule concentrations. The biomolecules are detected in the gate stack device architecture, when the biomolecule species are used in the gate oxide. By tailoring the biomolecule concentrations in the oxide, the detection of DNA biomolecules becomes attainable. It implies from the Fig., that Biotube FET illustrates a higher shift in the drain current over Nanowire FET for detecting the various biomolecules. This higher change in the drain current owes to the Biotube FET structure which has two gates. The potential elevates the lateral electric field and also the gate transport efficiency, thereby enhancing the drain current and biomolecule detection. Fig. 4(b) shows the analytical I_{ds} v/s V_{gs} for different biomolecule concentrations for Biotube FET. It can be clearly derived from the figure that the analytical results are in close accordance with the simulated results. It can also be inferred that when positive biomolecules are inserted the drain current increases and vice-versa. This is because of the change in the potential (upward/downward) on insertion of biomolecule (positive/negative).



(a)



(b)

Fig. 4(a) I_{ds} - V_{gs} Characteristics for different Biomolecule Concentrations for different device structures_(b) Analytical I_{ds} v/s V_{gs} for different biomolecule concentrations for Biotube FET

Fig. 4(c) shows I_{ON}/I_{OFF} ratio for the four contemplated device designs under different biomolecule concentrations. Efficiently working as a switch becomes an essential requisite for a device to be used for digital applications, thereby making the switching speed a crucial benchmark for evaluating the performance of the device, which can be expressed as:

$$\frac{I_{ON}}{I_{OFF}} = \frac{I_{ds(ON)atV_{gs} = 1.0V}}{I_{ds(OFF)atV_{gs} = 0.0V}} \quad (1)$$

As seen from the Fig., the shift in I_{ON}/I_{OFF} Ratio is higher in Biotube FET as compared to Nanowire FET because of superior control along the channel and thus making Biotube highly sensitive. It can be clearly derived from the Fig. and Table 3, that when Biotube FET is used for sensing different biomolecules, a larger drift in the ratio is tapped because of an additional inner core gate which manifests greater control over the channel. For negative biomolecules a change of 10% is observed in Nanowire FET based biosensor and 43% change is observed in Biotube FET based biosensor.

Fig. 4 (d) shows the drift in Conduction Band Energy (CBE) for Nanowire FET and Biotube FET for different biomolecule concentrations. As is evident from Fig. 4, the change in hole concentration for Biotube FET based Bio-sensor is higher than the change in Nanowire FET. This is because of the supplementary inner-gate in DM-JL-BT-FET Sensor. The presence of this inner cylindrical gate exerts an additional gate control over the channel, thereby enhancing the effective field and capacitance.

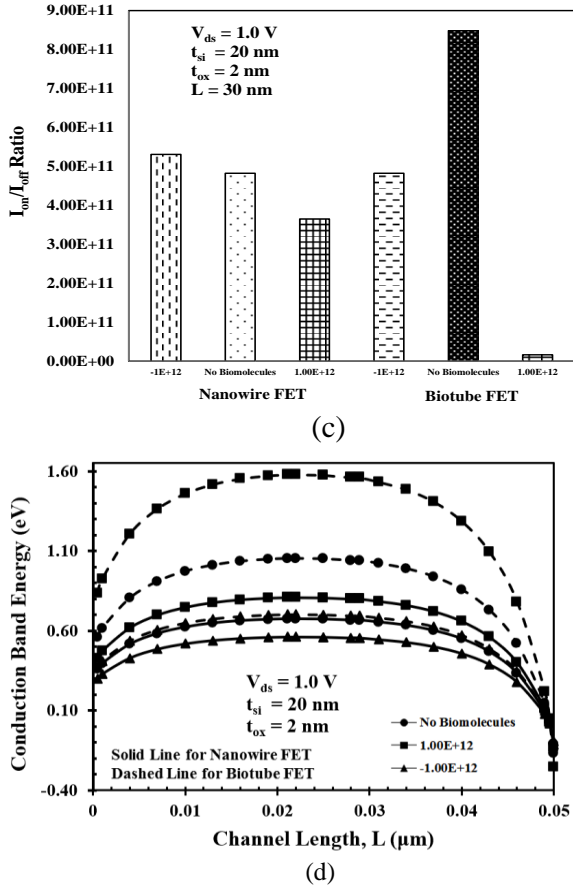


Fig. 4(c) I_{ON}/I_{OFF} Ratio for different Biomolecule Concentrations Fig. 4(d) Conduction Band Energy (CBE) contour plots for different biomolecule concentrations (i) Biotube FET (ii) Nanowire FET

Table 3. Sensitivity Parameters

Parameter	Nanowire FET			Biotube FET		
Biomolecule Concentrations \rightarrow	-1E+12	No Biomolecules	1E+12	-1E+12	No Biomolecules	1E+12
I_{ON}/I_{OFF}	5.31E+11	4.82E+11	3.65E+11	4.82E+11	8.49E+11	1.64E+10
V_{th} (V)	0.693	0.652	0.606	0.721	0.627	0.532
SS (V/Decade)	1.43E-01	1.55E-01	1.67E-01	6.50E-02	8.6E-02	2.63E-01

g_m is the first-order-derivative of I_{ds} w.r.t V_{gs} . Fig. 5(a) illustrates g_m (Transconductance) with change in V_{gs} for the contemplated designs. It is evident from the Fig., that Bio-tube FET shows a higher drift in g_m in contrast to Nanowire FET under different biomolecule concentrations. Owing to the benefits of the dual cylindrical gate architecture the charge carrier density and the capacitances increase by two fold times which further elevates the drain current. g_m being the derivative of the drain current also increases with an increase in I_{ds} under different biomolecule concentrations. Fig. 5(b) shows the analytical g_m v/s V_{gs} for different biomolecule concentrations for Biotube FET. It can be clearly noticed from the figure that the analytical results are in close accordance with the simulated

results. It can also be inferred that when positive biomolecules are inserted g_m increases and vice-versa. This is because of the change in the potential (upward/downward) on insertion of biomolecule (positive/negative).

V_{th} (Threshold voltage) of a device is expressed as the minimum V_{gs} at which the MOSFET turns ON and starts conducting. Fig. 5(c) shows threshold voltage for the contemplated device architectures under different biomolecule concentrations. As seen from the Fig. and Table 3, the shift in threshold voltage is higher in Biotube FET instead of Nanowire FET. For negative biomolecules a change of 6% is observed in Nanowire FET based biosensor and 15% change is observed in Biotube FET based biosensor and for positive biomolecules a change of 7% is observed in Nanowire FET based biosensor and 15.2% change is observed in Biotube FET based biosensor. This is because of superlative channel control, thus making Biotube highly sensitive.

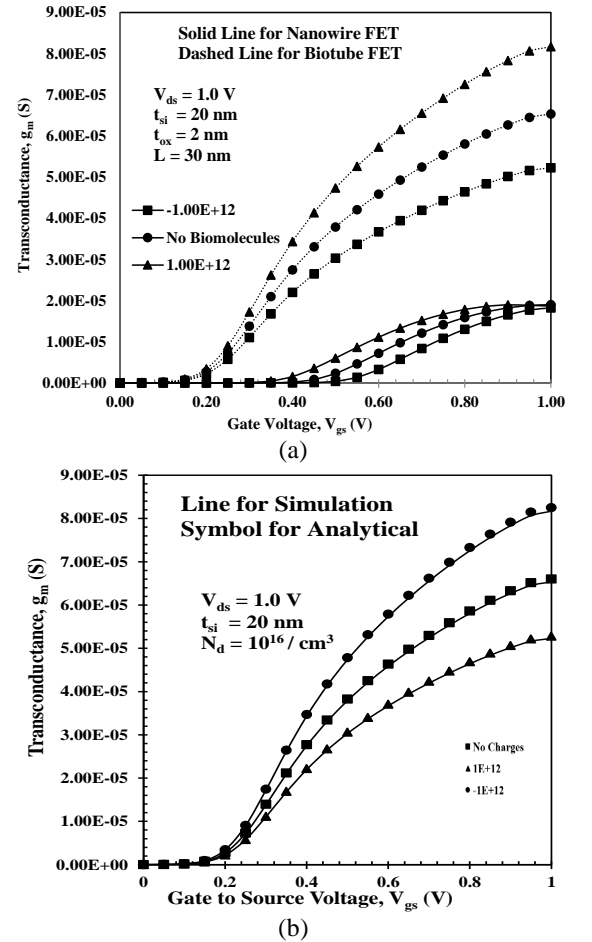


Fig. 5(a) g_m for different Biomolecule Concentrations (b) Analytical g_m v/s V_{gs} for different biomolecule concentrations for Biotube FET

Fig. 5 (c) shows the threshold voltage, V_{th} for Nanowire FET, Dual Metal Nanowire (DMNW) FET and Biotube FET for different biomolecule concentrations. As is evident from Fig. 5 (c), the change in hole concentration for Biotube FET based Bio-sensor is higher than the change in Nanowire FET and DMNW) FET. This is because of the supplementary inner-gate in DM-JL-BT-FET Sensor. The presence of this inner cylindrical gate exerts an additional gate control over the

channel, thereby enhancing the effective field and capacitance. Fig. 5 (d) shows the Sensitivity, S for Nanowire FET, Dual Metal Nanowire (DMNW) FET and Biotube FET for different biomolecule concentrations. Table 4 shows the Sensitivity Analysis Table for different devices.

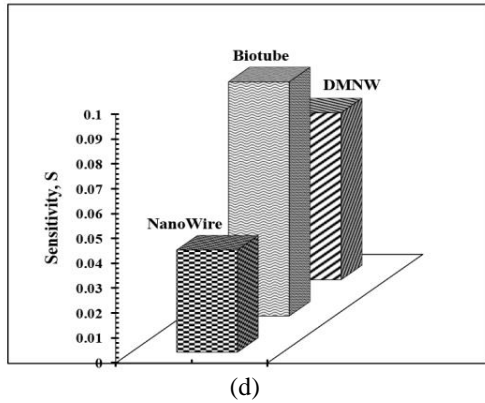
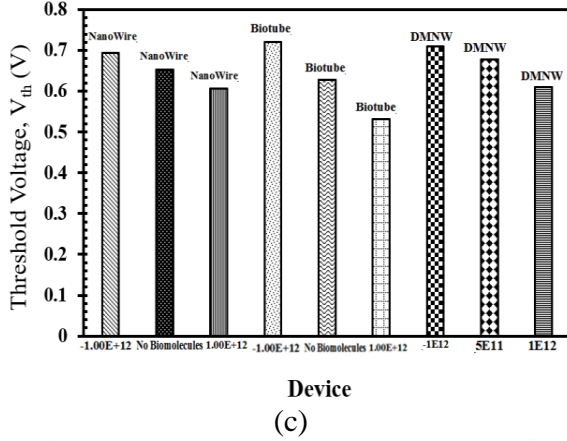


Fig. 5(c) V_{th} (d) Sensitivity of different Devices

Table 4: Sensitivity Analysis Table

Parameter	Nanowire FET			Biotube FET			DMNW FET		
Biomolecule Concentration \rightarrow	-1E+12	No Biomolecules	1E+12	-1E+12	No Biomolecules	1E+12	-1E+12	No Biomolecules	1E+12
V_{th} (V)	0.693	0.652	0.606	0.721	0.627	0.532	0.71	0.677	0.61
Sensitivity	0.041			0.094			0.067		

Fig. 6 (a) illustrates change in SS (Subthreshold Slope) for all the designs contrasted. It should ideally be approaching towards 60mV/decade as this also adds to the switching capacity of the device. Clearly evident from the Fig., the shift in subthreshold slope is higher in Biotube FET than in Nanowire FET. Fig. 6(b) shows the analytical SS v/s Channel Length for different biomolecule concentrations for Biotube FET. It can be clearly derived from the Fig. that the analytical results are in close accordance with the simulated results. It can also be inferred that when positive biomolecules are inserted the subthreshold slope increases and vice-versa. This is because of the change in the potential (upward/downward) on insertion of biomolecule (positive/negative) and also due to the nanotube structure which manifests a greater control over the

channel. It should also be noted that the sensitivity of NWFET/BIOTUBE FET sensors can be exponentially enhanced in the subthreshold regime where the gating effect of molecules bound on a surface is the most effective due to the reduced screening of carriers in NWs [48].

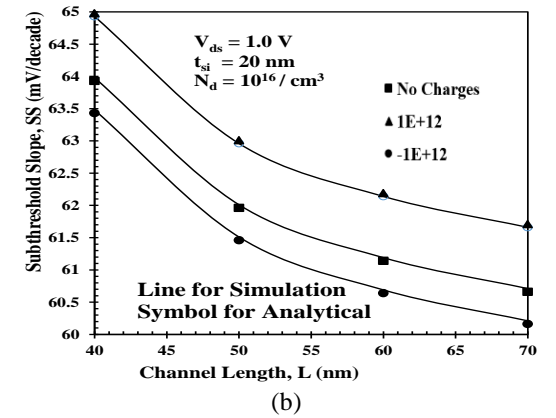
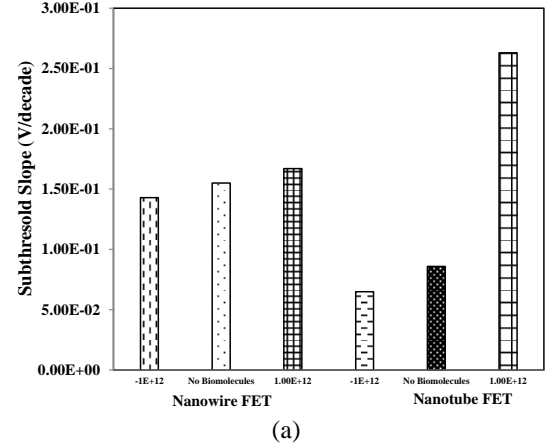
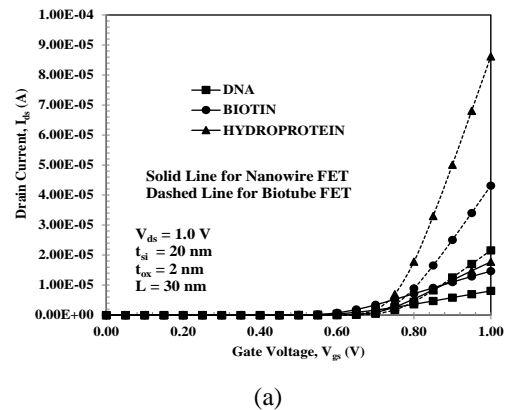


Fig. 6(a) Change in SS for different Biomolecule Concentrations (b)SS for different biomolecule concentrations for Biotube

B. Influence of Biomolecule Species

Fig. 7(a) illustrates the drift in I_{ds} with V_{gs} for Nanowire FET and Biotube FET with different biomolecule dielectric constants, viz. DNA, Biotin and Hydroprotein. The biomolecules are detected in the gate stack device architecture, when the biomolecule species are used in place of high K gate oxide.



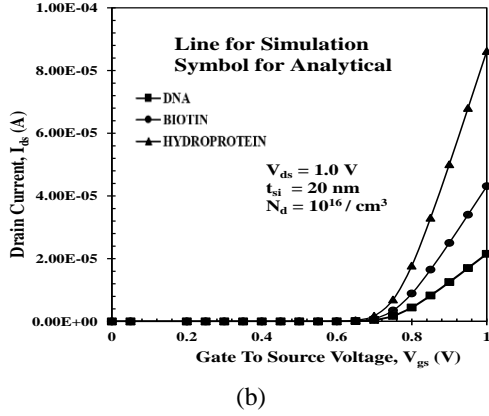


Fig. 7(a) Variation in I_{ds} (b) Analytical I_{ds} v/s V_{gs}

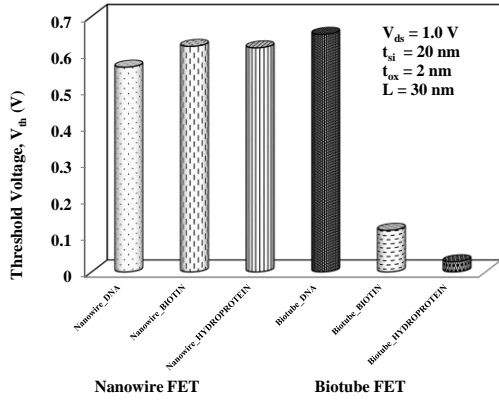


Fig. 7(c) Threshold Voltage, V_{th} for different Biomolecule Dielectric Constants

By tailoring the permittivity of the material that is inserted along with the dielectric-constant of neutral species like DNA = 1.0, Biotin = 2.1 and Hydroprotein = 5.0, the detection of DNA biomolecules becomes attainable. This becomes possible, as oxide is a gate stack, and when the permittivity changes, it changes the capacitance and hence the electric field also, changes as capacitance is given by the ratio of ϵ_{ox} and t_{ox} . It implies from the Fig. that Biotube FET illustrates a superior drain current over the Nanowire FET for detecting the various biomolecules viz. DNA, Biotin and Hydroprotein. This superiority in drain current owes to the fact that in the Biotube FET structure, two cylindrical gates exercises an enhanced gate control over the channel. Fig. 7(b) shows the analytical I_{ds} v/s V_{gs} for different biomolecules (DNA, Biotin and Hydroprotein) of Biotube FET. It can be clearly noticed from the figure that the analytical results are in close accordance with the simulated results. It can also be inferred that when the dielectric-constant of neutral species are inserted (DNA = 1.0, Biotin = 2.1 and Hydroprotein = 5.0) the drain current increases. This is because of a change in the potential on insertion of biomolecule (depending upon the dielectric permittivity of biomolecules) [42,49]. Fig. 7(c) shows threshold voltage for the contemplated device architectures under different biomolecules. As seen from the Fig., drift in V_{th} is higher in Biotube FET than that of Nanowire FET because of superior control along the channel and thus making Biotube highly sensitive.

Fig. 8(a) displays the Sensitivity of the device designs being contemplated for different biomolecule dielectric constants. It implies from the Fig. that, Biotube illustrates superior sensitivity as contrasted with other device designs for detection of the three biomolecules viz. DNA, Biotin and Hydroprotein.

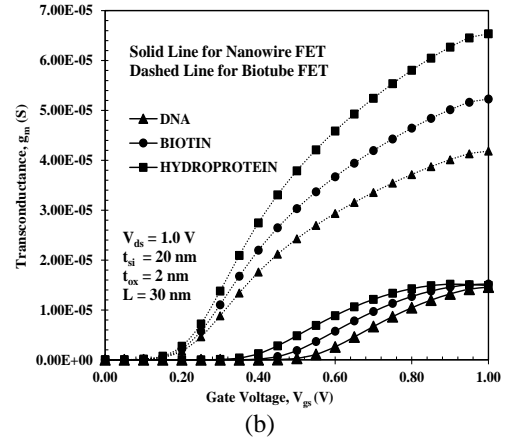
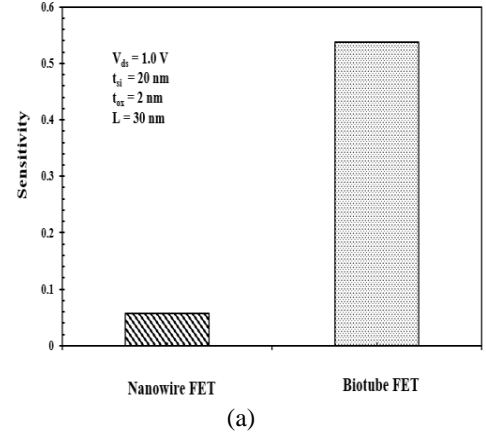


Fig. 8(a) Sensitivity for various Biomolecule Dielectric Constants (b) Transconductance, g_m for different Biomolecules

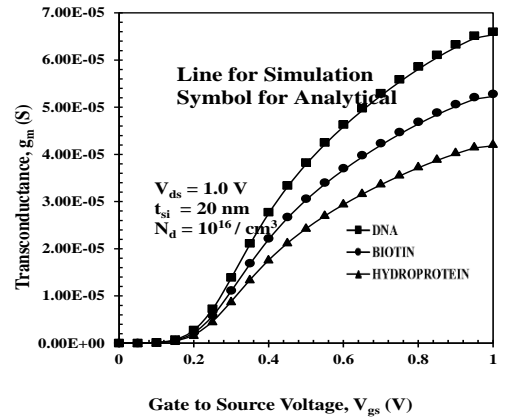


Fig. 8(c) Analytical g_m v/s V_{gs} for different biomolecules for Biotube FET

Fig. 8(b) illustrates the variation in g_m (Transconductance) with change in V_{gs} for the contemplated device architectures. It is clearly seen from the Fig. that Biotube FET shows a

higher drift in g_m in contrast to nanowire FET. The dual cylindrical gate architecture increases the charge carrier density and the capacitances by two fold times which further enhances the drain current. g_m being the derivative of the drain current also increases with an increase in I_{ds} for biomolecules like DNA, Biotin and Hydroprotein. Fig. 8 (c) shows the analytical g_m v/s V_{gs} for different biomolecules (DNA, Biotin and Hydroprotein) for Biotube FET. It can be clearly noticed from the figure that the analytical results are in close accordance with the simulated results. It can also be inferred that when the dielectric-constant of neutral species are inserted (DNA = 1.0, Biotin = 2.1 and Hydroprotein = 5.0) g_m increases. This is because of a change in the potential on insertion of biomolecule (depending upon the dielectric permittivity of biomolecules).

Fig. 9(a) shows I_{ON}/I_{OFF} ratio for the contemplated device designs under different biomolecule concentrations. As seen from the Fig., the shift in I_{ON}/I_{OFF} Ratio is higher in Biotube FET than Nanowire FET owing to superior control along the channel thus making Biotube highly sensitive for different biomolecules. Fig. 9(b) illustrates SS (Subthreshold Slope) for all the device designs considered and shows that the shift in subthreshold slope is higher in Biotube FET than in Nanowire FET and this is because of elevated control along the channel and thus making Biotube highly sensitive. Fig. 9 (c) shows the analytical SS v/s L for different biomolecules (DNA, Biotin and Hydroprotein) for Biotube FET. It can be clearly noticed from the Fig. that the analytical results are in close accordance with the simulated results. It can also be inferred that when the dielectric-constant of neutral species are inserted (DNA = 1.0, Biotin = 2.1 and Hydroprotein = 5.0) SS increases. This is because of a change in the potential on insertion of biomolecule.

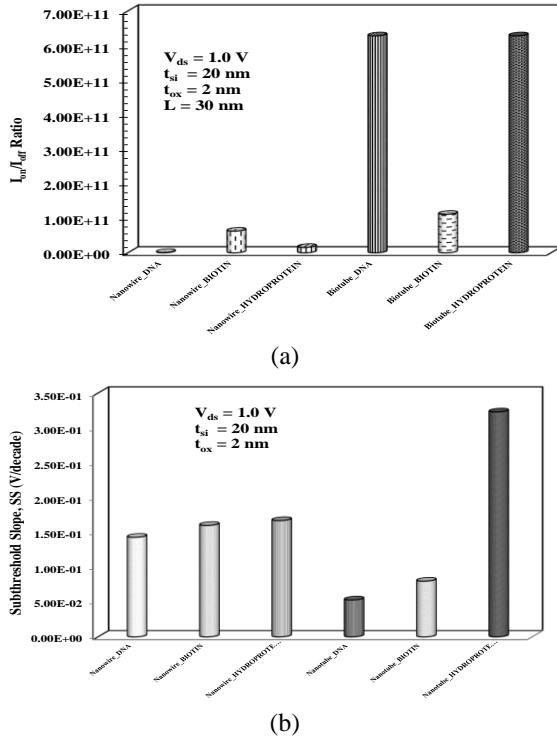


Fig. 9(a) I_{ON}/I_{OFF} Ratio (b) SS for different Biomolecule Dielectric Constants

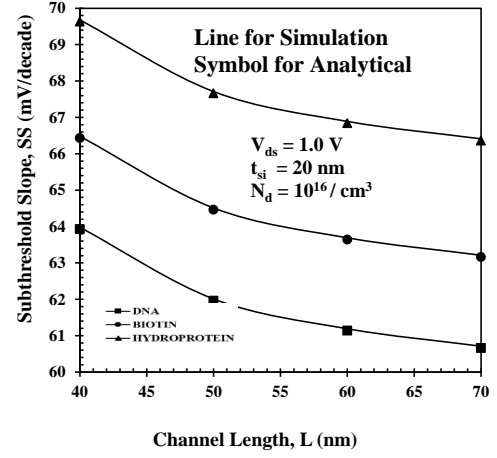


Fig. 9(c) Analytical SS v/s L for different biomolecules

Fig. 10(a) shows change in surface potential for different Biomolecule concentrations, which spawns over three categories i.e., No charged Biomolecules (neutral), positively charged biomolecule and negatively charged biomolecule. The change in the minima of surface-potential will modulate the V_{th} . When (-)vely charged biomolecules are infused inside the gate oxide layer, minima of the surface potential gets lowered, whereas for positively charged biomolecules, this minima of the surface potential increases. This owes to the fact that these biomolecules are electrically linked with the rooted silicon and attains a particular energy level, i.e. interface state energy level, E_{IT} which places itself according to the fermi level, E_F , as per their charges ($E_{IT} > E_F \rightarrow$ for (+)vely charged biomolecule insertions, $E_{IT} < E_F \rightarrow$ for (-)vely charged biomolecule insertion) [34]-[41]. The flat band voltage increases due to infusion of (-) vely charged biomolecules and reduces for the (+) vely charged biomolecule insertion by qN_f/C_G amount. Here, C_G can be expressed as capacitance per unit area of the gate-dielectric and N_f can be expressed as interface fixed charge density for charged biomolecules [47]. For larger channel radius, C_G becomes smaller, so for significant change in flat band potential, band bending occurs. Thus, effective V_{gs} varies and the minima of the surface-potential shifts. Conclusively, V_{th} varies distinctively for charged biomolecules as contrasted with neutral ones. It can be interpreted that for a smaller amount of charged biomolecules, change in V_{th} would be very small. For the higher amount of charged biomolecules, V_{th} shifts significantly. Fig. 10(b) shows variation of surface potential for different Biomolecules, specifically Hydro-protein, Biotin and DNA respectively. As is evident from the Fig., DNA shows the maximum surface potential as compared to the other two biomolecules viz., Hydro-protein and Biotin. DNA has been reportedly the greatest (-) vely charged molecule [31]. The operation of the FET can be modulated by this charge induced in the dielectric layer despite the proven dielectric constant effect [28].

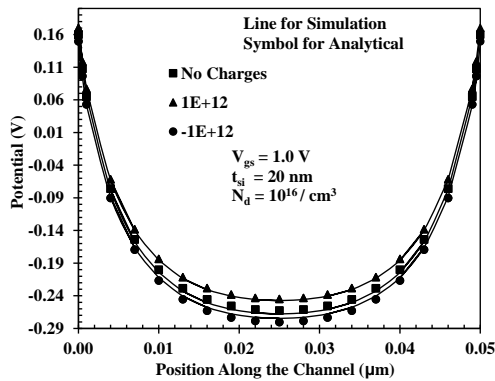


Fig. 10(a) Surface-Potential for Different Biomolecules

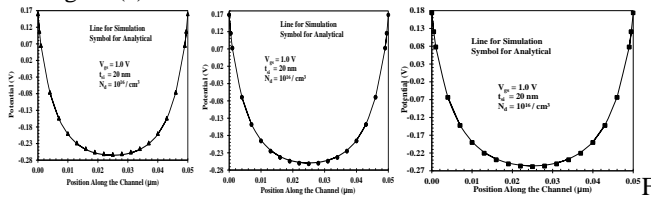


Fig. 10(b) Surface Potential for Different Biomolecules → Hydro-Protein, Biotin and DNA respectively.

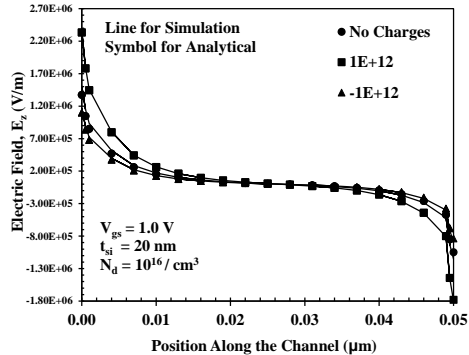
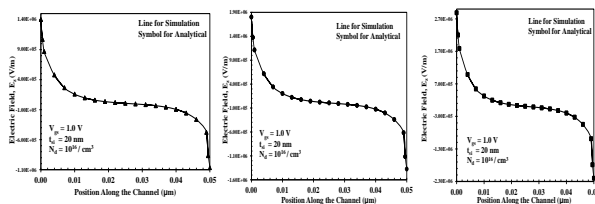
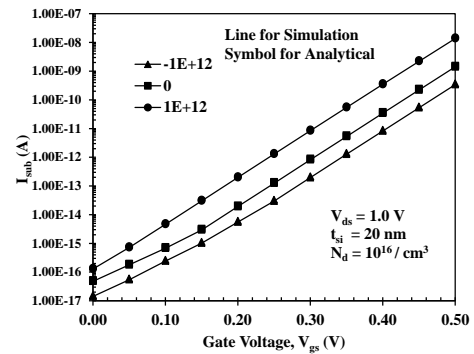
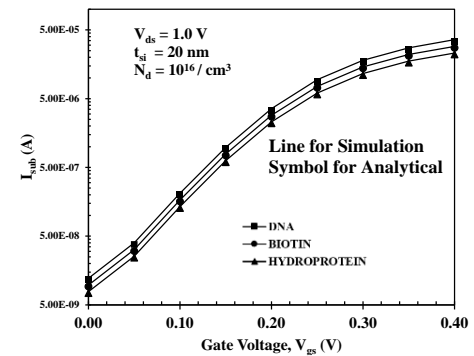
Fig. 11(a) E_z for Different Biomolecule Concentrations

Fig. 11(b) Electric Field for Different Biomolecules → Hydro-Protein, Biotin and DNA respectively

Fig. 11(a) shows variation in electric field for different biomolecule concentrations primarily over three categories i.e., NO charged Biomolecules (neutral), positively charged biomolecule and negatively charged biomolecule. As can be interpreted from the Fig., (+)vely charged biomolecules exhibits the maximum maxima for electric field as contrasted with the neutral and (-)vely charged biomolecules. Fig. 11(b) shows variation of electric field for different Biomolecules, specifically Hydro-protein, Biotin and DNA respectively. As is clearly evident from the Fig., the bio-sensor exhibits the maximum electric field when detecting DNA biomolecule.



(a)



(b)

Fig. 12 (a) Subthreshold Current for various Biomolecule Concentrations (b) Subthreshold Current for Different Biomolecules

Fig. 12 pictures variation of subthreshold current for (a) Different Biomolecule Concentrations and (b) Different Biomolecules. The different biomolecule concentrations includes the (+) vely charged biomolecules, neutrally charged biomolecules and (-) vely charged biomolecules. As interpreted from the Fig., (+) vely charged biomolecules and DNA exhibits the maximum variation in the maxima of the subthreshold current, thus advocating our device to be spell-fall for detection of biomolecules, both by inducing in the dielectric layer as well as by varying the dielectric constants.

V. CONCLUSION

In this manuscript, an analytical model has been demonstrated for Dielectric Modulated Junctionless Biotube FET (DM-JL-BT-FET) as a sensor. The Junctionless Biotube FET based sensor has been compared and contrasted with Nanowire FET under similar bio-molecule conditions. It has been well established by the comparisons, that Dielectric Modulated Junctionless Biotube FET shows much higher efficiency in Bio-sensing and poses superior device performance characteristic in terms of higher sensitivity, higher drift in drain current, transconductance, I_{on}/I_{off} ratio, Subthreshold Slope and Threshold voltage. The hole concentrations have also been investigated under different biomolecule conditions. Two different biomolecule conditions have been deeply considered in our analysis viz., firstly, varying the biomolecule concentrations and secondly, inserting different biomolecules namely DNA, Biotin and Hydroprotein. Improved bio sensing is observed in Junctionless Biotube FET

because of superlative gate control over the channel, owing to architecture of Biotube FET. The analytical results have also been modelled for DM-JL-BT-FET by finding a solution to the 2-D Poisson equation in accordance with the boundary conditions. The analytical results are much in coherence with the results obtained from the simulator.

APPENDIX

$$M_n = \frac{L_{2n} - L_{1n} \exp(-\frac{L}{\eta_n})}{2 \sinh(\frac{L}{\eta_n})} \quad (A1), \quad N_n = \frac{L_{1n} \exp(\frac{L}{\eta_n}) - L_{2n}}{2 \sinh(\frac{L}{\eta_n})} \quad (A2)$$

$$L_n = \frac{2}{i^2 J_1^2(\frac{t_{eff}}{\eta_n})} \left[\rho t_{eff} \eta_n^2 \left\{ t_{eff} J_2(\frac{t_{eff}}{\eta_n}) - (t_{eff} - \frac{t_{eff}}{2}) \left(\frac{J_1(\frac{t_{eff}}{\eta_n})}{\eta_n} - J_0(\frac{t_{eff}}{\eta_n}) \right) \right\} \right] \quad (A3)$$

$$L_{2n} = \frac{2}{i^2 J_1^2(\frac{t_{eff}}{\eta_n})} \left[\rho t_{eff} \eta_n^2 \left\{ t_{eff} J_2(\frac{t_{eff}}{\eta_n}) - (t_{eff} - \frac{t_{eff}}{2}) \left(\frac{J_1(\frac{t_{eff}}{\eta_n})}{\eta_n} - J_0(\frac{t_{eff}}{\eta_n}) \right) \right\} \right] \quad (A4)$$

REFERENCES

- [1] S. Vaddiraju, I. Tomazos, D.J. Burgess, F.C. Jain, F. Papadimitrakopoulos, "Emerging synergy between nanotechnology and implantable bio-sensors: a review", *Biosens. Bioelectron.*, Vol. 25, pp.1553–1565, 2010.
- [2] Y. Cui, Q. Wei, H. Park, and C. M. Lieber, "Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species," *Science*, vol. 293, no. 5533, pp. 1289–1292, Aug. 2001.
- [3] A. Goel, S. Rewari, S. Verma, R.S. Gupta, "Temperature-dependent gate-induced drain leakages assessment of dual-metal nanowire field-effect transistor—analytical model," *IEEE Transactions on Electron Devices*, Vol. 366, No. 5, pp. 2437–45, April 2019.
- [4] P. G. Collins, K. Bradley, M. Ishigami, and A. Zettl, "Extreme oxygen sensitivity of electronic properties of carbon Biotubes," *Science*, vol. 287, no. 5459, pp. 1801–1804, Mar. 2000.
- [5] S. R. Manalis, E. B. Cooper, P. F. Indermuhle, P. Kernen, P. Wagner, D. G. Hafeman, S. C. Minne, and C. F. Quate, "Microvolume field-effect pH sensor for the scanning probe microscope," *Appl. Phys. Lett.*, vol. 76, no. 8, pp. 1072–1074, Feb. 2000.
- [6] S. Purushothaman, C. Toumazou, and C. Ou, "Protons and single nucleotide polymorphism detection: A simple use for the ion sensitive field effect transistor," *Sens. Actuators B, Chem.*, vol. 114, no. 2, pp. 964–968, Apr. 2006.
- [7] G. Shekhawat, S. H. Tark, and V. P. Dravid, "MOSFET-embedded Microcantilevers for measuring deflection in biomolecular sensors," *Science*, vol. 311, no. 5767, pp. 1592–1595, Mar. 2006.
- [8] E. Stern, J. F. Klemic, D. A. Routenberg, P. N. Wyrembak, D. B. Turner-Evans, A. D. Hamilton, D. A. LaVan, T. M. Fahmy, and M. A. Reed, "Label-free immunodetection with CMOS-compatible semiconducting nanowires," *Nature*, vol. 445, no. 7127, pp. 519–522, Feb. 2007.
- [9] S. Rewari, V. Nath, S. Haldar, S.S. Deswal, R.S. Gupta, "Hafnium oxide based cylindrical junctionless double surrounding gate (CJLD SG) MOSFET for high speed, high frequency digital and analog applications," *Microsystem Technologies*, vol. 25(5), pp. 1527–36, May 2019.
- [10] S. Rewari, V. Nath, S. Haldar, S.S. Deswal, R.S. Gupta, "Novel design to improve band to band tunneling and gate induced drain leakages (GIDL) in cylindrical gate all around (GAA) MOSFET," *Microsystem Technologies*, vol. 25(5), pp. 1537–46, May 2019.
- [11] G. Zheng, F. Patolsky, Y. Cui, W. U. Wang, and C. M. Lieber, "Multiplexed electrical detection of cancer markers with nanowire sensor arrays," *Nat. Biotechnol.*, vol. 23, no. 10, pp. 1294–1301, Oct. 2005.
- [12] Z. Li, Y. Chen, T. I. Kamins, K. Nauka, and R. S. Williams, "Sequencespecific label-free DNA sensors based on silicon nanowires," *Nano Lett.*, vol. 4, no. 2, pp. 245–247, Jan. 2004.
- [13] F. Patolsky, G. Zheng, O. Hayden, M. Lakadamyali, X. Zhuang, and C. M. Lieber, "Electrical detection of single viruses," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 39, pp. 14 017–14 022, Sep. 2004.
- [14] N. Elfström, R. Juhasz, I. Sychugov, T. Engfeldt, A. E. Karlström, and J. Linnros, "Surface charge sensitivity of silicon nanowires: Size dependence," *Nano Lett.*, vol. 7, no. 9, pp. 2608–2612, Sep. 2007.
- [15] J. P. Colinge, C.W. Lee, A. Afzal, N. D. Akhavan, R. Yan, I. Ferain, P. Razavi, B. O'Neill, A. Blake, M. White, A. Kelleher, B. McCarthy, and R. Murphy, "Nanowire Transistors Without Junctions", *Nature Nanotechnology*, vol. 5, no. 3, pp. 225–229, 2010.
- [16] A. Kranti, R. Yan, C.-W. Lee, I. Ferain, R. Yu, N. D. Akhavan, P. Razavi, and J. P. Colinge, "Junctionless nanowire transistor (JNT): Properties and Design Guidelines", In *Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, pp. 357–360, 2010.
- [17] D. Ghosh, M. S. Parihar, G. A. Armstrong, and A. Kranti, "High Performance Junctionless MOSFETs for Ultralow-Power Analog/RF Applications", *IEEE Electron Device Letters*, vol. 33, no. 10, pp. 1477–1479, Oct. 2012.
- [18] T. Wang, L. Lou, and C. Lee, "A Junctionless Gate-All-Around Silicon Nanowire FET of High Linearity and Its Potential Applications", *IEEE Transactions on Electron Devices*, vol. 34, no.4, pp.478–484, May 2013.
- [19] Nitin Trivedi, Manoj Kumar, Subhasis Haldar, S.S. Deswal, Mridula Gupta and R. S. Gupta, "Analytical Modelling of Junctionless Accumulation Mode MSOFET (JAM-CSG)", *International Journal Of Numerical Modelling: Electronic Networks, Devices And Fields*, Feb. 2016.
- [20] V. Nathan and N. C. Das, "Gate-induced drain leakage current in MOS devices," *IEEE Trans. Electron Devices*, vol. 40, no. 10, pp. 1888–1890, Oct. 1993.
- [21] Hoffmann, T., Doornbos, G., Ferain, I., Collaert, N., Zimmerman, P., Goodwin, M., Rooyackers, R., Kottantharayil, A., Yim, Y., Dixit, A. and De Meyer, K., "GIDL (gate-induced drain leakage) and parasitic Schottky barrier leakage elimination in aggressively scaled HfO₂/TiN FinFET devices," in *IEDM Tech. Dig.*, pp. 725–729, 2005.
- [22] A. Sharma, A. Jain, Y. Pratap, and R. S. Gupta, "Effect of high-K and vacuum dielectrics as gate stack on a junctionless

- cylindrical surrounding gate (JL-CSG) MOSFET," *Solid-State Electron.*, vol. 123, pp. 26–32, Sep. 2016.
- [23] J. P. Duarte, S.-J. Choi, D.-I. Moon, and Y.-K. Choi, "Simple analytical bulk current model for long-channel double-gate Junctionless transistors," *IEEE Trans. Electron Devices*, vol. 32, no. 6, pp. 704–706, Jun. 2011.
- [24] S. Rewari, V. Nath, S. Haldar, SS Deswal, R.S. Gupta, "Improved analog and AC performance with increased noise immunity using Biotube junctionless field effect transistor (NJLFET)," *Applied Physics A.*, vol. 122(12), pp. 1049, Dec 2016.
- [25] S. Rewari, S. Haldar, V. Nath, SS Deswal, RS Gupta, "Numerical modeling of Subthreshold region of junctionless double surrounding gate MOSFET (JLDSG)," *Superlattices and Microstructures*, vol. 90, pp. 8-19, Feb 2016.
- [26] P. Marconcini, G. Fiori, M. Macucci, G. Iannaccone, "Hierarchical simulation of transport in silicon nanowire transistors," *Journal of Computational Electronics*, vol. 7, pp. 415, 2008.
- [27] M. Luisier, G. Klimeck, "Atomistic full-band simulations of silicon nanowire transistors: Effects of electron-phonon scattering," *Phys. Rev. B*, vol. 80, pp. 155430, 2009.
- [28] A. Bafekry, S. Farjami Shayesteha, M. Ghergherehchi, F. M. Peetersb" Adsorption of molecules on C3N nanosheet: A first-principle calculations" *Chemical Physics*, Volume 526, pp. 110442, Oct. 2019.
- [29] A. Bafekry, Asadollah, Saber Farjami Shayesteh, Mitra Ghergherehchi, and Francois M. Peeters. "Tuning the bandgap and introducing magnetism into monolayer BC3 by strain/defect engineering and adatom/molecule adsorption." *Journal of Applied Physics* 126, no. 14, pp. 144304, 2019.
- [30] A. Bafekry, Asadollah, Catherine Stampfl, Mitra Ghergherehchi, and Saber Farjami Shayesteh. "A first-principles study of the effects of atom impurities, defects, strain, electric field and layer thickness on the electronic and magnetic properties of the C2N nanosheet." *Carbon* 157, pp. 371-384, 2020.
- [31] A. Bafekry, Asadollah, Mohammed Obeid, Chuong Nguyen, Meysam Bagheri Tagani, and Mitra Ghergherehchi. "Graphene hetero-multilayer on layered platinum mineral Jacutingaite (Pt2HgSe3): Van der Waals heterostructures with novel optoelectronic and thermoelectric performances." *Journal of Materials Chemistry A*, 2020.
- [32] A. Bafekry, A., M. Yagmurcukardes, M. Shahrokhi, and M. Ghergherehchi. "Electro-optical properties of monolayer and bilayer boron-doped C3N: Tunable electronic structure via strain engineering and electric field." *Carbon* 168, pp. 220-229, 2020.
- [33] A. Bafekry, A., M. Yagmurcukardes, B. Akgenc, M. Ghergherehchi, and C. V. Nguyen. "Van der Waals heterostructures of layered Janus transition-metal dichalcogenides (MoS2 and Janus MoSSe) on graphitic boron-carbon-nitride (BC3, C3N, C3N4 and C4N3) nanosheets: A First-Principles study." *J. Phys. D: Appl. Phys* 53, pp. 355106, 2020.
- [34] A. Bafekry "Graphene-like BC6N single-layer: Tunable electronic and magnetic properties via thickness, gating, topological defects, and adatom/molecule" *Physica E: Low-dimensional Systems and Nanostructures*, Volume 118, pp. 113850, April 2020.
- [35] A Bafekry, B Akgenc, M Ghergherehchi and F M Peeters "Strain and electric field tuning of semi-metallic character WCrCO2 MXenes with dual narrow band gap" *Journal of Physics: Condensed Matter*, Volume 32, Number 35, pp. 355504, 2020.
- [36] P. Bergveld, "Thirty years of ISFETOLOGY: What happened in the past 30 years and what may happen in the next 30 years," *Sens. Actuators B, Chem.*, vol. 88, no. 1, pp. 1–20, Jan. 2003.
- [37] L. Bousse, N. F. De Rooij, and P. Bergveld, "Operation of chemically sensitive field-effect sensors as a function of the insulator-electrolyte interface," *IEEE Trans. Electron Devices*, vol. 30, no. 10, pp. 1263–1270, Oct. 1983.
- [38] H. Im, X. J. Huang, B. Gu, and Y. K. Choi, "A dielectric-modulated field effect transistor for biosensing," *Nat. Nanotechnol.*, vol. 2, no. 7, pp. 430–434, Jul. 2007.
- [39] C.-H. Kim, C. Jung, H. G. Park, and Y.-K. Choi, "Novel dielectric modulated field-effect transistor for label-free DNA detection," *Biochip J.*, vol. 2, no. 2, pp. 127–134, Jun. 2008.
- [40] B. Gu, T. J. Park, J.-H. Ahn, X.-J. Huang, S. Y. Lee, and Y.-K. Choi, "Nanogap field-effect transistor bio-sensors for electrical detection of avian influenza," *Small*, vol. 5, no. 21, pp. 2407–2412, Aug. 2009.
- [41] J. M. Kinsella and A. Ivanisevic, "Biosensing: Taking charge of biomolecules," *Nature Nanotechnol.*, vol. 2, pp. 596–597, Oct. 2007.
- [42] S. Kanungo, S. Chattopadhyay, K. Sinha, P.S. Gupta, H. Rahaman, "A Device Simulation-Based Investigation on Dielectrically Modulated Fringing Field-Effect Transistor for Biosensing Applications," *IEEE Sensors Journal*, Vol. :17(5), pp. :1399-406, Dec. 2016.
- [43] N. Shafi, J.S. Parmar, A. Porwal, A. M. Bhat, C. Sahu, and C. Periasamy, "Gate All Around Junctionless Dielectric Modulated BioFET Based Hybrid Biosensor," *Silicon*, pp.: 1-12, July 2020.
- [44] S.J. Choi, D.I. Moon, S. Kim, J.P. Duarte, Y.K. Choi, Sensitivity of threshold voltage to nanowire width variation in Junctionless transistors. *IEEE Electron Device Lett.*, vol. 32, pp. 125–132, 2011.
- [45] ATLAS User's Manual: 3-D Device Simulator, SILVACO International, Version 5.14.0.R, 2018.
- [46] A. Goel, S. Rewari, S. Verma, R.S. Gupta, "Modeling of shallow extension engineered dual metal surrounding gate (SEE-DM-SG) MOSFET gate-induced drain leakage (GIDL)," *Indian Journal of Physics (Springer Publication)*, pp. 1-10, March 2020.
- [47] R. Gautam, M. Saxena, R. S. Gupta, and M. Gupta. "Numerical model of gate-all-around MOSFET with vacuum gate dielectric for biomolecule detection." *IEEE electron device letters* 33, Vol. no. 12, pp.: 1756-1758, 2012.
- [48] X. P. Gao, G. Zheng, C. M. Lieber, "Subthreshold Regime Has the Optimal Sensitivity for Nanowire FET Biosensors," *Nano Lett*, Vol.: 10, pp.: 547-552, 2010.
- [49] C. H. Kim, J. H. Ahn, K. B. Lee, C. Jung, H. G. Park and Y. K. Choi, "A new sensing metric to reduce data fluctuations in a nanogap-embedded field-effect transistor biosensor," *IEEE transactions on electron devices*, Vol.: 59(10), pp.: 2825-2831, Aug. 2012.

PAPER

Effect of fly ash and graphite addition on the tribological behavior of aluminium composites

To cite this article: Vipin Kumar Sharma *et al* 2021 *Surf. Topogr.: Metrol. Prop.* **9** 025027

View the [article online](#) for updates and enhancements.



IOP | ebooks™

Bringing together innovative digital publishing with leading authors from the global scientific community.

Start exploring the collection—download the first chapter of every title for free.

Surface Topography: Metrology and Properties



PAPER

Effect of fly ash and graphite addition on the tribological behavior of aluminium composites

Vipin Kumar Sharma¹ , Ramesh Chandra Singh² and Rajiv Chaudhary²

¹ Deptt. of M.A.E., Maharaja Agrasen Institute of Technology, Rohini Sector-22, Delhi-110086, India

² Deptt. of M.E., Delhi Technological University, Delhi-110042, India

E-mail: vipin.dtu@gmail.com

Keywords: wear, flyash, composite, friction, lubrication

RECEIVED
2 December 2020

REVISED
17 April 2021

ACCEPTED FOR PUBLICATION
6 May 2021

PUBLISHED
13 May 2021

Abstract

The present work discusses the wear and friction studies of aluminium graphite fly ash composite. Three aluminium composites, Al with 10 wt. % fly ash, Al with 10 wt. % graphite, and Al with 5 wt. % fly ash & 5 wt. % graphite were fabricated using the stir casting technique. A linear reciprocating tribometer was used to evaluate the wear and frictional behavior at two distinctive temperatures of 36 °C and 100 °C in dry and lubricated conditions. The aluminium with 10 wt. % fly ash resulted in the least amount of mass loss and coefficient of friction in dry as well as a lubricated condition at 36 °C, however, at a higher temperature of 100 °C the aluminium with 10% graphite exhibits the lowest mass loss and coefficient of friction value. It is concluded that for dry sliding conditions, aluminium fly ash composite could be used as a potential material for applications that are to be operated at room temperature and for high-temperature applications aluminium graphite composite is more suitable.

1. Introduction

Aluminium alloys are used in many industrial applications. It has an excellent weight to strength ratio, thermal conductivity, and corrosion resistance which makes it one of the most used material in the automobile sector. In aluminium alloys, AA 6061 exhibits very good mechanical properties which support its use in aircraft, electronics, and food packaging industries as well [1, 2]. Yield, ultimate, shear, and fatigue strength are the main factors that greatly influence the mechanical properties of a material. AA 6061 has a good yield and ultimate strength which makes its use in applications where static loading is required and the high shear strength of AA 6061 enables its use in different applications where the torsional load is required. The shear strength value helps in the application where there is repetitive loading like in axels and pistons. The other special property that AA 6061 possesses is corrosion resistance. In the presence of water or atmospheric air, it makes a layer of oxide which makes the material non-reactive to corrosive elements. Despite good mechanical and corrosion properties, the wear and friction properties of these alloys are not good [3, 4]. It is difficult to make components where sliding motion is

required. So, to enhance the wear and friction properties, aluminium is reinforced with different elements to fabricate aluminium metal composite (AMC).

In the recent past, silicon carbide, boron carbide, tungsten carbide, and aluminium oxide particle reinforcements have been successfully used to produce the AMCs. These hard ceramics improved the hardness and wear resistance of the different matrix materials [5–7]. However, with the introduction of these hard particles, higher coefficient of friction (COF) had been reported above specific load values. For reducing this effect soft phase particles like graphite, molybdenum has been added to the Al matrix. The layer of atoms on these elements is very weakly bonded and gets dislodged easily with low shearing forces which resulted in low friction. With these advantages of hybrid composites, a numerous amount of research is going on in its fabrication and processing. Sharma *et al* (2018) performed wear and friction tests on the Al-Gr composites and reported that the presence of graphite in the Al matrix greatly enhanced the tribological properties of the Al matrix [8]. Rajesh *et al* (2019) developed the Al-fly ash composites with varying graphite contents. With the increasing graphite contents in the Al-fly ash composite, the mechanical and tribological properties get increased [9]. Similarly, Dirisenapul *et al* (2020)

prepared a hybrid aluminium metal composite by reinforcing boron carbide and nitride in the aluminium 7010 alloys. No intermetallic compounds were detected by the authors and these reinforcements used up to 2% w/w improved the tensile properties of the aluminium alloy.

Palanikumar *et al* (2019) fabricated a hybrid composite of B₄C and mica with aluminium 6061 and compared the wear behavior of the hybrid composite with a metal matrix composite of aluminium 6061-B₄C and aluminium 6061 alloys. The authors reported that the hybrid composite exhibited better wear properties as well as surface finish [10]. Pitchayappillai *et al* (2016) developed a hybrid composite using alumina (Al₂O₃) and molybdenum disulfide powder as the reinforcement in Al 6061 alloy. The introduction of these reinforcements improved the wear resistance of the aluminium alloy [11]. Gopinath *et al* (2020) used the graphite, Al₂O₃, and boron nitride reinforcements in Al 6061 alloy to improve its wear, mechanical, and corrosion behaviors [12].

It is reviewed that fly ash particles improve the wettability and mechanical properties of aluminium alloys. The fly ash particles also improved the wear and friction properties as reported by Sharma *et al*. It contains oxides of the mineral contents from the coals used. The presence of oxides of aluminium, silicon, and carbide makes the nature of the fly ash ceramic.

Fly ash is the one industrial waste that has many important elements like Fe, Zn, Si, Ni. These elements might prove to be useful if mixed with aluminum alloys. The introduction of fly ash particles to the metal matrix was first proposed by Pont (1982) [134] and Rohtagi *et al* (1995) [13]. Numerous amounts of studies are available in which fly ash has been used as a possible reinforcement to aluminum alloys [14–16].

With these benefits of fly ash and to reuse the waste by-product of coal burning, fly ash was selected as one of the main reinforcement of the hybrid aluminium composite, and graphite was selected as the other reinforcement owing to its excellent friction reduction properties. It is also revealed from the literature review that individual use of fly ash and graphite particles in the aluminium matrix are proven to improve the tribological properties of the composite. However, most of the works related to composite materials lacked to show the combined effect of both the reinforcements. So, in the present paper, three composites were prepared using the stir casting with 10 wt. % fly ash, 10 wt. % graphite and 5 wt. % graphite-5 wt. % fly ash reinforcement. Wear and Friction tests were performed to evaluate the wear behaviour at two temperature settings in dry and lubricated conditions. The selection of temperatures for the tribological tests was done based on the literature survey and tribological testing machine specifications. For providing the lubrication SAE 15W40 multigrade lubricant was used as the mono-grade oils are not capable of working at higher temperatures [17]. SAE 15W40 lubricant

Table 1. Elemental composition of Al 6061 alloy.

Si	Fe	Mg	Cu	Cr	Zn	Mn	Al
0.61	0.68	0.81	0.21	0.05	0.27	0.17	97.2

performs well at high-temperature applications. The recorded density and viscosity of the used oil were 0.83 g cm⁻³ and 106.39 mm² s⁻¹ (at 100 °C).

2. Materials and methods

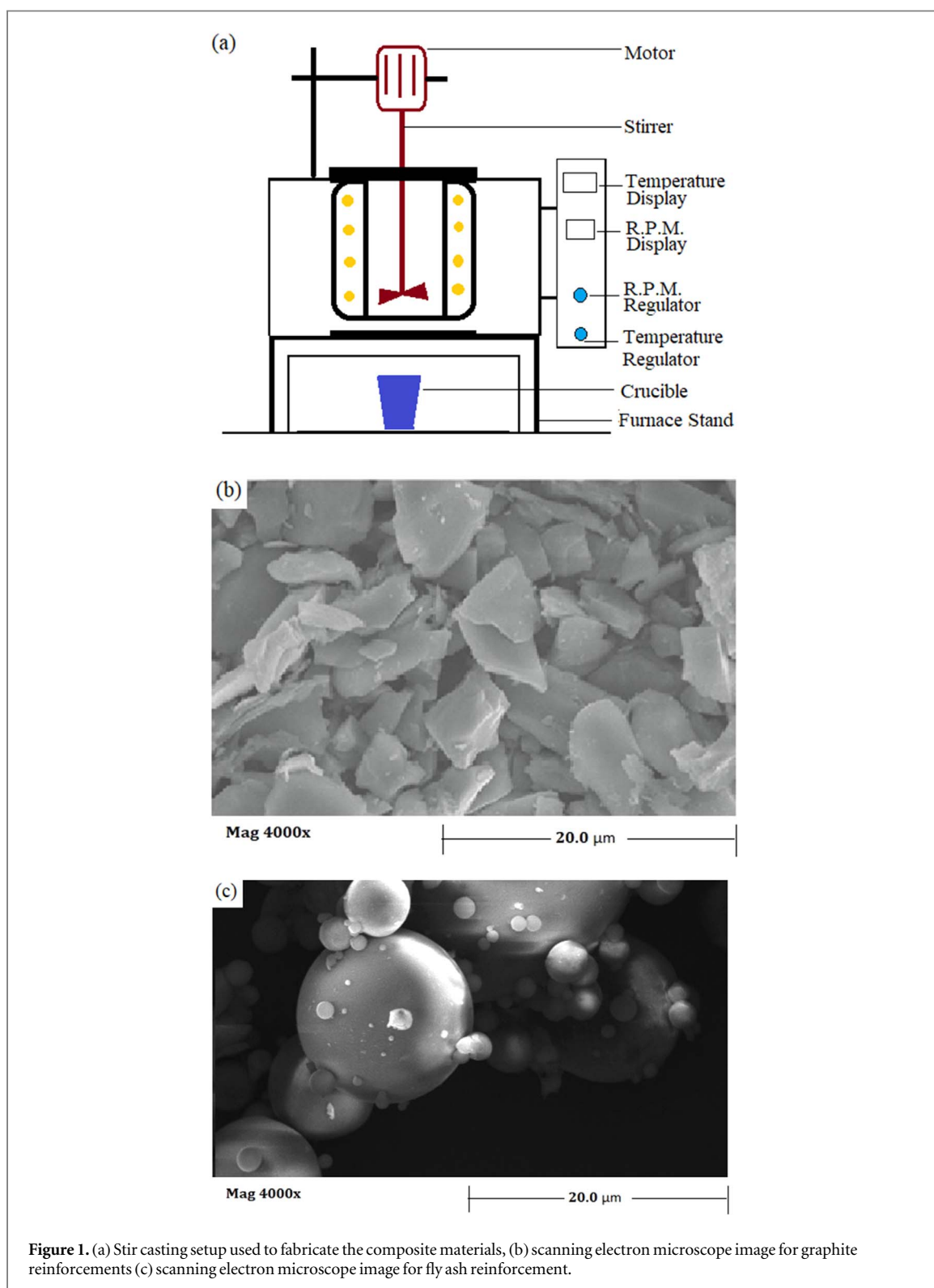
2.1. Materials

Al 6061 was used as the matrix material. The elemental composition of Al 6061 is given in table 1. Si, Mg, and Zn are the prime elements in it. These elements provide high strength and hardness.

2.2. Stir casting

There are several different methods to fabricate the composite materials [18–21]. The most common one is stir casting [22, 23]. The stir casting machine is an advanced version of the conventional electrical furnace with a digital interface for controlling input parameters (such as temperature, stirrer speed, as well as crucible setup) and a mechanical stirrer connected to the motor. The speed of the stirrer is controlled using a regulator provided at the input panel of the machine. An open space is provided at the top of the furnace for easy installation and movement of the stirrer to achieve uniform distribution of reinforcements into the parent material while performing the mixing operation. Figure 1 presents the basic parts of the stir casting setup.

The casting process begins with the heating of Aluminium (6061) at 800 °C and pre-heating of the reinforcement for 30 min to catalyze uniform distribution and strong bond formation between the Al 6061 and reinforcement. Al 6061 were filled in the crucible and the crucible was carefully placed on the crucible platform of the machine and the platform was directed to the furnace using the feed button. The temperature was set to 800 °C using a temperature-regulating switch. The reinforcement was preheated for 30 min in another conventional electric furnace. The Aluminium (6061) melts after some time as the furnace temperature reaches 800 °C, the stirrer was dipped into the melted aluminium and preheated reinforcements were added into the molten aluminium in form of very small balls of aluminium foil with reinforcement wrapped inside it. The fly ash contents were limited to 10% as it was noticed during the literature survey that, till 10% of fly ash particles the improvement in hardness value was very much appreciable, however with further increment in percentage, the improvement was not that high. Also, the higher percentage of reinforcements in the matrix results in the lower yield strength of the composite. For these



reasons, the amount was limited to 10% only. The stirrer was switched on and the rotation speed of the stirrer was set at 400 rpm. After the addition of the whole reinforcement, the mixing was done for the next 10 min and the stirrer was switched off. The stirrer was carefully taken out of the furnace and the molten mixture was heated at the set temperature (800 °C) for the

next 10 min. The furnace was switched off and the crucible platform of the machine was lower down through the regulating switch. The crucible was held carefully using tongs and the molten mixture was carefully poured into a preformed cavity of cylindrical shape with 100 mm diameter and 120 mm length. The casting was allowed to cool down and it was cautiously

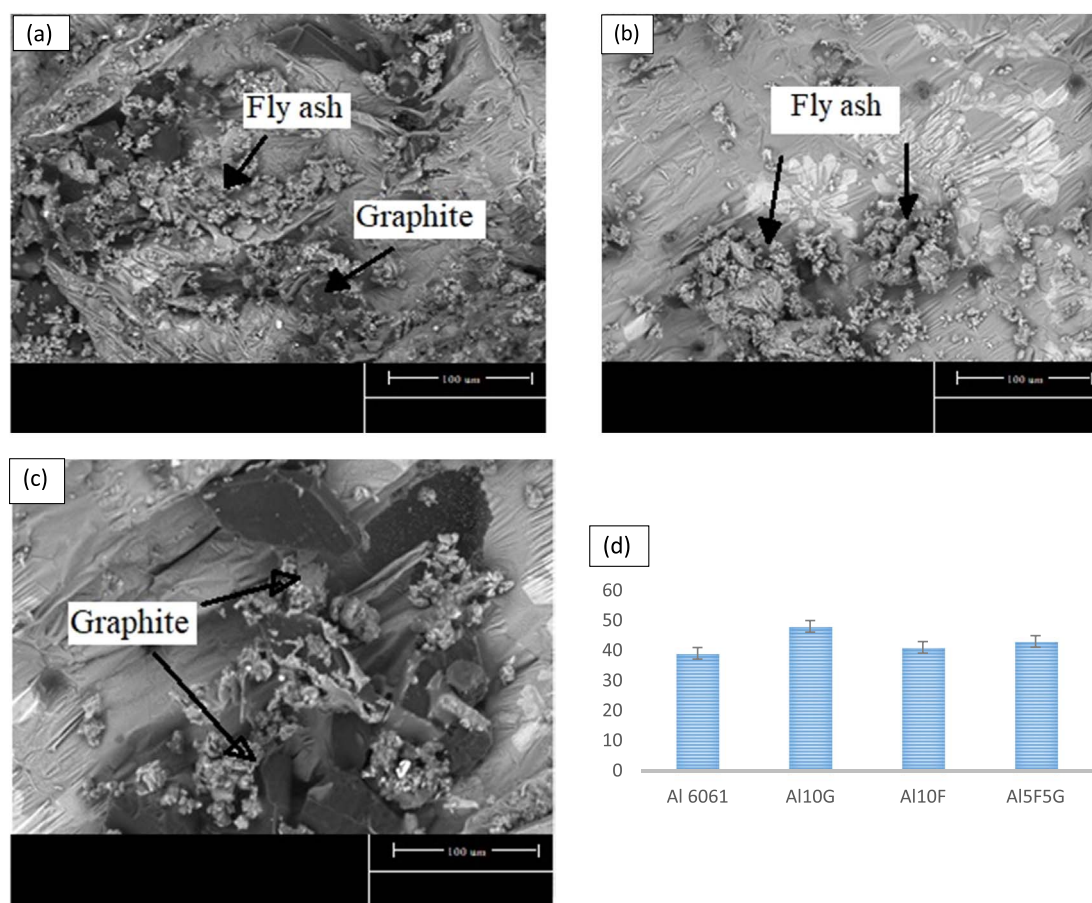


Figure 2. Optical micrograph of (a) Al5F5G (b) Al10F (c) Al10G (d) variation of hardness for the prepared composites.

extracted out from the metal mold. Table 2 provides the stir casting parameters used to fabricate the composites.

2.3. Measurements

The Vicker hardness of all the prepared materials was evaluated using a Fischer-made micro-hardness tester as per ASTM E-384. The measurements were performed at 3000 mN load for 15 s. Five observations were made and an average of these was used to evaluate the results.

The wear test was performed on the polished samples of aluminium metal matrix composites (surface roughness value of $1.1 \mu\text{m}$ (R_a value)) using the reciprocating tribometer. The set-up consists of the reciprocating head at which load was applied and a small platform of dimension $40 \text{ mm} \times 40 \text{ mm} \times 5 \text{ mm}$ with a thermocouple attached to it, for the installation of sample workpieces for wear testing as well as to read the variation in temperature of the workpiece while testing. The wear tests on the samples were performed in two conditions, dry condition and lubricated condition on the test samples of dimensions $20 \text{ mm} \times 20 \text{ mm}$. For the wear test in lubricated condition multi-grade, SAE 15W40 engine oil was used as the lubricant, for a reciprocating body, hardened steel ball having surface roughness value of $0.7 \mu\text{m}$ (R_a value) was used at a

Table 2. Parameters for stir casting.

Weight of aluminium (Kg)	1.5
Stirrer speed (rpm)	400
Temperature ($^{\circ}\text{C}$)	800
Reinforcement preheating (min)	30
Casting process duration (min)	90

Table 3. Parameters for wear testing.

Parameter/Lubrication	Dry	Lubricated
Temperature ($^{\circ}\text{C}$)	36, 100	36, 100
Duration (min)	5	60
Stroke length (mm)	6	6
Load (N)	10	10
Frequency (Hz)	2	2

load of 10 N. As per the design of the reciprocating tribo-meter, during the lubricated tribo-testing, the lubricant was applied before stating the experiment and there was no continuously supply of it. The frequency of the reciprocating stroke was set at 2 Hz. The tests in unlubricated conditions were performed for 6 min and for lubricated the duration of the test was 60 min to obtain the appreciable weight loss from the prepared samples. For each set of experiments, three observations

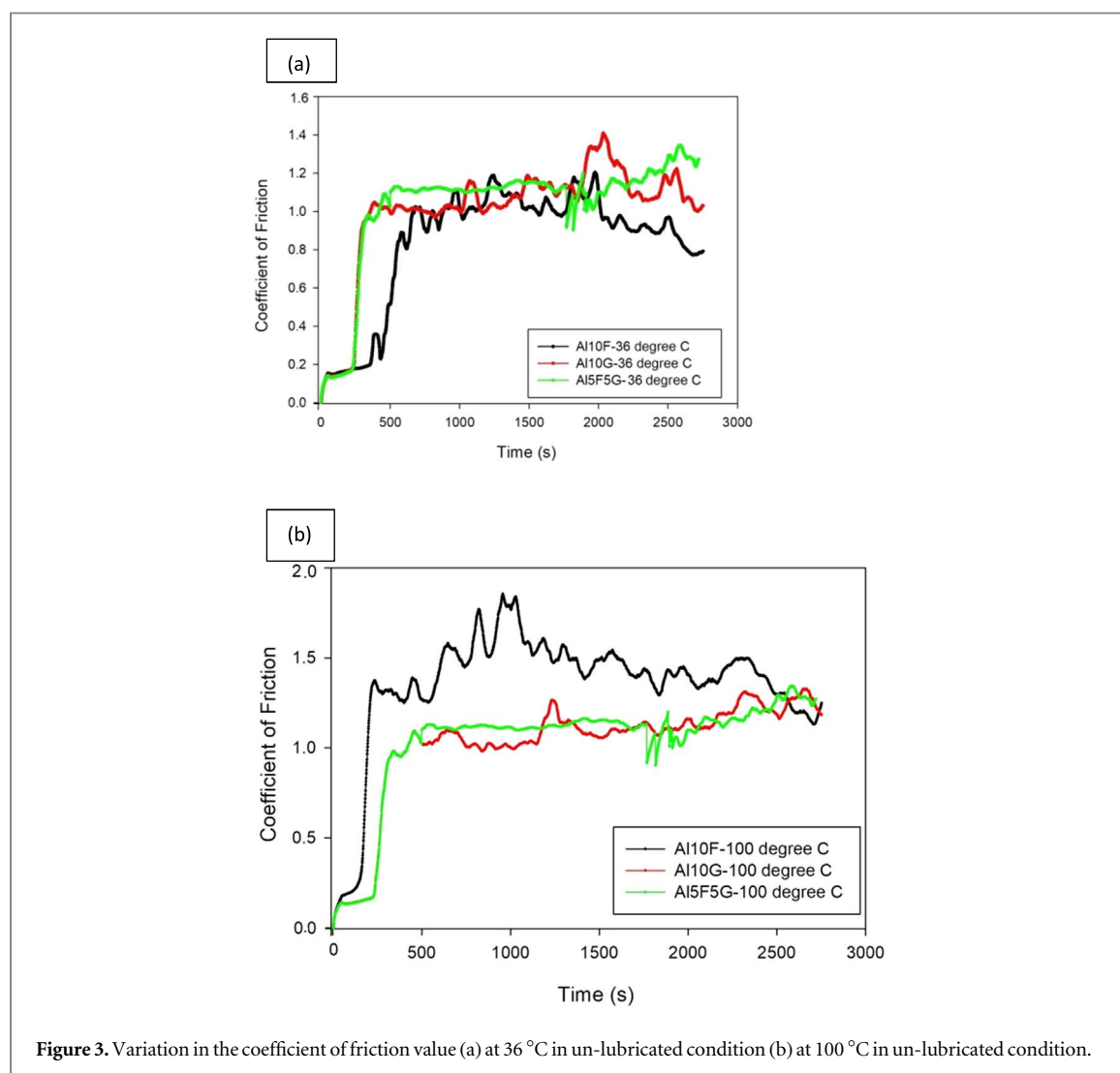


Figure 3. Variation in the coefficient of friction value (a) at 36 °C in un-lubricated condition (b) at 100 °C in un-lubricated condition.

were taken and an average of these was used to plot the results. The details about the working parameters are given in table 3. The variation in workpiece temperature, the frictional force was recorded for the whole duration of the wear test. The recorded data were analyzed to understand the worn nature of the fabricated composites. The sample were weighed before and after the test to evaluate the amount of mass loss in gram using a weighing balance having 0.0001 g accuracy.

3. Results and discussions

3.1. Microstructure and hardness

Figures 2(a)–(c) shows the scanning electron micrographs of fabricated composite materials obtained at 400 \times with accelerating voltage of 20 kV, an emission current of 47800 nm, and at a working distance of 6900 μ m with a ZEISS EVO Series Scanning Electron Microscope. The polished samples were etched in the solution prepared with the mixing of nitric acid

(5ml), hydrochloric acid (3 ml), hydrofluoric acid (2 ml), and distilled water (190 ml). The etchant helped in clearly identifying the grain boundaries. In the microstructure, it is observed that graphite and fly ash reinforcements were uniformly distributed throughout the aluminium matrix and some intermetallic compounds were also seen in the microstructure. The presence of the intermetallic compound was more prominent in the Al with 10 wt. % fly ash (Al10F) composite as compared to the other samples.

Figure 2(d) presents the variation of the surface hardness of the prepared composites. It is found that Al with 10 wt. % graphite (Al10G) possessed maximum hardness in comparison to other samples. The presence of graphite in the form of fine particles might have increased the surface hardness of the composite. Similar trends of hardness variation was observed by Shankar *et al* (2016) [24]. ElGhazaly *et al* (2017) [25] also reported that in the aluminium composites, the formation of aluminium carbides takes place which

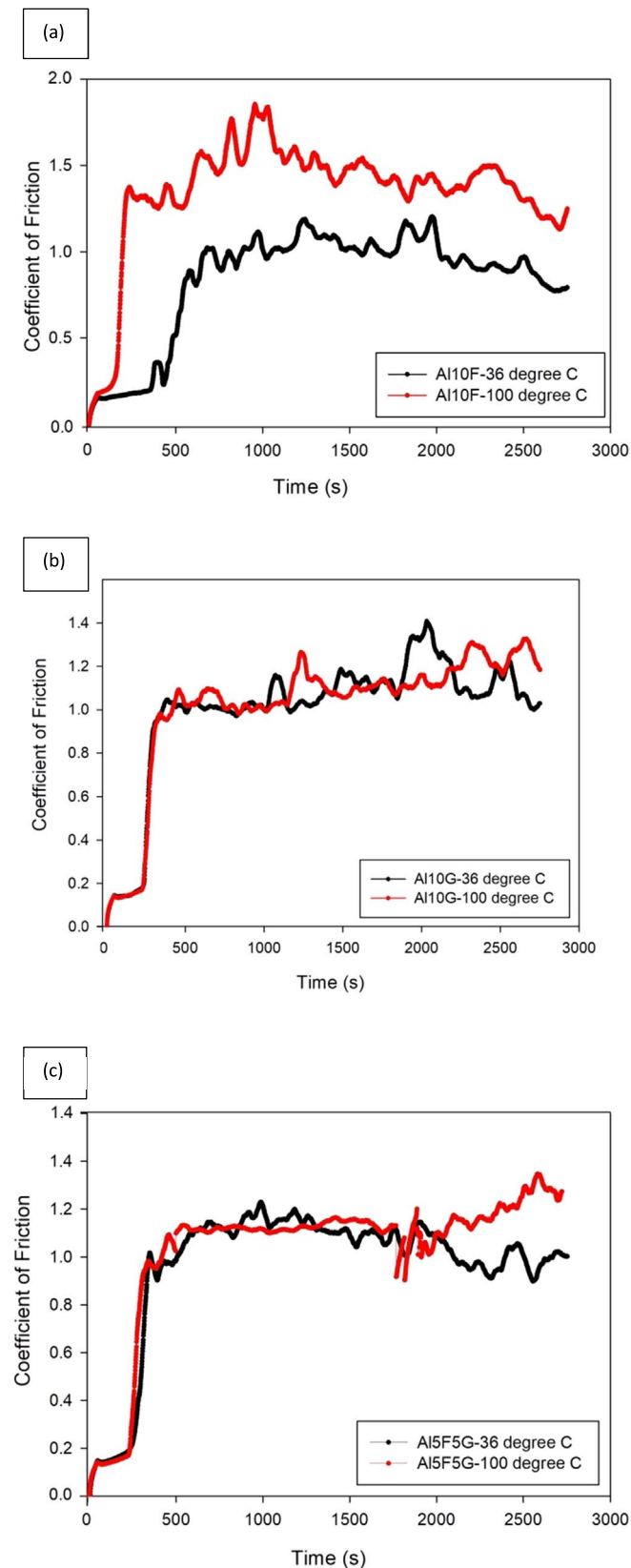


Figure 4. Comparison of the coefficient of friction value at 36 °C and 100 °C in un-lubricated condition for (a) Al10F composite (b) Al10G composite (c) Al5F5G composite.

reduces the ductility of the composite. With this, the hardness gets improved. It is also observed that the increase in hardness is not appreciable as compared with the Al with 10% fly ash contents.

3.2. Friction behaviour

3.2.1. Coefficient of friction in dry condition

The wear and friction experiments were performed in dry as well as in lubricated conditions at 36 °C and

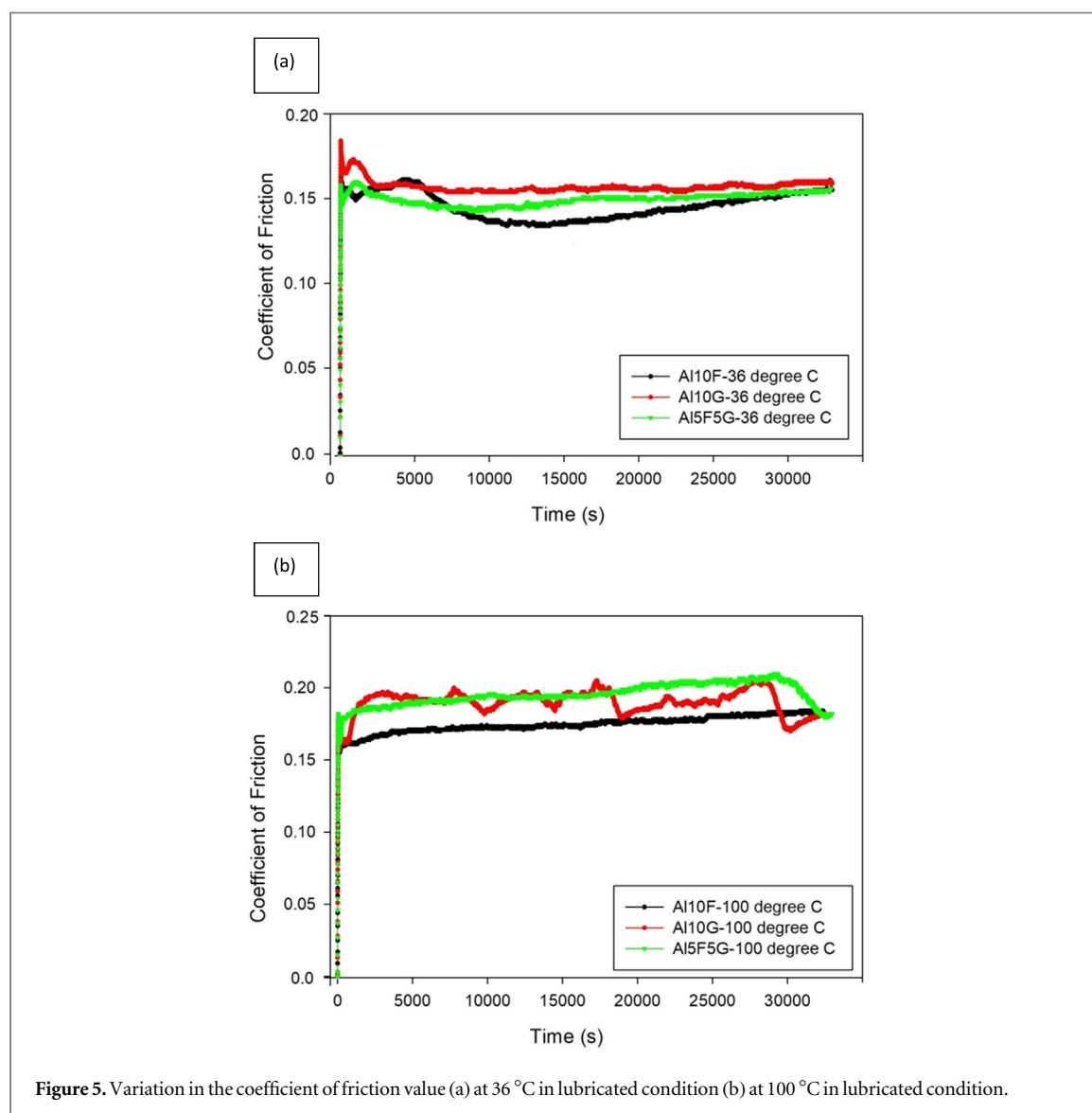


Figure 5. Variation in the coefficient of friction value (a) at 36 °C in lubricated condition (b) at 100 °C in lubricated condition.

100 °C temperature. Figure 3 presents the variation of coefficient of friction for the prepared hybrid composite as a function of sliding time at 36 °C working temperature. The friction force between the tribo-pairs was recorded with the help of a friction force sensor, and later the friction force was converted to the coefficient of friction as per the classical Coulomb's law [26, 27]. It was reported that at 36 °C, Al10F material produced the lowest COF value, followed by Al5G5F and Al10G. This low value of the COF presents a good bond between the aluminium and fly ash particles. The spherical shape of the fly ash particles helped in reducing the friction values. The dry sliding behaviour of the graphite particles in the aluminium matrix did not reduce the COF values for Al5F5G and Al10G as compared to the Al10F. The solid lubricating characteristics of the graphite particle depend on the environmental contaminations between the tribo-pairs [28]. At lower temperature values (36 °C) these contaminations affect the lubricating nature of graphite particles and did not allow to lower the COF value for graphite-reinforced Al5F5G and Al10G

Table 4. Average of COF values at 36 °C and 100 °C temperature in dry condition.

Composite	COF at 36 °C	COF at 100 °C	% increase
Al5F5G	0.9639	1.0306	6.9198
Al10F	0.8386	1.3503	61.0183
Al10G	1.0003	1.0173	1.69949

Table 5. Average of COF values at 36 °C and 100 °C temperature in lubricated condition.

Composite	COF at 36 °C	COF at 100 °C	% Increase
Al5F5G	0.1497	0.1954	30.5277
Al10F	0.1455	0.1746	20.00
Al10G	0.1569	0.1899	21.0325

composite. Hence the COF for Al10F is lower than Al4F4G and Al10G. However, at a higher temperature value (100 °C), the surface contaminations get released from the surface and graphite provides the required solid lubrication.

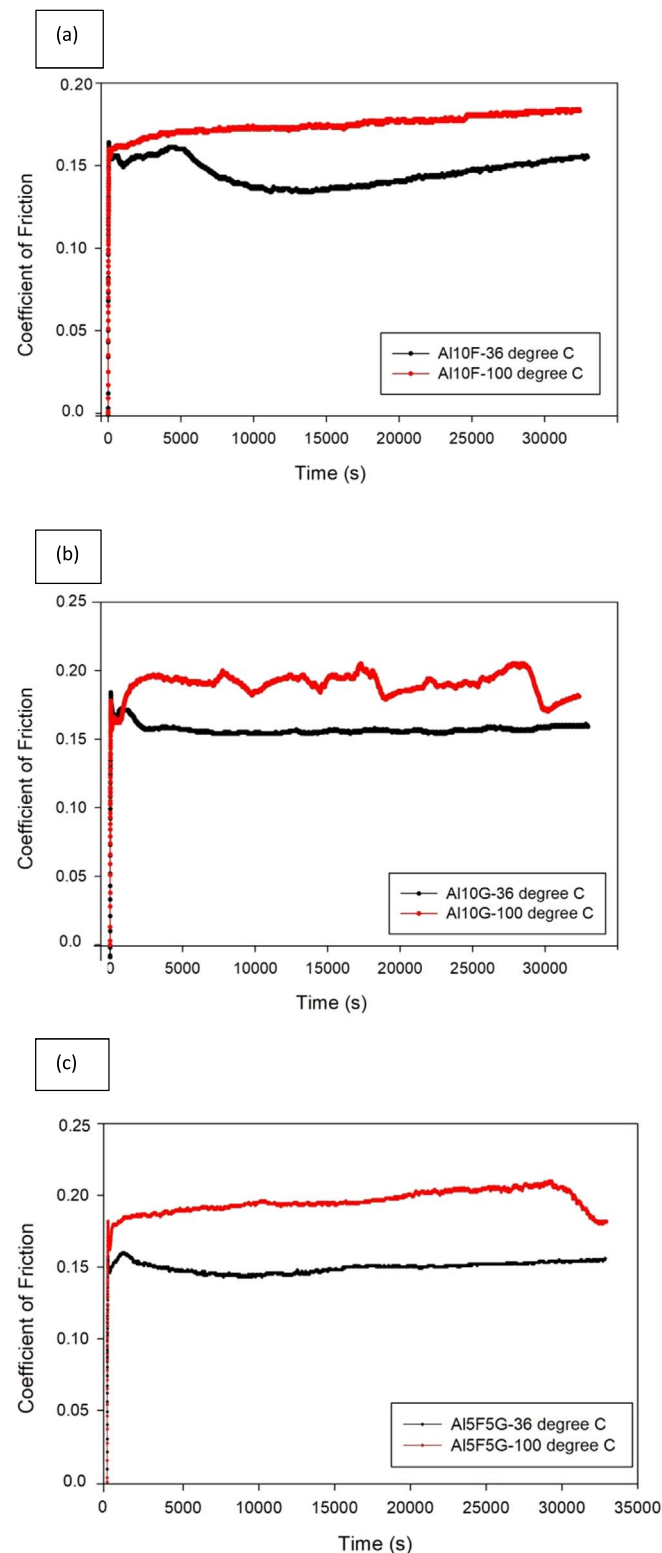
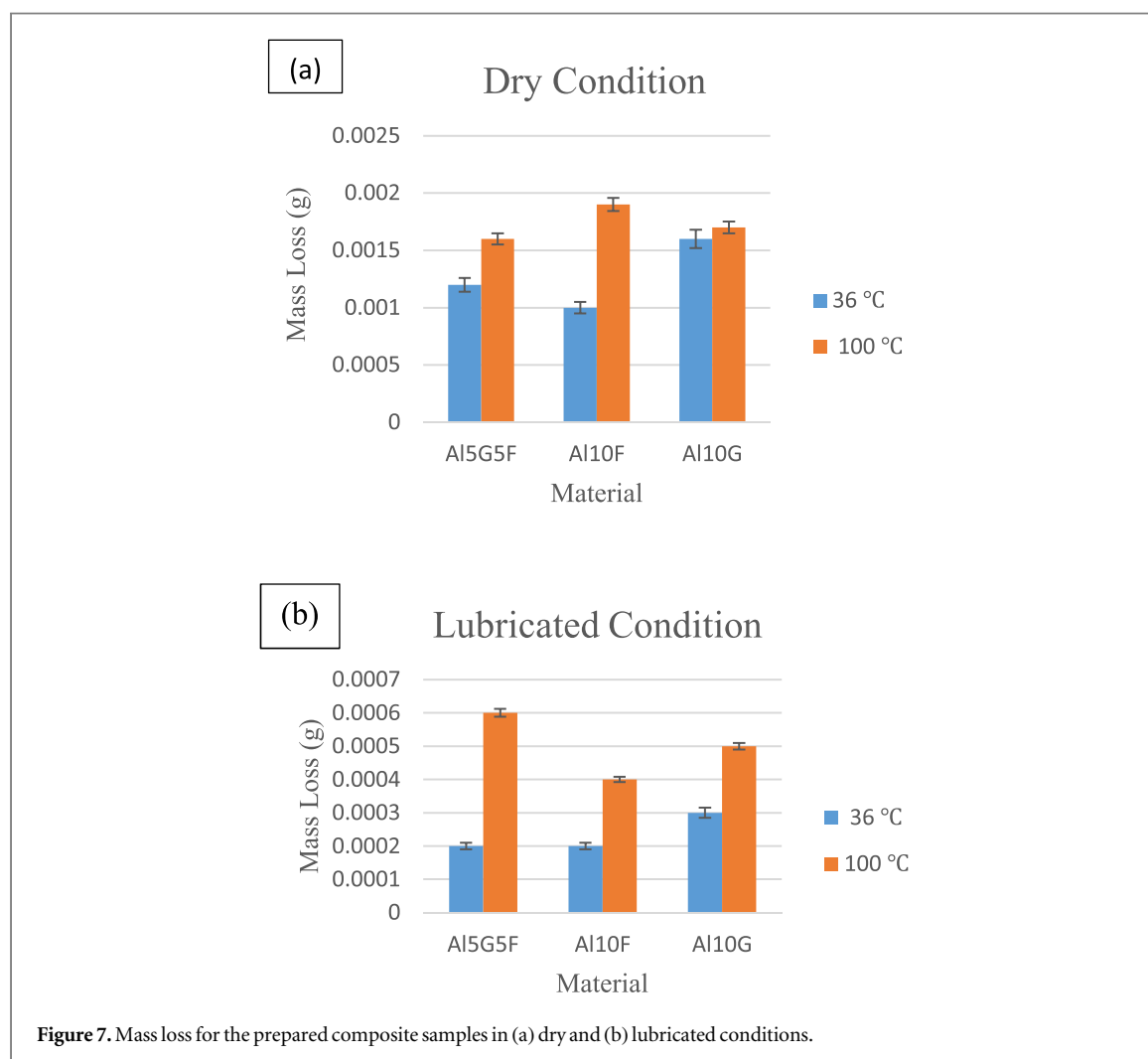


Figure 6. Comparison of the coefficient of friction value at 36 °C and 100 °C in lubricated condition for (a) Al10F composite (b) Al10G composite (c) Al5F5G composite.

The Al10G represents better tribological properties as compared to other composites at 100 °C. It is worth mentioning here that the Al10F produced the lowest COF value at 36 °C however, at a higher temperature 100 °C it resulted in the highest COF value.

The high hot hardness value of the graphite particles and release of environment contaminations might help in improving the overall COF value of the Al10G composite. Also, at high temperatures, the Al matrix gets soften and interfacial stress gets surpasses the



bonding strength between the matrix and reinforcement. This might lead to the generation of cracks and hence the mass loss increases.

The sliding started when the asperities of the hardened chromium steel ball surface makes a contact with the asperities of the composite surface. In the initial stages of the experimentation, the wear from the composite surface generates by the disintegration of the surface asperities by the hard asperities of the ball surface. As the sliding progresses, the surface of the composites becomes soft and asperities of the ball surface easily penetrate through the surface of the composites, and hence the friction coefficient increases.

For a better understanding of the effects of graphite and fly ash, the individual graphs for the prepared composites examined at 36 °C and 100 °C are presented in figures 4(a)–(c) and the average COF values for testing of different composites at 36 °C and 100 °C is given in table 4.

3.2.2. Coefficient of Friction in Lubricated Condition

In the lubricated condition, SAE 15W-40 lubricating oil was used at the interface of the ball and composite plate [23]. Lubricating oil was applied only once at the interface of the plate and ball. Figure 5 presents the

variation of the coefficient of friction with time. In this lubrication regime, Al10F material performed better as compared to the other composites at 36 °C and 100 °C. The average COF values obtained from the wear and friction test between the friction pairs in lubricated conditions are presented in table 5 and figure 6. It is observed that the coefficient of friction value rises to the maximum for all the prepared composites thereafter it stabilizes. The Al10F composite yields the lowest value of COF value during the tribo-testing at 36 °C and 100 °C.

The performance of the composite material gets affected by the working temperature. With increasing temperature, the lubrication applied between the prepared composite and ball surface gets heats up, and chances of shear thinning of the lubricating oil arise. Since there was no continuous supply of the lubricant at the interface of the tribo-pairs, so starved lubrication regime would get created. This phenomenon would reduce the layer of lubrication between the tribo-pairs and the effects of lubrication.

3.3. Wear behaviour

The wear loss from the composites was measured by weighing the specimens before and after the testing. A

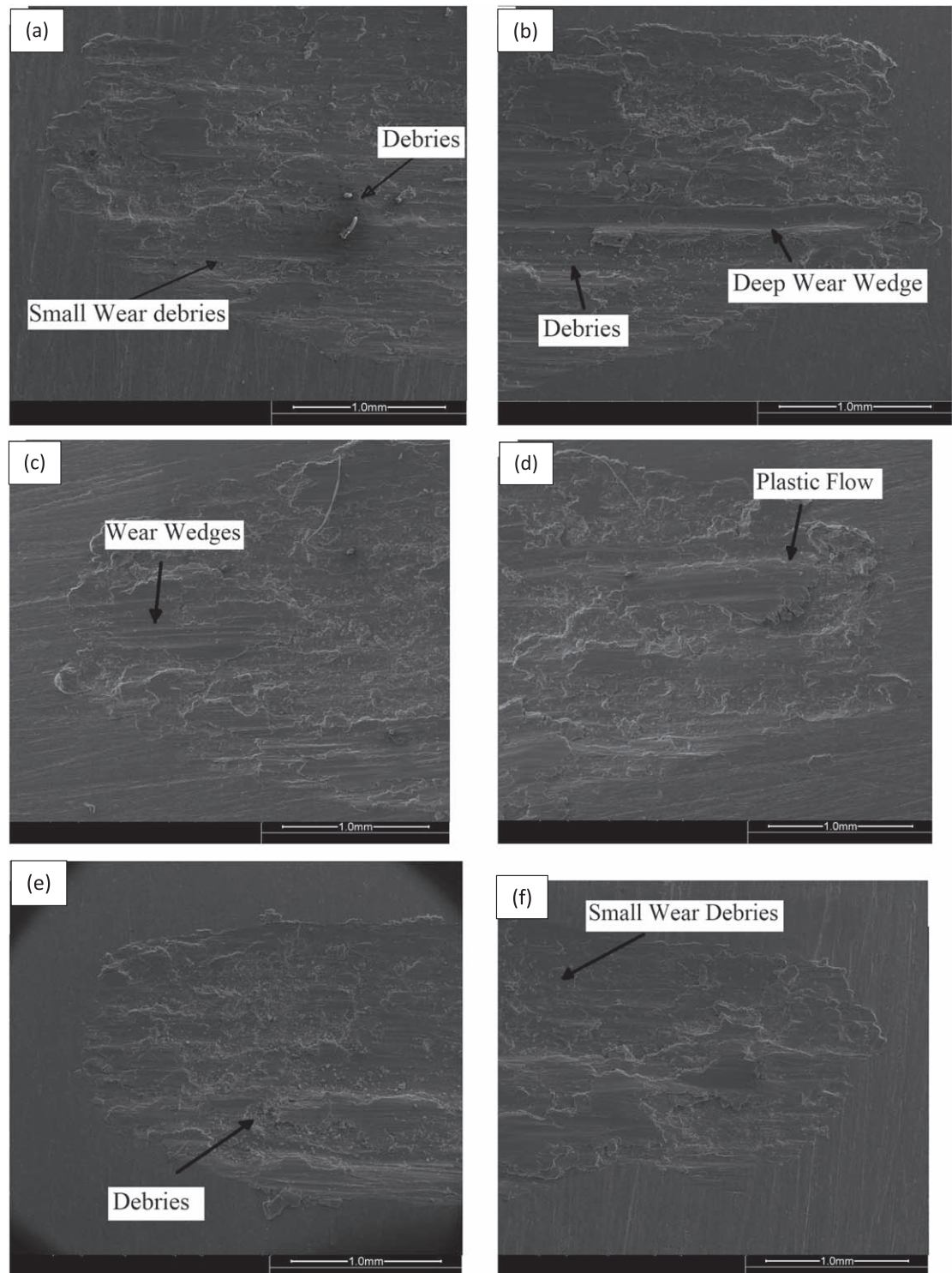


Figure 8. The scanning electron microscopic images (a) Al10F at 36 °C (b) Al10F at 100 °C (c) Al5F5G at 36 °C (d) Al5F5G at 100 °C (e) Al10G at 36 °C (f) Al10G at 100 °C.

weighing balance with 0.0001g accuracy was used to weigh the samples. Figure 7 presents the variation of mass loss for the prepared composite materials in dry and lubricated conditions. In the dry condition, the Al10F composite exhibits the lowest amount of mass loss at 36 °C. The fly ash particles prevent the wear of

the aluminum matrix. At higher testing temperatures, the Al10G composite yields minimum wear loss. For graphite-reinforced composites, the effects of solid lubricant and wear protective layer comes into play, which reduced the wear from the composite materials [29]. Also, as the sliding between the composite

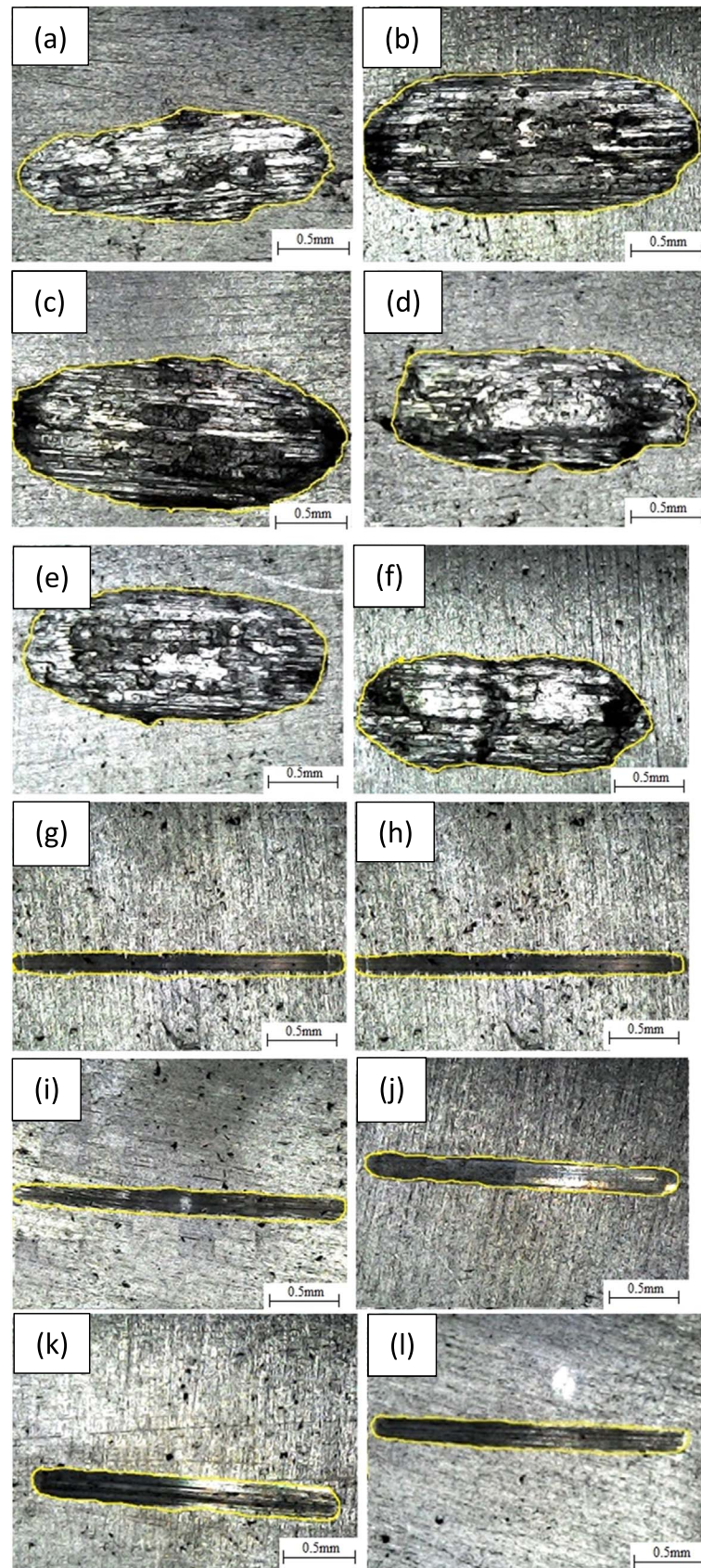


Figure 9. Presentation of wear surfaces for, dry sliding test of (a) Al5F5G at 36 °C, (b) Al10F at 36 °C, (c) Al10G at 36 °C. Dry sliding test of (d) Al5F5G at 100 °C, (e) Al10F at 100 °C, (f) Al10G at 100 °C. Lubricated sliding test of (g) AL5G5F at 36 °C, (h) Al10F at 36 °C, (i) Al10G at 36 °C. Lubricated sliding test of (j) Al5F5G at 100 °C, (k) Al10F at 100 °C, (l) Al10G at 100 °C.

material and hardened ball started, the material from the composite material starts shearing off by the asperities of balls. With this, the wear debris gets transferred to the ball surface. In the case of graphite-reinforced composite, this debris gets agglomerated between the composite and ball and prevents wear.

Figure 7(b) presents the variation of mass loss in the lubricated condition. In lubricated conditions, the COF value was reduced for all the composite materials as compared to the dry sliding. The amount of mass loss from Al10F is low as compared to other composites at 36 °C as well as 100 °C. The thin lubricating film between the composite and ball helped in reducing the mass loss [30]. As the sliding time is increased, the lumps of the wear debris get to mix with the lubricating oil and prevent the adhesion wear.

3.4. Worn surface analysis

A wear scar on the surface of the composite materials characterized the worn surface. This wear scar indicated the contact point of the ball surface. For the easy identification of the wear scar, scanning electron microscope images were obtained for the wear surfaces. The images were obtained at an accelerating voltage of 20 kV, an emission current of 47800 nm, and at a working distance of 6900 μm with a ZEISS EVO Series Scanning Electron Microscope. All the images were captured at 100 X magnification. All the worn surfaces had small pits, grooves, and plastic flow of the material. The worn surfaces of composites obtained during dry sliding testing at 36 °C and 100 °C are presented in figure 8. Figures 8(a)–(b) presents the scanning electron microscope image for Al10F composite, some loose debris was seen on the worn surface which was embedded into the composite surface. Small wear wedges were also visible for the 36 °C tested specimen. These wear wedges were larger as the testing temperature increases to 100 °C. The worn surface of the Al5F5G composite tested at 36 °C and 100 °C was presented in figures 8(c)–(d). The debris was seen on the surface with more plastic flow of the material. Figures 8(e)–(f) represents the worn surface for the Al10G composite. The plastic flow of material with a small cavity was seen for the Al10G composite at 36 °C. A small amount of debris was also visible on the surface. Very fine debris was seen during the dry sliding testing of the Al10G composite, which indicates abrasive wear.

The optical microscopic images of the wear specimens are shown in figures 9(a)–(l). These were obtained at 10 \times magnification. Figures 9(a)–(c) presents the wear region for the composite samples examined at 36 °C in dry lubricated conditions. Some grooves and major mass loss were reported for these samples. The wear region at 100 °C for dry lubricated conditions is presented in figures 9(d)–(f). At 100 °C the composite became slightly soft and deep grooves were present on the wear region. Figures 9(g)–(l)

shows the wear region in lubricated conditions at 36 °C and 100 °C temperature. A similar trend of wear was noticed for the wear region.

4. Conclusions and future scope

In this experimental study, a hybrid aluminium composite with fly ash and graphite as reinforcements was fabricated using the stir casting method. Dry and lubricated sliding tests were conducted at 36 °C and 100 °C temperatures to evaluate the wear and friction behaviors. The following conclusions could be drawn from the present study.

1. The tribological study in dry conditions revealed that at low temperatures Al10F composite resulted in the lowest coefficient of friction value, however, at higher temperatures, Al10G resulted in the lowest coefficient of friction value.
2. During the lubricating conditions, Al10F yields the lowest friction coefficient value at 36 °C as well as 100 °C.
3. The Al10F composite resulted in a minimum amount of mass loss at 36 °C during the dry sliding testing as well as in lubricated conditions.
4. The scanning electron microscopic images indicated abrasive wear nature with small debris and plastic flow for all the fabricated composites.
5. The wear performance of a composite material also depends on its strength and toughness values. The proposed composites may be tested to evaluate their tensile, flexural, and compressive strength as future scope of the present study.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

ORCID iDs

Vipin Kumar Sharma  <https://orcid.org/0000-0002-7356-4590>

References

- [1] Kheirkhah S, Imani M, Aliramezani R, Zamani M H and Kheilnejad A 2019 Microstructure, mechanical properties and corrosion resistance of Al6061/BN surface composite prepared by friction stir processing *Surf. Topogr.: Metrol. Prop.* **7** 035002
- [2] Sharma V K, Singh R C and Chaudhary R 2017 Effect of fly ash particles with aluminium melt on the wear of aluminium metal matrix composites *Eng. Sci. Technol. an Int. J.* **20** 1318–23
- [3] David Raja Selvam J, Dinaharan I and Mashinini P M 2017 High temperature sliding wear behavior of AA6061/fly ash

- aluminum matrix composites prepared using compocasting process *Tribology - Materials, Surfaces & Interfaces* **11** 39-46
- [4] Hashemi R and Hussain G 2015 Wear performance of Al/TiN dispersion strengthened surface composite produced through friction stir process: a comparison of tool geometries and number of passes *Wear* **324–325** 45–54
- [5] Banerjee S, Poria S, Sutradhar G and Sahoo P 2020 Abrasive wear behavior of WC nanoparticle reinforced magnesium metal matrix composites *Surf. Topogr.: Metrol. Prop.* **8** 025001
- [6] Farghadani M, Karimzadeh F, Enayati M H, Naghshehkhesh N and Ostovari Moghaddam A 2020 Fabrication of AZ91D/Cu/Mg₂ Cu and AZ91D/Mg₂ Cu/MgCu₂/MgO in-situ hybrid surface nanocomposites via friction stir processing *Surf. Topogr.: Metrol. Prop.* **8** 045002
- [7] Sharma V K, Singh R C and Chaudhary R 2018 Wear and friction behaviour of aluminium metal composite reinforced with graphite particles *Int. J. Surf. Sci. Eng.* **12** 419–32
- [8] Sharma V K and Singh R C 2018 Wear and friction behaviour of aluminium metal composite reinforced with graphite particles *Int. J. Surf. Sci. Eng.* **12** 419–32
- [9] Rajesh and Mahendra K V 2019 Development of Al-6Mg fly ash-graphite hybrid metal matrix composites by stir casting and evaluation of mechanical properties *International Journal of Current Engineering and Technology* **9** 669–72
- [10] Palanikumar K, Eaben Rajkumar S and Pitchandi K 2019 Influence of primary B₄C particles and secondary mica particles on the wear performance of Al6061/B₄C/mica hybrid composites *J. Bio- Tribo- Corrosion.* **5** 1–12
- [11] Pitchayappillai G, Seenikannan P, Raja K and Chandrasekaran K 2016 Al6061 hybrid metal matrix composite reinforced with alumina and molybdenum disulphide *Adv. Mater. Sci. Eng.* **2016** 1-9 Article ID 6127624
- [12] Gopinath S, Prince M and Raghav G R 2020 Enhancing the mechanical, wear and corrosion behaviour of stir casted aluminium 6061 hybrid composites through the incorporation of boron nitride and aluminium oxide particles *Mater. Res. Express* **7** 016582
- [13] Rohtagi P K 1993 Synthesis and metal matrix composites containing fly ash, graphite *Glass, Ceramics or Other Metals. Patent No.* 5228494
- [14] Magibalan S, Senthilkumar C, Prabu M, Yuvaraj S and Balan A V 2020 Optimization and effect of load, sliding velocity, and time on wear behavior of AA8011 - 8 wt.% fly-ash composites *Surf. Topogr.: Metrol. Prop.* **8** 045022
- [15] Kumar V M and Venkatesh C V 2019 A comprehensive review on material selection, processing, characterization and applications of aluminium metal matrix composites *Mater. Res. Express* **6** 072001
- [16] Mahanta S, Chandrasekaran M, Samanta S and Arunachalam R 2019 Multi-response ANN modelling and analysis on sliding wear behavior of Al7075/B₄C/fly ash hybrid nanocomposites *Mater. Res. Express* **6** 0850h4
- [17] Sharma V, Joshi R, Pant H and Sharma V K 2020 Improvement in frictional behaviour of SAE 15W-40 lubricant with the addition of graphite particles *Material Today: Proceedings.* **25** 719–23
- [18] Shabani M O, Mazahery A, Bahmani A, Davami P and Varahram N 2011 Solidification of A356 Al alloy: experimental study and modeling *Kovove Mater.* **49** 253258
- [19] Shabani M O and Mazahery A 2013 Suppression of segregation, settling and agglomeration in mechanically processed composites fabricated by a semisolid agitation processes *Trans. Indian Inst. Metals.* **66** 5–70
- [20] Shabani M, Mazahery A, Davami P and Razavi M 2012 Silicon morphology modelling during solidification process of A356 Al alloy *Int. J. Cast. Metals Res.* **25** 53–8
- [21] Prakash K S, Moorthy R S, Gopal P M and Kavimani V 2016 Effect of reinforcement, compact pressure and hard ceramic coating on aluminium rock dust composite performance *Int. J. Refract Metal Hard Mater.* **54** 223–9
- [22] Singh R C, Chaudhary R and Sharma V K 2019 Fabrication and sliding wear behavior of some lead-free bearing materials *Mater. Res. Express* **6** 066533
- [23] Lal R and Singh R C 2018 Experimental comparative study of chrome steel pin with and without chrome plated cast iron disc in situ fully flooded interface lubrication *Surf. Topogr.: Metrol. Prop.* **6** 035001
- [24] Latha Shankar B, Anil K C and Karabasappagol P J 2016 A Study on effect of graphite particles on tensile, hardness and machinability of aluminium 8011 matrix material *IOP Conf. Series: Materials Science and Engineering* **149** 012060
- [25] ElGhazaly A, Anis G and Salem H G 2017 Effect of graphene addition on the mechanical and tribological behavior of nanostructured AA2124 self-lubricating metal matrix composite *Composites Part A: Applied Science and Manufacturing* **95** 325–36
- [26] Lu J, Song Y, Hua L, Zhou P and Xie G 2019 Effect of temperature on friction and galling behavior of 7075 aluminum alloy sheet based on ball-on-plate sliding test *Tribol. Int.* **140** 105872
- [27] Ghiotti A, Bruschi S, Sgarabotto F and Medea F 2014 Novel wear testing apparatus to investigate the reciprocating sliding wear in sheet metal forming at elevated temperatures *Key Eng. Mater.* **622–623** 1158–65
- [28] Monikandan V V, Rajendrakumar P K and Joseph M A 2020 High temperature tribological behaviors of aluminium matrix composites reinforced with solid lubricant particles *Trans. Nonferrous Met. Soc. China* **30** 1195–210
- [29] Mehta V R and Sutaria M P 2020 Effect of temperature on wear and friction behavior of as-cast and heat treated LM25/SiC aluminum matrix composites *World Journal of Engineering* **18** 206–16
- [30] Kumar Singh K, Singh S and Kumar Shrivastava A 2016 Study of tribological behavior of silicon carbide based aluminum metal matrix composites under dry and lubricated *Environment, Advances in Materials Science and Engineering* **1–11**



E-FUCA: enhancement in fuzzy unequal clustering and routing for sustainable wireless sensor network

Pawan Singh Mehra¹

Received: 13 August 2020 / Accepted: 29 April 2021
© The Author(s) 2021

Abstract

With huge cheap micro-sensing devices deployed, wireless sensor network (WSN) gathers information from the region and delivers it to the base station (BS) for further decision. The hotspot problem occurs when cluster head (CH) nearer to BS may die prematurely due to uneven energy depletion resulting in partitioning the network. To overcome the issue of hotspot or energy hole, unequal clustering is used where variable size clusters are formed. Motivated from the aforesaid discussion, we propose an enhanced fuzzy unequal clustering and routing protocol (E-FUCA) where vital parameters are considered during CH candidate selection, and intelligent decision using fuzzy logic (FL) is taken by non-CH nodes during the selection of their CH for the formation of clusters. To further extend the lifetime, we have used FL for the next-hop choice for efficient routing. We have conducted the simulation experiments for four scenarios and compared the propound protocol's performance with recent similar protocols. The experimental results validate the improved performance of E-FUCA with its comparative in respect of better lifetime, protracted stability period, and enhanced average energy.

Keywords Clustering · Cluster head · Energy efficient · Fuzzy logic · Lifetime · Wireless sensor network

Introduction

The convergence of massive advancement in embedded computing, wireless communication and diverse sensor technology has fostered the emergence of WSN very swiftly. A WSN consists of enormous tiny devices called sensors to monitor the required field. A simple pictorial representation of WSN is shown in Fig. 1. There are numerous WSN applications, e.g. industrial monitoring, structural monitoring, climatic monitoring, defence, environmental monitoring, and health care [1, 2]. With the miniature size of Sensor Node (SN), there is a restriction of limited energy, storage, communication and computation. WSN has constraints in energy, computation and communication [3, 4]. Battery-operated SN depletes energy because of long-distance transmission to BS and redundant data processing. A single node failure may throw the network into an unreliable state. Thus, reducing energy consumption is a challenging issue that has attracted many researchers. Routing techniques which are

capable of reducing energy consumption are highly desirable. Cluster-based routing has proved to be a promising approach [5–7]. Election of CH and formation of clusters are crucial parts in dragging out the lifespan of the network.

The incorporation of FL helps in efficiently handling the decision-making behaviour of human solving uncertainty. Since there are several overlapping parameters that affect energy consumption, thus, this uncertainty can be driven by FL. Furthermore, FL possesses the potential to deal with imprecision in data and conflicting situations using heuristic human reasoning without requiring a complex mathematical model [8]. Regardless of the evidential advantages of FL by its widespread successful deployment in diverse domains, there is a comparatively limited number of fuzzy-based routing algorithms than fuzzy-based clustering algorithm. Since most of the cluster-based routing protocols require only a simple decision-making process (i.e. single-hop transmission of data from CH to BS), and hence the use of FL is unnecessary. However, for energy-aware clustering and routing demand comprehensive decision-making, FL represents an effective approach [9].

A fuzzy-based system has four primary modules: Fuzzifier maps the crisp input value to fuzzy linguistic value along with the assignment of membership function (MF).

✉ Pawan Singh Mehra
pawansinghmehra@gmail.com

¹ Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

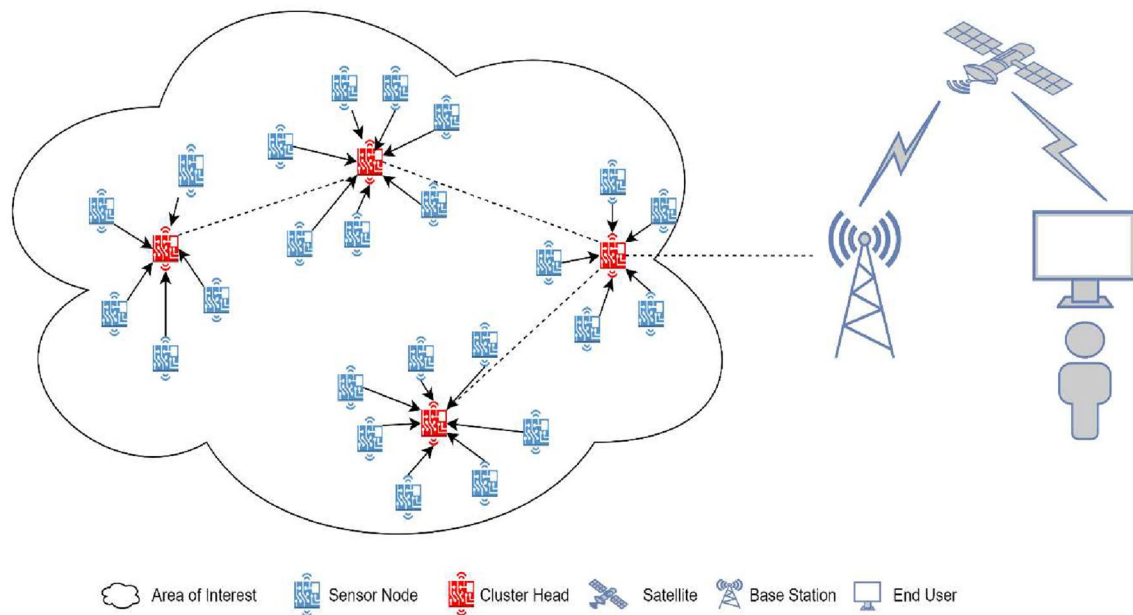


Fig. 1 Representation of WSN

Knowledge Base has a set of IF–THEN rules or conditions made by the user, which is considered by Inference Engine while making decisions and inferring or drawing conclusions. A fuzzy set is acquired by the Defuzzifier mapping it into a crisp output value.

Contribution of paper

This paper propounds E-FUCA, which is a distributed protocol for unequal clustering approach for protracting the stability span by balancing the load. The contributions made in this paper are as follows.

- Maximal clustering protocols are probabilistic and elect CH based on larger residual energy, its aloofness from BS and density of node, which is insufficient for electing the suitable candidate for CH.
- E-FUCA is an enhancement over the FUCA [10] protocol.
- FUCA contemplates the remnant power, nearness to BS and node density for calculating rank and competition radius during CH election whereas, in E-FUCA, instead of node density, the average distance to communicating nodes is considered because node density is incapable of giving a complete insight of energy expenditure by CH. Still, average distance can provide a clear idea of communication cost to be carried by CH if selected.
- In FUCA, during the formation of clusters, non-CH nodes make the greedy decision of choosing the nearest CH without any consideration of CH's existing load.

In contrast, in E-FUCA, non-CH nodes do not make a greedy decision but choose their CH intelligently based on its rank obtained during CH selection, closeness to that CH and number of nodes in its cluster radius obtained during CH selection.

- FUCA considers BS near/corner of the target field, whereas E-FUCA considers BS's location at the centre and at far off place.
- For unequal clustering, we consider FL in the routing of information from CH to BS.
- We have designed FIS for the selection of next-hop so that energy efficiency can further be enhanced.
- For gauging the performance of E-FUCA, simulation experiments are performed and obtained results are contrasted with the state of the art approaches such as FUCA [10], LEACH [11] and URBD [12] protocol. Experimental results validate the prolonged stability period, larger average energy with load balancing.
- The complexity of the proposed E-FUCA in terms of time and message is discussed and computed.

The rest of this paper is summarised as follows: discussion on pertinent work is done in “[Pertinent work](#)”. System and Energy model description is provided in “[Wireless sensor network model](#)”; the description of the proposed E-FUCA protocol is done in “[The proposed approach: E-FUCA](#)”. Simulation experiment and evaluation of performance is shown in “[Simulation experiment and result analysis](#)”. Lastly, a summary of the proposed E-FUCA protocol with concluding remarks is discussed in “[Conclusion](#)”.

Pertinent work

Energy efficiency is the demanding task of WSN, which can be provided by the clustering approaches. Some of the pertinent unequal clustering approaches are discussed in this section. The literature survey of any cluster-based proposed work is incomplete without the discussion of the LEACH [11] protocol. In the year 2000, Heinzelman et al. propound LEACH, which makes local decisions by adopting a probabilistic method for the selection of CH. For balancing the network load, CHs are rotated in each round as static CH prematurely expires in comparison to non-CH nodes in the network. The data aggregation is done at the CH level to minimise communication cost. Limitations of this protocol are that the CH selection is purely randomised, and crucial factors such as residual energy and aloofness from BS, which affect energy, are not put into consideration.

PRODUCE [13] protocol is proposed for eliminating the hot spot problem makes use of local probabilities for cluster formation of unequal size. CH nearer to BS focuses on inter-cluster communication, whereas CH at a distant place may focus on intra-cluster communication. It successfully balances the network load and prolongs the lifetime. EDUC [14] protocol, which is a distributed algorithm, evades the hot spot problem and energy dissipation in heterogeneous WSN. It involves the energy-driven rotation method of clusters. Every node in this protocol gets an opportunity to be CH in its lifetime. This method is not useful in multi-hop networks. LUCA [15] is based on probability to prevent the hotspot problem. The size of clusters varies with remoteness to BS. GPS is bundled with SN and is location-aware. A backoff timer is there with the randomised initial value. If an SN receives a message from CH, it joins it; else, it will proclaim its candidature. EADUC [16] protocol is designed to gather data periodically in WSN. The weight for CH candidature is based on remnant energy along with the degree of node and exhibits better performance in terms of lifetime. CHEF [17] is a fuzzy-based protocol wherein there are two inputs for FIS: local distance and remnant energy of node. For evaluating the fuzzified inputs and calculating the chance of a node to be chosen as coordinator of the cluster, there are nine fuzzy rules. CHUFL [18] is distributed protocol in which fuzzy inputs are remnant energy, reachability and distance to the BS. The non-CHs choose the nearest CHs to form clusters. A distributive clustering protocol, namely FBECs [19], is proposed, which assigns a pre-defined probability to SN based on distance from BS. It uses FL for the selection of adequate nodes for the role of CH.

An FL-based clustering algorithm is proposed in EAUCF [20], which uses remnant energy and remoteness to BS for electing CH. Nine IF–THEN fuzzy rules are used for selecting tentative CH. Competitive radius is calculated by each

tentative CH its candidature. But this proposed work does not anticipate energy exhaustion due to large intra-communication resulting in fading the protocol performance. An improvement over EAUCF is FBUC [21], in which the tentative CHs are selected on a probabilistic method. Competition radius and Chance are computed during the CH election. For cluster formation, the non-CH nodes calculate the chance of each CH on the basis of density and distance to CH. The protocol achieves a better lifetime in comparison to LEACH and EAUCF. The proposed IFUC [22] protocol is capable of reducing energy consumption and lengthening network lifetime. For nominating CHs and computing the range of the cluster, FL is used. The factors considered are remnant energy, closeness to BS and density of nodes. SN with a greater chance is selected as the final CH. DUCF [23] protocol makes load balancing certain by cluster formation using FL. The inputs to the fuzzifier are remoteness to BS, node density and remnant energy. There are two output variables, namely size and chance. The size of the cluster is dependent on the chance obtained. Mamdani method is used for inference.

A distribution independent unequal clustering is propounded in MOFCA [24], which contemplates remnant energy, calculated density and remoteness to BS for clustering. For reducing the intra-cluster relay, the cluster radius is varied as per the remoteness to BS. It successfully addresses the hotspot and energy hole problem. A diverse approach for cluster formation is proposed in MCFL [25]. Those candidates who are most eligible are chosen as CH, and no re-clustering takes place for a few rounds so as to reduce the message exchange for forming clusters. Experimental results exhibit the good performance of the proposed work than its comparatives.

FUCA [10] is a probabilistic approach for unequal clustering. Input variables considered are closeness to BS, remnant energy and density. There are two output variables: rank and competition radius. Higher ranking nodes in the competition radius are elected as CH. It achieves better performance than its counterparts. URBD protocol [12] is another unequal clustering protocol that is based on the density of nodes for clustering. There are two phases in this protocol: CH selection and Member-Join. In this protocol, density and distance parameters are used in collaboration for cluster formation resulting in a longer lifetime.

Tian et al. proposed LEACHEN [26], which introduces a multi-hop clustering cum routing method, which takes into account the fuzzy output as well multipath tree for improving the efficiency of the network. For the routing of information in multipath mode, three input variables are considered, i.e. remaining energy, traffic load and minimum hops. However, in a real-world scenario, other factors should be taken into consideration. AlShawi et al. proposed a routing algorithm [27], which includes FL and an A* algorithm for

lifetime enhancement. Remaining energy and Traffic load as the input parameter to the fuzzy system. Leabi and Abdalla [28] proposed a routing protocol using FL and an immune system that contemplates remnant energy and shortest hop for determining a route for communication to sink. This approach improves the efficiency of the network.

Jiang et al. proposed FLEOR [29] for optimised routing. FLEOR considers three factors in the fuzzy-based routing process. The inputs to the inference engine are the degree of closeness to sink (DCS), degree of closeness to the shortest path (DCSP) and degree of energy balance. NORIA [30] is a fuzzy-based routing protocol that considers a fuzzy rule set for parent election and role assignment in routing. The parameters fed to fuzzy systems are the number of hops to the BS and the remaining battery level. A fuzzy-based routing from node to sink is proposed in [28], which contemplates remaining energy and shortest hop as input variables to the fuzzy system for computing edge cost. The simulation results are compared with the Dijkstra routing technique. Haider and Yusuf proposed an energy-optimised approach based on FL [31]. They considered six input variables for the fuzzy system and computed the cost of the same. The simulation results exhibit a reliable and efficient approach, but if the size of the network grows, then the fuzzy system with six input variable will become more complex.

In the aforementioned approaches, a greedy decision is made by the non-CH nodes by choosing the nearest CH candidate for cluster formation, whereas some of these approaches use FL for calculating chance so that non-CH nodes may choose their respective CH based on the chance obtained. Most of these protocols do not consider routing along with clustering, which may limit the performance of the protocol. We have considered FL for all three cases, i.e. CH selection, cluster formation and routing for extending the lifetime of the protocol. Table 1 depicts the summary of the aforementioned approaches.

Wireless sensor network model

Maximising lifetime problem

Designing the architecture of WSN is a very challenging task as the SNs have limited power, computational capability and memory [3]. Energy consumption is the most significant among the three factors as the power source (battery) is irreplaceable. One of the promising solution to achieve energy efficiency is clustering [11]. In the cluster-based routing, deployed SNs are divided into clusters, and one of the SN plays the role of CH. If the CH is inefficient, then the protocol could not maximise energy efficiency [14]. FL has been used in the selection of efficient CHs [32], which has improved the lifetime of the WSN. Even if we select the

efficient CHs, then also in most of the protocols, the data are transmitted directly to BS, which limits the performance of the protocol. For maximising lifetime, energy-efficient clustering and routing algorithm have to be in place.

System model

In the proposed protocol, the network has homogeneous nodes with battery level at par, i.e. all the SNs are having the same energy level when they are deployed. The SNs have dispersed arbitrarily over the target field. Once the network gets operational, BS and SNs are immobile, i.e. neither the BS nor the SNs will change their location. There is a continuous power supply to the BS. The radio in SNs is capable of directional communication to conserve energy. The battery of SN is irreplaceable/non-rechargeable as in typical deployment, and the SNs are left unattended once deployed. Once the SNs are deployed, each SN will broadcast a hello_message. The separation distance between the two SNs is computed by the Received Signal Strength Index (RSSI). With RSSI, the SNs can estimate the location of other nearby SNs. SN is presumed to be lifeless only if the battery supply is fully drained. There is no constraint for BS in terms of processing and storage.

Energy consumption model

In E-FUCA, the radio energy model used in FUCA [10] is adapted. The energy of the WSN may get drained in transmission, amplification, reception, sensing, aggregation.

The energy dissipated for transmitting (E_{Tx}) and receiving (E_{Rx}) s bits over distance d is given by the following equations:

$$E_{Tx}(s, d) = \begin{cases} sE_{elec} + s\epsilon_{fs}d^2, & d < d_o \\ sE_{elec} + s\epsilon_{mp}d^4, & d \geq d_o \end{cases}, \quad (1)$$

$$E_{Rx}(s) = E_{Rx-elec}(s) = sE_{elec}, \quad (2)$$

where E_{elec} is the energy dissipated in electronic circuitry, d_o is a threshold that determines either free space (ϵ_{fs}) or multipath (ϵ_{mp}) model adopted and it can be calculated by the following equation:

$$d_o = \sqrt{\frac{\epsilon_{fs}}{\epsilon_{mp}}}. \quad (3)$$

In amplification of the signal, energy (E_{amp}) dissipated is calculated by the following equation:

Table 1 Comparison of pertinent work

Protocol	Parameters for CH selection	Parameters for cluster formation	Methodology for routing	Parameters for routing	Communication to BS	Remarks
CHEF [17]	Energy Local distance	Nearest CH is chosen by non-CH nodes	Not considered	–	CHs directly communicate to BS	It does not consider energy efficiency in cluster formation and routing
IFUC [22]	Energy Distance Density	Nearest CH is chosen by non-CH nodes	Ant colony optimisation	Remnant Energy Remoteness to BS	Multihop	Cluster formation could have been addressed efficiently
EAUCF [20]	Remnant energy Remoteness to BS	Nearest CH is chosen by non-CH nodes	Not considered	–	CHs directly communicate to BS	It does not consider energy efficiency in cluster formation and routing
CHUFL [18]	Remnant energy Distance Reachability	Nearest CH is chosen by non-CH nodes	Dijkstra algorithm	Distance	Multihop	Energy efficiency in cluster formation could have been addressed effectively
MOFCA [24]	Remnant energy Remoteness to BS Node degree	Nearest CH is chosen by non-CH nodes	Not considered	–	CHs directly communicate to BS	It does not consider energy efficiency in cluster formation and routing
DUCF [23]	Remnant Energy Remoteness to BS Node degree	Nearest CH is chosen by non-CH nodes	Not considered	–	CHs directly communicate to BS	It does not consider energy efficiency in cluster formation, routing and multihop communication
FBUC [21]	Remnant energy Remoteness to BS Node degree	Closeness to CH CH node degree	Not considered	–	CHs directly communicate to BS	It does not consider energy efficiency in routing and multihop communication
MCFL [25]	Remnant energy Node density	Closeness to CH Remnant energy of CH	Not considered	–	CHs directly communicate to BS	It does not consider energy efficiency in cluster formation, routing and multihop communication
CAFL [32]	Remnant energy Closeness to sink	Closeness to CH Remnant energy	Not considered	–	CHs directly communicate to BS	It does not consider energy efficiency in routing and multihop communication
FUCA [10]	Remnant energy Remoteness to BS Node density	Nearest CH is chosen by non-CH nodes	Not considered	–	CHs directly communicate to BS	It does not consider energy efficiency in cluster formation, routing and multihop communication
E-FUCA(Proposed)	Remnant energy Remoteness to BS Average distance for communication	CH rank Closeness to CH Total nodes in CH radius	Fuzzy logic	Next hop rank Nearness to next-hop Distance reduced to BS	Multihop	Energy efficiency issues in CH selection, cluster formation and routing are considered for prolong lifetime

$$E_{\text{amp}} = \begin{cases} \epsilon_{fs} d^2, & \text{if } d < d_o \\ \epsilon_{mp} d^4, & \text{if } d \geq d_o \end{cases} \quad (4)$$

For a CH, the amount of energy (E_{CH}) exhausted in a round is computed by the following equation:

$$E_{CH} = ns(E_{\text{elec}} + \epsilon_{fs} d_{CM} + E_{DA}), \quad (5)$$

where d_{CM} is the distance to cluster members and E_{DA} is the energy exhausted in data aggregation.

For a non-CH node, the energy (E_{nCH}) dissipated is computed by the following equation, in which d_{CH} is the distance from its CH:

$$E_{nCH} = s(E_{\text{elec}} + \epsilon_{fs} d_{CH}) \quad (6)$$

Decision variables

Residual energy is considered because a lower energy node is not suitable for CH candidature as it is a resource-intensive task. Residual energy can be calculated by

$$R_E(\text{Node}(i)) = \bar{E} - \bar{e}, \quad (7)$$

where \bar{E} is the initial energy level during deployment, and \bar{e} is the energy dissipated till now.

Closeness to BS is vital for consideration as CH candidates need to forward the accumulated data. If this distance is too long, then the node will dissipate more energy. The competition radius is inversely proportional to this distance as a closer node will have a larger radius as compared to the node at far off place from BS. The closeness to BS can be calculated as

$$\delta_{BS}(\text{node}(i)) = \sqrt{(\text{node}(i).x - BS.x)^2 + (\text{node}(i).y - BS.y)^2}. \quad (8)$$

Average distance is crucial in calculating rank because the intra-cluster communication cost is dependent on the

separation distance. The average distance from a node (i) can be computed as

$$\text{Avg_dis}(\text{node}(i)) = \frac{1}{m+1} \left\{ \sum_{k=1}^m d_k + \delta_{BS} \right\}, \quad (9)$$

where d_k is the distance to communicating nodes, δ_{BS} is the remoteness to BS.

Rank determines the candidature weight of an SN to become CH. The higher the rank of SN, the higher will be the probability of the SN to be selected as CH. The rank of each SN can be computed using FL, as shown in Fig. 2.

The closeness to CH is considered because, to reduce the intra-cluster communication cost, cluster members should be closer to CH. It can be calculated by

$$CN_CH(\text{node}(i)) = \sqrt{(\text{node}(i).x - CH_i.x)^2 + (\text{node}(i).y - CH_i.y)^2}, \quad (10)$$

where $\text{node}(i).x$, $\text{node}(i).y$ are x and y coordinates of the node and $CH_i.x$, $CH_i.y$ are the coordinates of CH under consideration.

The number of nodes in a cluster radius is useful to determine if the cluster is overcrowded, then it will increase the burden on CH as it has to expend more power in receiving data from a large number of SNs. It can be calculated by

$$ND(\text{node}(i)) = \sum_{j=1}^n \text{node}(j) \text{ s.t. } d(\text{node}(i), \text{node}(j)) \leq K'_i, \quad (11)$$

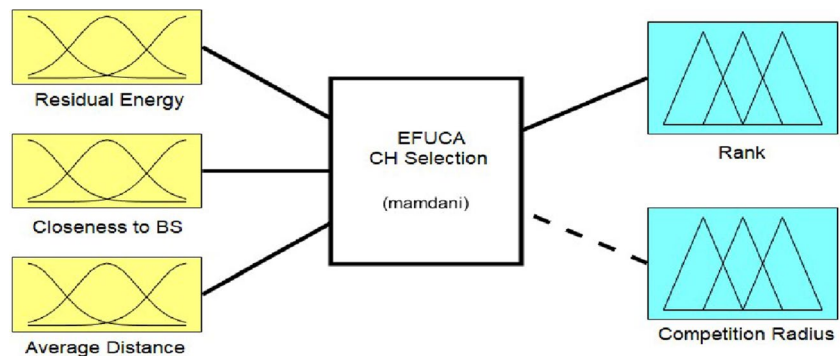
where $d()$ represents the distance between two nodes, and K'_i is the cluster radius of a node (i).

The distance reduced to BS is considered because we need to ensure that there is a significant reduction of distance after each hop. It can be calculated as

$$\dot{D}(\text{node}(i)) = \delta_{BS}(\text{node}(i)) - d(\text{node}(i), \text{node}(j)), \quad (12)$$

where δ_{BS} is the distance to BS, and $d()$ represents the distance between two nodes.

Fig. 2 FIS designed for CH selection in E-FUCA



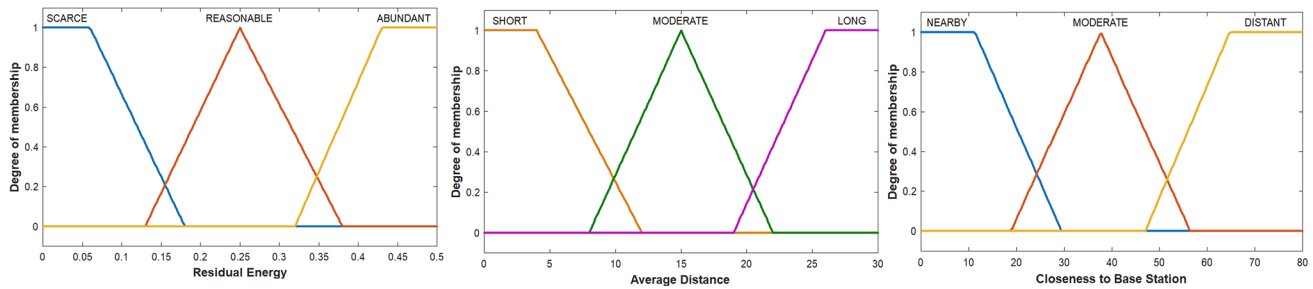


Fig. 3 MF for input variables in CH selection

The proposed approach: E-FUCA

The proposed protocol “E-FUCA” is designed to enhance the stability period to make the network more reliable as well as achieving a load-balanced network. E-FUCA is an improvement over the FUCA [10] protocol. The improvements can be enumerated in the following ways: first, In E-FUCA, for computing the rank of a node for CH candidature, the average distance to communicating nodes is calculated as one of the parameters together with remnant energy and aloofness to BS, unlike FUCA which considers node density, residual energy and aloofness from BS. Merely calculating the node density does not fulfil the requirement as the communication cost cannot be calculated only on the basis of node density. Some SNs may be nearer, and some SNs may be at far off place. Thus, to determine the nearest approximation of communication cost, the average distance may serve the purpose instead of node density. Second, in FUCA, during the cluster formation, non-CH nodes select the closest CH without determining the overall load on the CH candidate. In our protocol, the non-CH node will calculate the CH chance to determine which cluster must be joined. This CH chance is calculated on the basis of three parameters; the rank of CH, closeness to that CH and number of nodes in CH competition radius. This chance will help in minimising the extraneous energy dissipation in intra-cluster communication. Third, in FUCA, there is no focus on the routing of data, but in the proposed E-FUCA protocol, the fuzzy-based routing algorithm is designed to further prolong the network’s lifetime. The working of the designed protocol is partitioned into rounds. In each round,

there are three stages, selection of CH, Cluster formation and Data dissemination stage.

Selection of CH

In this phase, the selection of the CHs is decided on the basis of their characteristics. At the initiation of a round, a random number is generated by every node for becoming tentative CH. The threshold probability (T_{Prob}) is compared with the generated number. If the number is less than T_{Prob} , then the node becomes a tentative CH. Once the tentative CHs are determined, these nodes calculate their rank using designed FIS, as shown in Fig. 2.

The calculation of rank and competition radius is done using three input variables: remnant energy, the average distance to communicating nodes and closeness to BS.

Table 2 Linguistic variables for input and output in CH selection

Parameters	Linguistic variables
Residual energy	Scarce, reasonable, abundant
Distance to BS	Nearby, moderate, distant
Average distance	Sparse, medium, dense
Rank	Poor, below average, average, satisfactory, good, very good, extra ordinary
Competition radius	Very large, large, medium, medium large, medium, medium small, small, very small

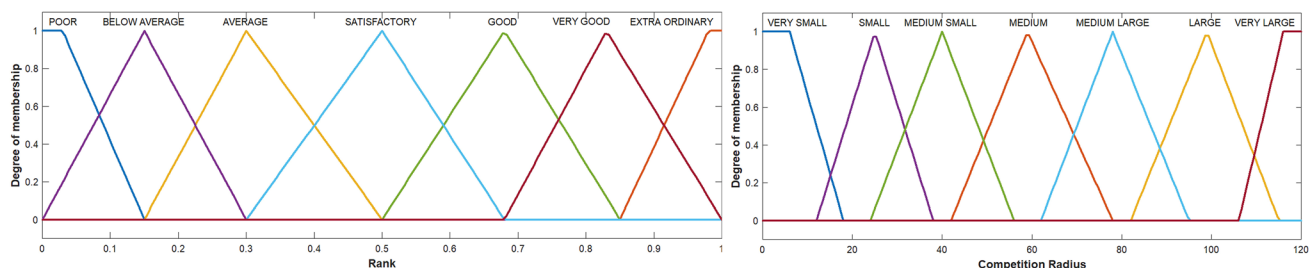


Fig. 4 MF for output variables in CH selection

Table 3 Fuzzy rules for rank and competition radius

Residual energy	Closeness to BS	Average distance	Rank	Competition radius
Scarce	Near	Long	Average	Medium small
Scarce	Near	Moderate	Average	Medium small
Scarce	Near	Short	Satisfactory	Medium
Reasonable	Near	Long	Satisfactory	Medium
Reasonable	Near	Moderate	Satisfactory	Medium
Reasonable	Near	Short	Good	Medium large
Abundant	Near	Long	Very good	Large
Abundant	Near	Moderate	Extra ordinary	Very large
Abundant	Near	Short	Extra ordinary	Very large
Scarce	Moderate	Long	Below average	Small
Scarce	Moderate	Moderate	Below average	Small
Scarce	Moderate	Short	Average	Medium small
Reasonable	Moderate	Long	Satisfactory	Medium
Reasonable	Moderate	Moderate	Satisfactory	Medium
Reasonable	Moderate	Short	Good	Medium large
Abundant	Moderate	Long	Good	Medium large
Abundant	Moderate	Moderate	Very good	Large
Abundant	Moderate	Short	Very good	Large
Scarce	Distant	Long	Poor	Very small
Scarce	Distant	Moderate	Poor	Very small
Scarce	Distant	Short	Below average	Small
Reasonable	Distant	Long	Below average	Small
Reasonable	Distant	Moderate	Below average	Small
Reasonable	Distant	Short	Satisfactory	Medium
Abundant	Distant	Long	Good	Medium large
Abundant	Distant	Moderate	Good	Medium large
Abundant	Distant	Short	Very good	Large

In FUCA, node density is considered. Node density cannot determine exactly the energy consumption by the CH node for intra-communication. This can be illustrated in Example 1.

Example 1 Suppose there are two nodes N1 and N2, competing for CH candidature. Their current energy level is 0.3 J, and closeness to BS is 150 m with equal node density as 10 (i.e. there are ten neighbouring nodes). FUCA protocol will generate equal rank for both the nodes N1 and N2 as all the values passed on to the FIS are the same because it does not consider the distance to the neighbouring nodes. In the case of the E-FUCA protocol, for N1 and N2, it will compute the average distance to all the communicating nodes. Thus, the rank generated for both the nodes N1 and N2 will be different, which will give a better perspective for CH candidature.

There are two output variables: rank and competition radius. *Rank* determines the candidature weight of an SN. The higher the rank of SN, the higher will be the probability of the SN to be selected as CH. *Competition radius* determines the radio range of a node within which it can communicate. It may vary according to the rank obtained by SN as a low energy node ought not to communicate to a longer

radio range as it will lead to quicker energy dissipation in intra-cluster communication. The MF plots for input and output variables are shown in Figs. 3 and 4.

We have used Trapezoidal and Triangular MF for boundary and intermediate variables, respectively, because they provide faster calculation and are simpler to implement. Each MF has to satisfy one condition that its degree of membership should range from 0 to 1. There are other MFs that can also be used like Sigmoid, Bell, Gaussian etc. but proposed E-FUCA depicted better results with Triangular and Trapezoidal MF. The linguistic variables which are used are shown in Table 2. The input variables are fed to the designed fuzzy inference system, and IF–THEN rules are applied to calculate rank and competition radius, which are described in Table 3. Here, the Mamdani Inference method [33] is applied, which is most commonly used [19, 34] because of its simplicity and characteristics.

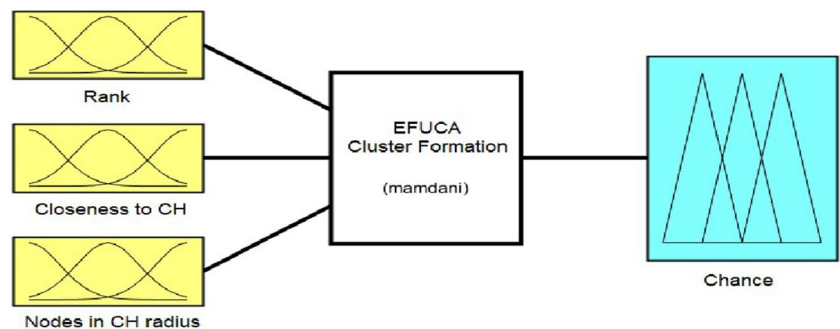
For defuzzification, the centre of area method is used to obtain crisp value from output linguistic variables. After the calculation of rank and competition radius, tentative CH nodes broadcast their candidature within the competition radius, and a higher ranking node is selected as CH. The selection procedure of CH is described in Algorithm 1.

Algorithm1: Selection of CH in E-FUCA**Begin:**

```

1 : S_EFUCA  $\leftarrow$  Total alive nodes in the network
2 : j  $\leftarrow$  SN Identity
3 : S_EFUCA (j).Energy  $\leftarrow$  current SN energy level
4 : S_EFUCA (j).AD  $\leftarrow$  Average distance to nearby nodes:
5 : S_EFUCA(j).Type  $\leftarrow$  member
6 : S_EFUCA (j).DTBS  $\leftarrow$  Distance of SN to BS
7 : S_EFUCA (j).Rank  $\leftarrow$  0 // initially rank set to 0
8 : S_EFUCA (j).Tent_CH  $\leftarrow$  False
9 : CH_List  $\leftarrow$  0 // Initially, there is no CH
10 : T_Prob  $\leftarrow$  Threshold probability for becoming tentative CH
11 :   For j=1 to S_EFUCA
12 :     S_EFUCA (j).R  $\leftarrow$  rand(0,1)
13 :     If S_EFUCA (j).R < T_Prob then
14 :       S_EFUCA (j).Tent_CH  $\leftarrow$  True
15 :       Compute rank and competition radius using FIS
16 :       Broadcast CH_MSG(ID, Rank, Comp_Radius)
17 :     End If
18 :   End For
19 :   For j=1 to S_EFUCA
20 :     If S_EFUCA (j).Tent_CH==True then
21 :       For k=1 to S_EFUCA
22 :         If S_EFUCA (k).Tent_CH==True then
23 :           If S_EFUCA (j).Rank > S_EFUCA (k).Rank
24 :             S_EFUCA(j).Type  $\leftarrow$  CH
25 :             CH_List  $\leftarrow$  j
26 :           End If
27 :         End If
28 :       End For
29 :     End If
30 :   End For

```

Terminate**Fig. 5** FIS designed for cluster formation in E-FUCA

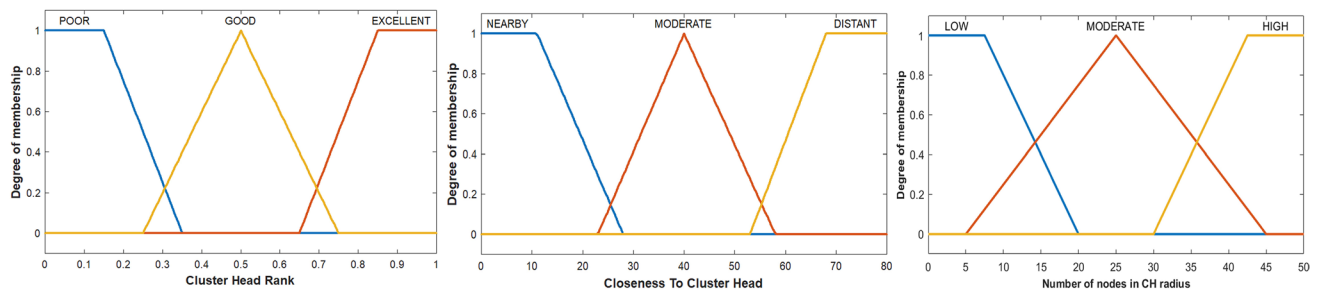
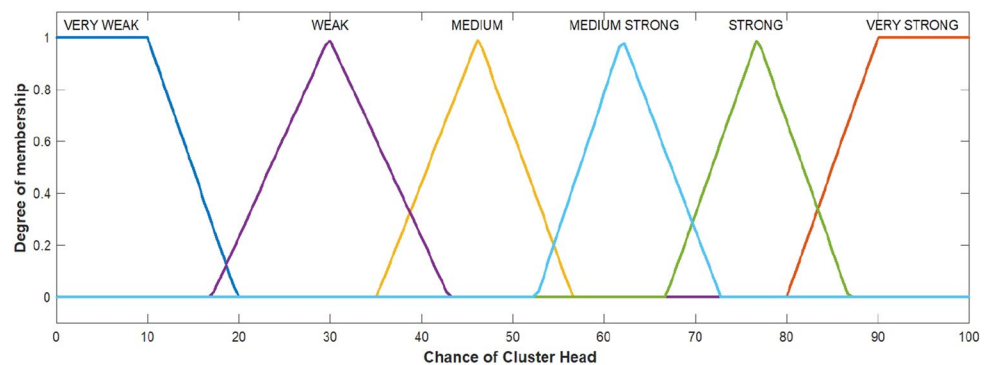


Fig. 6 MF for input variables in cluster formation

Fig. 7 MF for output variables in cluster formation



Cluster formation

After the completion of the CH selection procedure, all the nodes which are not selected for the CH role need to make the decision to join the appropriate cluster. In most of the protocols, the non-CH nodes make a greedy decision of joining the nearest CH without determining the load on that CH. In this proposed protocol, the overall load of the CH candidate is already determined by its rank, which is computed during the CH election on the basis of its closeness to BS, its current energy level and average distance to nearby nodes. The decision of choosing the CH by the non-CH node is supported by the designed FIS as shown in Fig. 5, and the MF functions used for the input variables and output variables are presented in Fig. 6 and Fig. 7.

Non-CH nodes calculate the chance of each CH on the basis of IF–THEN rules applied to the inputs: CH_Rank, number of nodes in the competition radius of CH and distance to that CH. Explanation to support this intelligent decision is described in Example 2.

Table 4 Linguistic variables for input and output in cluster formation

Parameters	Linguistic variables
Rank	Poor, good, excellent
Closeness to CH	Nearby, moderate, distant
Nodes in CH radius	Low, moderate, high
Chance	Very weak, weak, medium, medium strong, strong, very strong

Example 2 Suppose there are two CH nodes C1 and C2. A non-CH node (N1) needs to choose a CH between C1 and C2. Suppose rank of C1 = 2 and rank of C2 = 98. Distance from N1 to C1 is 14 m, and C2 is 16 m. According to FUCA protocol, N1 will take a greedy decision and directly choose C1 as its CH without considering its low rank, which could be due to low energy, a large number of neighbouring nodes and a large distance to BS. If all the nodes make greedy decisions like this, then it could result in more power dissipation

Table 5 Fuzzy rules for computing chance of CH

Rank	Closeness to CH	Nodes in CH radius	CH's chance
Poor	Distant	High	Very weak
Poor	Distant	Moderate	Very weak
Poor	Distant	Low	Very weak
Poor	Moderate	High	Weak
Poor	Moderate	Moderate	Weak
Poor	Moderate	Low	Weak
Poor	Near	High	Medium
Poor	Near	Moderate	Medium
Poor	Near	Low	Medium
Good	Distant	High	Very weak
Good	Distant	Moderate	Weak
Good	Distant	Low	Weak
Good	Moderate	High	Medium
Good	Moderate	Moderate	Medium strong
Good	Moderate	Low	Medium strong
Good	Near	High	Medium strong
Good	Near	Moderate	Strong
Good	Near	Low	Very strong
Excellent	Distant	High	Medium
Excellent	Distant	Moderate	Medium
Excellent	Distant	Low	Medium
Excellent	Moderate	High	Medium strong
Excellent	Moderate	Moderate	Strong
Excellent	Moderate	Low	Strong
Excellent	Near	High	Strong
Excellent	Near	Moderate	Very strong
Excellent	Near	Low	Very strong

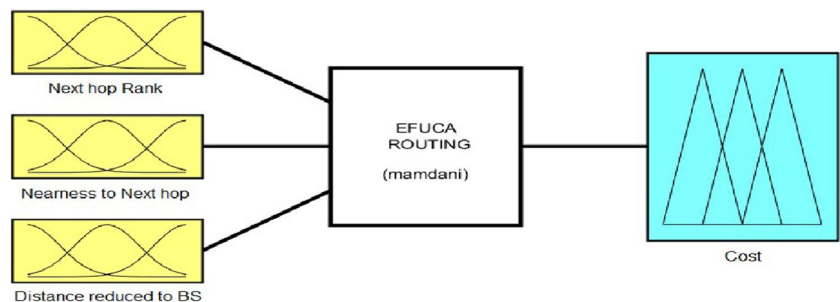
Table 6 Linguistic variables for input and output in routing

Parameters	Linguistic variables
Next hop rank	Low, average, high
Nearness to next-hop	Nearby, moderate, distant
Distance reduced to BS	Low, moderate, high
Cost	Very large, large, medium, medium large, medium, medium small, small, very small

as CH responsibility is a resource-intensive task. In the case of the E-FUCA protocol, node N1 will take this decision intelligently by considering the rank of CH, the number of nodes in the competition radius of CH and closeness to CH before choosing its CH. Finally, it will choose C2 as its CH, although C1 is closer to N1. This will result in reducing the load on low-rank CH nodes and balancing the energy dissipation by the CH nodes, thereby contributing an extension of the lifetime of the network.

The linguistic variables used in input and output variables are depicted in Table 4.

The IF–THEN rules applied for determining the chance of CHs are described in Table 5. After the calculation of the chance of each CH node, the non-CH node joins the CH, which is having the highest chance value by transmitting a join request (JOIN_REQ) message. The CH node accepts the request received from all non-CH nodes and forms the cluster. The cluster formation procedure is explained in Algorithm 2.

Fig. 8 FIS designed for routing in proposed E-FUCA protocol

Algorithm 2: Cluster formation in E-FUCA**Initiate:**

```

1 : TN_CH_LIST  $\leftarrow$  Total CH nodes whose packet is received by SN.
2 : j  $\leftarrow$  SN Identity
3 : For each CH_Node, calculate Chance using Fuzzy_Logic(CH_Rank, Distance_to_CH)
4 : OPTIMUM_CH  $\leftarrow$  0
5 : OPTIMUM_CH_CHANCE  $\leftarrow$  0
6 :   For k=1 to TN_CH_LIST
7 :     If CH_Node(k).Chance > OPTIMUM_CH_CHANCE then
8 :       OPTIMUM_CH  $\leftarrow$  k // ID of CH node
9 :       OPTIMUM_CH_CHANCE  $\leftarrow$  CH_Node(k).Chance
10 :    End If
11 :   End For
12 : S_EFUCA (j).CH= OPTIMUM_CH
13 : S_EFUCA (j) will send a packet (JOIN_REQ) to the CH node
14 : CH node will send ACK to S_EFUCA (j) with TDMA slot.

```

Terminate**Data dissemination**

Once the clustering process gets completed, the data dissemination stage begins. SNs sense the target area and generate the data on a periodic basis. SNs forward the collected data to their respective CH as per the TDMA slot for preventing loss of data in a collision. Once the CHs collect data from all their cluster members, it compresses the data prior to forwarding it to the BS. In most of the protocols, CHs forward data directly to BS, which depletes a large amount of energy of the CHs. For conserving the energy in forwarding the data from CH to BS, CH will make a decision using a designed FIS, which takes three inputs, namely, next-hop rank, nearness to next-hop and distance reduced to BS as shown in Fig. 8 for computing the cost of next-hop. The next-hop can

be one of the chosen CHs or BS. The LV for input and output variables are shown in Table 6. The MF for input and output variables is depicted in Figs. 9 and 10. The CH calculates the eligibility of every other next-hop CH nodes, which are in the direction of BS, using the IF-THEN rules designed for mapping inputs to output which are shown in Table 7.

With an objective to minimise the distance as well as preventing the intermediate CH nodes from overburden during data forwarding, the CH selects the next-hop CH node having maximum eligibility. Once the best next-hop CH node is selected, the current CH checks its remoteness to BS as well as the distance to the next-hop. If the distance to BS is shorter, then it will forward the data to the BS; else, it will forward the data to the next-hop. The process of forwarding the data is elaborated in Algorithm 3.

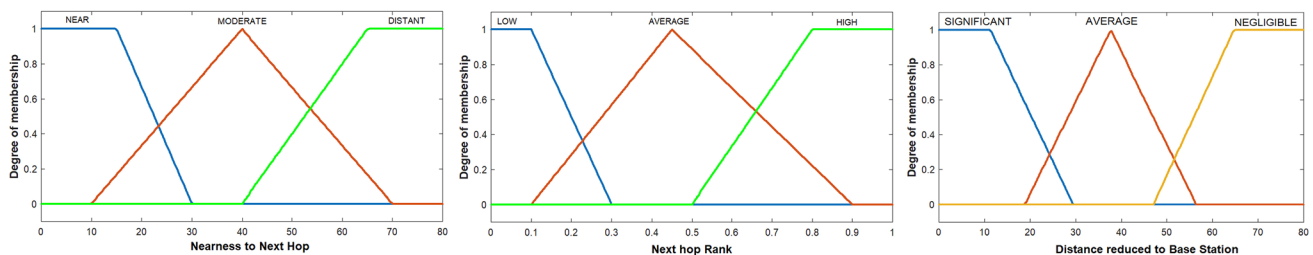
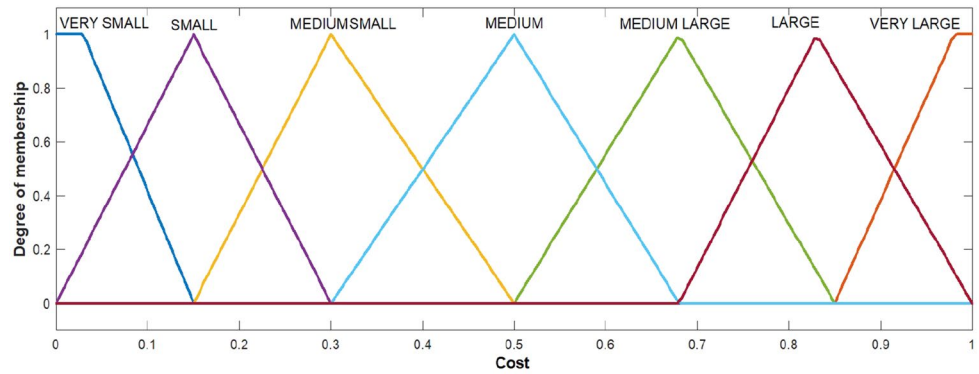


Fig. 9 MF for input variables in routing

Fig. 10 MF for output variables in routing**Algorithm 3:** Routing in EFUCA**Initiate:**

```

1 :  $TN\_CH \leftarrow$  Total CH nodes in one round.
2 :  $m, n \leftarrow$  index representing the ID of CHs
3 :  $CH(m) \leftarrow$  Cluster head with index  $m$  as its ID
4 :  $CH(m).DTBS \leftarrow$  distance from  $CH(m)$  to BS
5 :  $d(CH(m), CH(n)) \leftarrow$  distance between  $CH(m)$  and  $CH(n)$ 
6 : For each CH node  $m$  in  $TN\_CH$  do
7 :     For each CH node  $n$  in  $TN\_CH$  do
8 :         Broadcast  $CH\_MSG$  ( $SN(k).ID$ ,  $SN(k).Member\ Nodes$ ,  $SN(k).FF1$ )
9 :         If the direction of  $CH(n)$  is towards BS, then
10 :             Compute  $cost$  using FUZZY LOGIC (Next hop rank, nearness to next hop, distance reduced to BS)
11 :         End if
12 :     End For
13 :      $CH(n)$  with the lowest  $cost$  is selected
14 :     If  $CH(m).DTBS > d(CH(m), CH(n))$  then
15 :         Forward data to  $CH(n)$ 
16 :     Else
17 :         Forward data to  $CH(n)$ 
18 :     End if
19 : End For

```

Terminate

In this manner, all the CHs forward data to BS for further processing and completes one round of proposed work. For a better understanding of the complete flow of the proposed work, we have drawn a flow chart, as shown in Fig. 11, describing the steps involved in clustering and routing of proposed work.

Simulation experiment and result analysis

For evaluation of the proposed E-FUCA protocol, simulation experiments are performed extensively for E-FUCA, DEFL [8], URBD [12], FUCA [10] and LEACH [11] under

four scenarios in MATLAB, and experimental results are obtained. In scenario-1, the field size is chosen as 200×200 m² with 100 SNs having 1 J of initial energy and the position of BS is kept at a distant position from the field, i.e. (100, 300). In scenario-2, the field size is similar to scenario-1, and the BS is kept at the centre of the field, i.e. (100, 100). There are 200 SNs with an initial energy of 0.5 J. In scenario-3, the field size is 300×300 with 300 nodes with 0.5 J. The position of BS is kept at the bottom centre, i.e. (150, 0). In scenario-4, the field size is 500×500 with BS located at (0, 500), i.e. at the top-left position of the field. 500 SNs are deployed with 0.5 J of energy. All four scenarios are shown in Fig. 12. The reason behind choosing these four scenarios

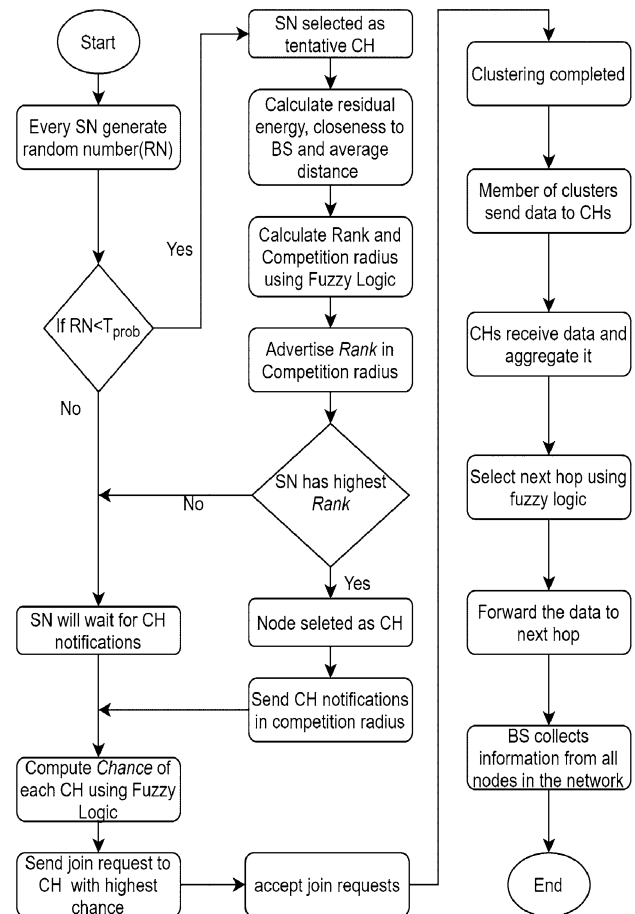
Table 7 Fuzzy rules for computing cost of next-hop

Next hop rank	Nearness to next-hop	Distance reduced to BS	Cost
Low	Distant	Negligible	Very large
Low	Distant	Average	Very large
Low	Distant	Significant	Large
Low	Moderate	Negligible	Very large
Low	Moderate	Average	Very large
Low	Moderate	Significant	Medium large
Low	Near	Negligible	Large
Low	Near	Average	Large
Low	Near	Significant	Medium large
Average	Distant	Negligible	Large
Average	Distant	Average	Medium large
Average	Distant	Significant	Medium
Average	Moderate	Negligible	Medium large
Average	Moderate	Average	Medium small
Average	Moderate	Significant	Small
Average	Near	Negligible	Medium
Average	Near	Average	Medium small
Average	Near	Significant	Small
High	Distant	Negligible	Medium
High	Distant	Average	Medium small
High	Distant	Significant	Medium small
High	Moderate	Negligible	Small
High	Moderate	Average	Very small
High	Moderate	Significant	Very small
High	Near	Negligible	Small
High	Near	Average	Very small
High	Near	Significant	Very small

is that the proposed protocol can be applied to any type of application wherein the position of BS either can be at the centre of the field or beyond the boundaries of the target area at a remote place. The experimental values considered for different parameters are stated in Table 8. For the evaluation and comparison of the E-FUCA with FUCA, LEACH, URBD and DEFL, the performance metrics chosen are Stability period, Total Average Energy, Total Alive nodes, Quarter Node Death (QND) and Half Node Death (HND).

Since the objective of WSN is to collect surrounding information, it is necessary that all the SN deployed should be alive so that cent per cent coverage is guaranteed. Reliability, in terms of coverage, is directly proportional to the stability period [35, 36]. Figure 13 exhibits the performance of E-FUCA, FUCA, LEACH, URBD and DEFL protocols in terms of Stability period for four scenarios.

The stability period determines the round in which the death of the first node occurred in the network [36]. The larger the stability period, the more the protocol will be

**Fig. 11** Flow chart of proposed E-FUCA protocol

reliable because of the complete coverage. We can see that for Scenario-1, the stability period of the E-FUCA protocol is 147.47%, 87.9%, 70.24% and 26.10% better than LEACH, FUCA, URBD and DEFL protocols, respectively. Similarly, for Scenario-2, it is 157.89%, 99.8%, 84.21% and 42.03% enhanced as compared to LEACH, FUCA, URBD and DEFL, respectively. The stability period of E-FUCA over LEACH, FUCA, URBD and DEFL is protracted by 282.50%, 130.94%, 59.38% and 47.12% for scenario-3 and 983.33%, 490.91%, 136.36% and 85.71% for scenario-4, respectively. The proposed E-FUCA has performed tremendously well in terms of stability period because not only the best candidate is chosen for the CH role, but also non-CH nodes take the intelligent decision of selecting the appropriate CH.

In Fig. 14, a graph for QND is plotted for four scenarios. In this graph, an assessment of the performance of the proposed E-FUCA in terms of the first quarter of nodes death can be seen. For scenario-1, E-FUCA has performed 84.33%, 72.15%, 29.38% and 21.19% better than LEACH, FUCA, URBD and DEFL protocols, respectively, and for

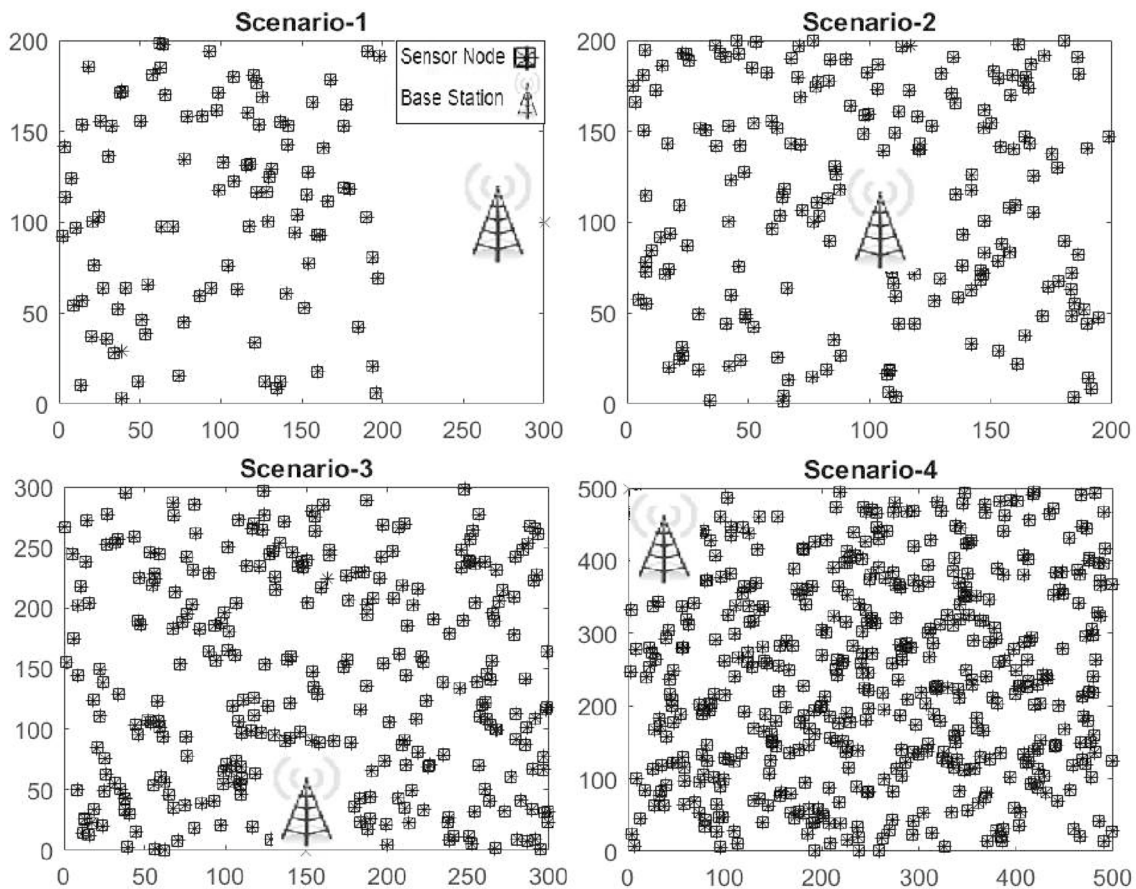


Fig. 12 Network scenarios for E-FUCA

Table 8 Description of parameters used for simulation

Parameters	Symbol	Values for Scenario-1	Values for Scenario-2	Values for Scenario-3	Values for Scenario-4
Total SN	N	100	200	300	500
Area	A	(200, 200)	(200, 200)	(300, 300)	(500, 500)
BS location	BS	(300, 100)	(100, 100)	(150, 0)	(500, 0)
Free-space model	ϵ_{fs}	10 pJ/bit/m ²	10 pJ/bit/m ²	10 pJ/bit/m ²	10 pJ/bit/m ²
Multipath model	ϵ_{mp}	0.0013 pJ/bit/m ⁴	0.0013 pJ/bit/m ⁴	0.0013 pJ/bit/m ⁴	0.0013 pJ/bit/m ⁴
Initial battery level	E_o	1 J	0.5 J	0.5 J	0.5 J
Size of packet	M	4000 bits	4000 bits	4000 bits	4000 bits
Data aggregation	E_{DA}	5 nJ/bit/report	5 nJ/bit/report	5 nJ/bit/report	5 nJ/bit/report
Electronic circuitry	E_{elec}	50 nJ/bit	50 nJ/bit	50 nJ/bit	50 nJ/bit

Scenario-2, it is 79.96%, 42.92%, 24.28% and 15.10%. Likewise, in scenario-3, E-FUCA has shown improvement of 158.57%, 123.46%, 57.39% and 39.23% over LEACH, FUCA, URBD and DEFL protocols, respectively. Significant enhancement in QND can be seen for scenario-4, where E-FUCA boosted QND by 212.73%, 177.42%, 82.01% and 77.32% over LEACH, FUCA, URBD and DEFL, respectively.

Figure 15 depicts the performance of E-FUCA, LEACH, FUCA, URBD and DEFL protocols in terms of HND. We have contemplated HND only because once half of the nodes are dead; then the complete coverage cannot be guaranteed in most of the cases. In scenario-1, the HND of the proposed E-FUCA protocol is enhanced by 31.81%, 26.03%, 16.95% and 13.11% over LEACH, FUCA, URBD and DEFL protocols, and for Scenario-2, it is extended by 52.21%, 43.20%,

Fig. 13 Stability period

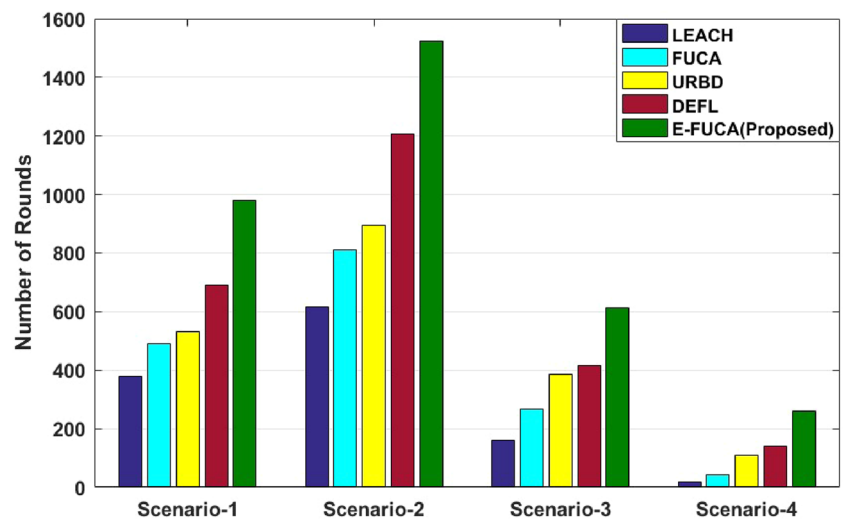


Fig. 14 QND for four scenarios

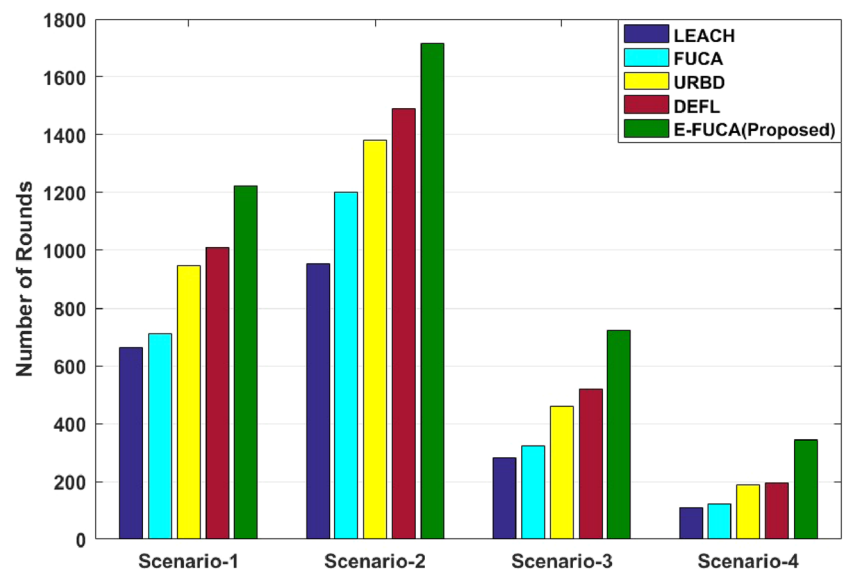
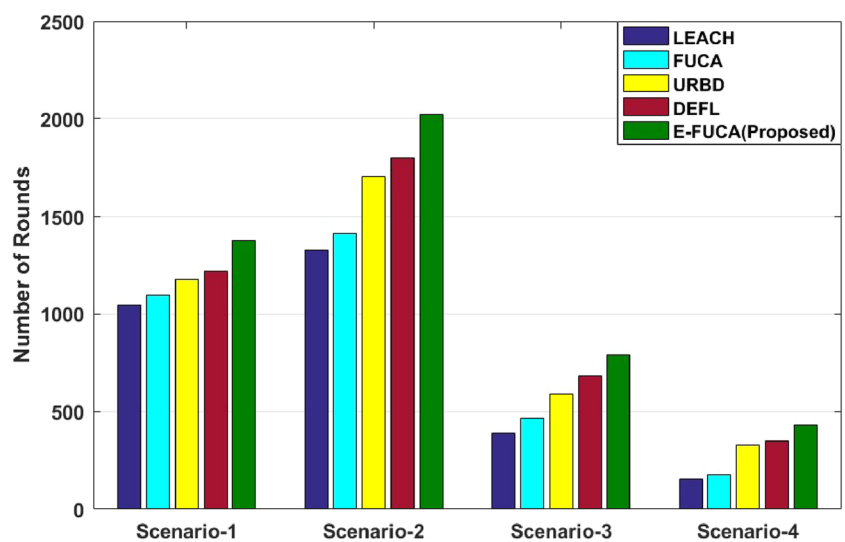


Fig. 15 HND for four scenarios



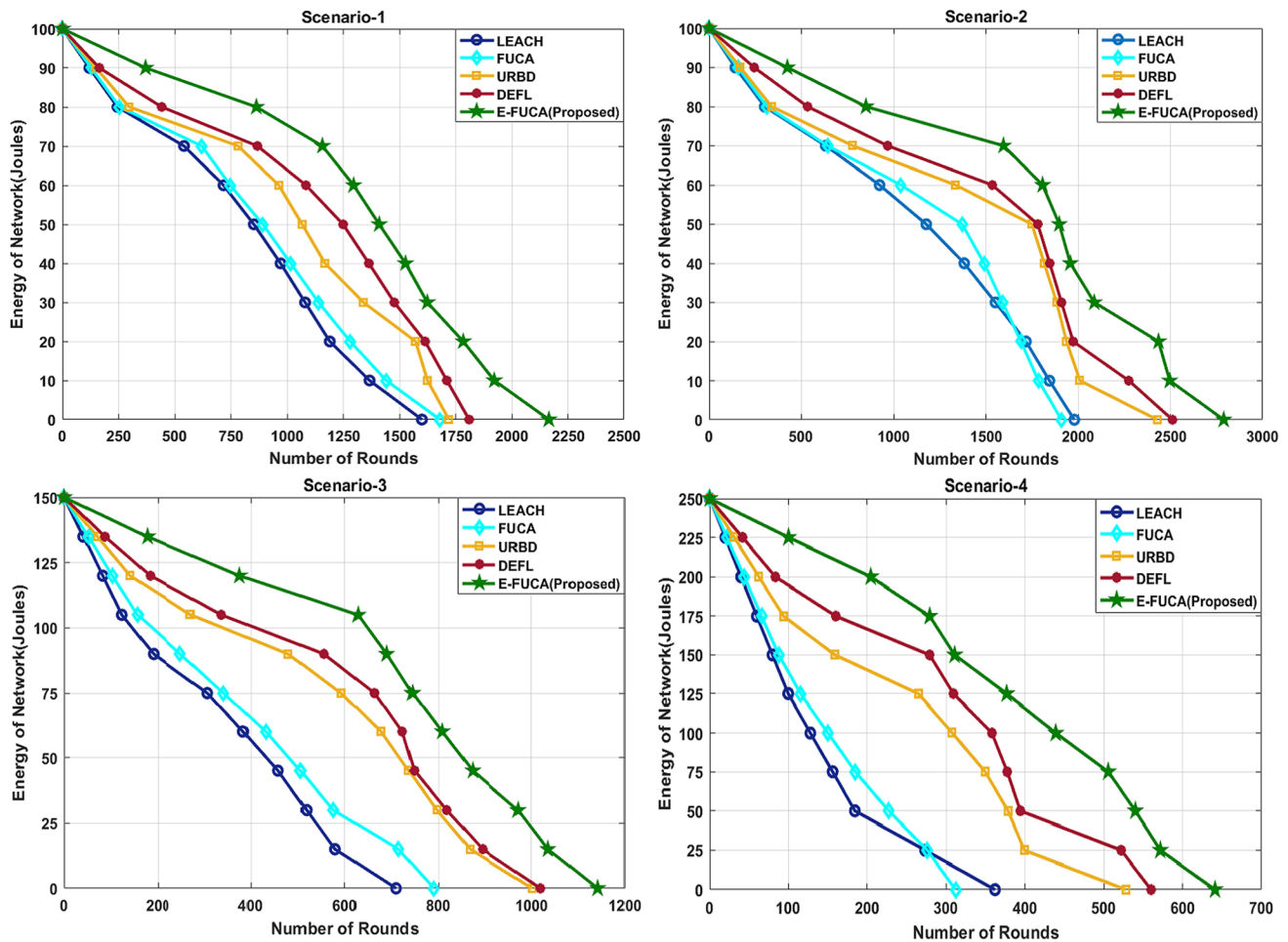


Fig. 16 The total energy of the network for four scenarios

18.66% and 12.27% more than LEACH, FUCA, URBD and DEFL protocols, respectively. In the case of scenario-3, E-FUCA increased HND by 102.56%, 68.09%, 33.90% and 16% over LEACH, FUCA, URBD and DEFL protocols. Similarly, for scenario-4, it is incremented by 176.92%, 144.07%, 35.42% and 22.73% over LEACH, FUCA, URBD and DEFL protocols, respectively.

Figure 16 presents the total average energy of the network for four scenarios. We can observe that the total average energy of the E-FUCA is dissipating at a very slow rate as compared to LEACH and FUCA. It can be clearly observed that the LEACH protocol poorly performed as compared to FUCA, URBD, DEFL and E-FUCA because it does not consider the crucial parameters during CH selection that affect the energy of the network. FUCA protocol has performed better than LEACH but poor in comparison to E-FUCA because it adapts a greedy approach in cluster formation as non-CH nodes choose the closest CH irrespective of considering its existing load. URBD has better performance

than FUCA and LEACH because it considers density and distance in cluster formation but has poor performance than E-FUCA because E-FUCA considers average distance instead of node density.

In Fig. 17, total alive nodes for different round slices are presented for the four scenarios considered. For scenario-1, we can observe that all nodes are alive in E-FUCA protocol approximately up to 950 rounds, whereas in LEACH and FUCA protocol, the count of the alive node is merely 55%, and for URBD and DEFL protocols, almost 30% of nodes are dead. It can be clearly observed in scenario-2 that E-FUCA performs extremely better than its comparatives. Up to 1500 rounds in E-FUCA protocol, all the nodes are alive, whereas, in the case of LEACH and FUCA, less than 50 per cent of the nodes are alive in the network. In the URBD and DEFL protocols, almost a quarter of nodes are dead in the network, which is poorer as compared to E-FUCA. In scenario-3, almost all the nodes are alive up to 800 rounds in E-FUCA protocol, whereas in case LEACH

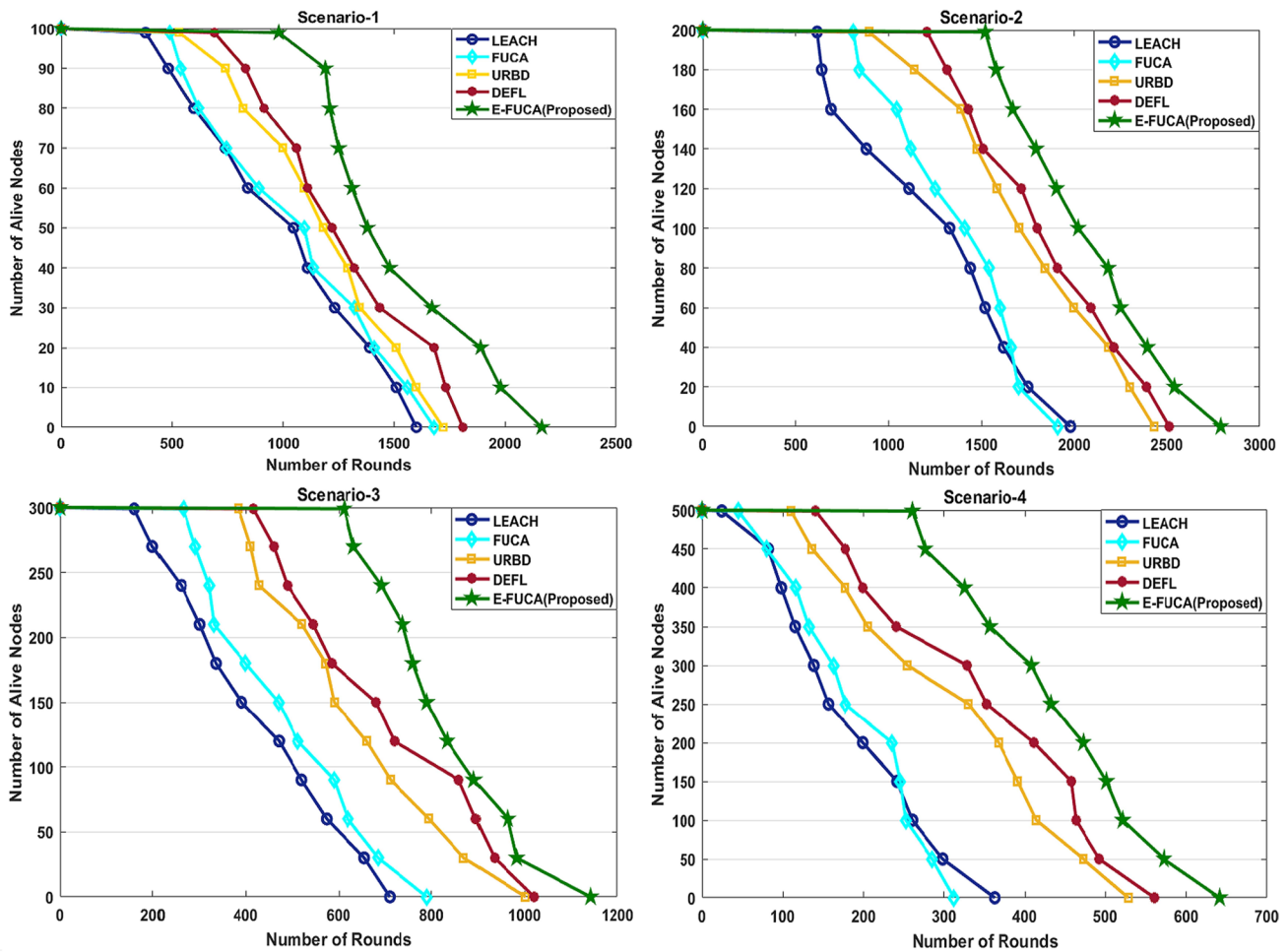


Fig. 17 Total alive nodes for four scenarios

and FUCA, the network has expired. If we talk about the URBD protocol, less than a quarter of nodes are alive, but in the case of DEFL, one-third of nodes are alive in the network. In scenario-4, at 300 round, LEACH and FUCA lost more than three-fourth of deployed nodes, URBD lost more than three-fifth nodes, and DEFL lost a quarter of nodes, whereas the proposed E-FUCA protocol lost only one-tenth nodes. E-FUCA has shown better performance because it considers influential parameters during the CH election. In addition, at the time of cluster formation, non-CH nodes make an intelligent decision of choosing their CH by determining its existing load.

Complexity analysis of E-FUCA

Time complexity

There are total n nodes deployed in the network. For the CH selection, each node will compute its rank and competition radius independently. In the worst case, an SN will

make $(n - 1)$ number of comparisons of rank for getting itself elected as CH, as shown in Algorithm 1. Therefore, for n nodes, a total $n(n - 1)$ number of comparisons occur for CH selection. For the formation of the cluster, every non-CH node will calculate the chance of each node in the CH_NODE list. Thus, in the worst case, there will be $(n - 1)$ comparisons. If there are k CHs, then in the case of routing, there will be k comparisons. Therefore, the complexity of the E-FUCA Protocol in terms of BIG-OH will be $O(n^2)$.

Message complexity

At the commencement of each round, all the SN generate an RN and if that $RN < T_{prob}$, then that SN broadcasts a message (CH_MSG). Let the number of CH be k for each round. Therefore, the total CH_MSG messages will be k . The non-CH nodes will transmit a message (JOIN_REQ) to CH, which will be $(n - k)$. TDMA schedule will be broadcast to cluster members who will be equal to k . Thus, the total number of messages exchanged for a selection of CH and the

formation of clusters in a round will be $k + (n - k) + k = n + k$. In the case of routing, the total messages forwarded will be k . Thus, the message complexity of the proposed protocol will be $O(n)$.

Conclusion

While designing WSN, the proliferation of energy efficiency is a key concern. Distributing the load among all nodes at par may result in a better stability period. E-FUCA is designed to enhance the performance of FUCA protocol by considering remnant energy, closeness to BS and average distance to nearby nodes instead of node density during CH election. In addition, in the E-FUCA protocol, non-CH nodes intelligently determine the prevailing load of CH before making a decision of selecting its CH. Energy-efficient Fuzzy-based next-hop selection is proposed for protracting network lifetime. The experimental evaluation of the propound work is carried out for four different cases wherein the BS position is kept at various places in the area of interest, meeting the requirement of all kinds of applications. The simulation results proclaim remarkable performance of E-FUCA over LEACH, FUCA, URBD and DEFL in all four scenarios in context to stability period, QND, HND, total average energy and total alive nodes.

Declarations

Conflict of interest No conflict of interest exists.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Rawat P, Singh KD, Chaouchi H, Bonnin JM (2014) Wireless sensor networks: a survey on recent developments and potential synergies. *J Supercomput* 68:1–48. <https://doi.org/10.1007/s11227-013-1021-9>
2. Pal R, Yadav S, Karnwal R, Aarti Y (2020) EEWC: energy-efficient weighted clustering method based on genetic algorithm for HWSNs. *Complex Intell Syst* 6:391–400. <https://doi.org/10.1007/s40747-020-00137-4>
3. Yick J, Mukherjee B, Ghosal D (2008) Wireless sensor network survey. *Comput Netw* 52:2292–2330. <https://doi.org/10.1016/J.COMNET.2008.04.002>
4. Bhushan S, Kumar M, Kumar P, Stephan T, Shankar A, Liu P (2021) FAJIT: a fuzzy-based data aggregation technique for energy efficiency in wireless sensor network. *Complex Intell Syst*. <https://doi.org/10.1007/s40747-020-00258-w>
5. Mehra PS, Doja MN, Alam B (2015) Low energy adaptive stable energy efficient (LEASE) protocol for wireless sensor network. *Int Conf Futuristic Trends Comput Anal Knowl Manag*. <https://doi.org/10.1109/ABLAZE.2015.7155044>
6. Afsar MM, Tayarani-N MH (2014) Clustering in sensor networks: a literature survey. *J Netw Comput Appl* 46:198–226. <https://doi.org/10.1016/j.jnca.2014.09.005>
7. Deebak BD, Al-Turjman F (2021) Secure-user sign-in authentication for IoT-based eHealth systems. *Complex Intell Syst* 1:3. <https://doi.org/10.1007/s40747-020-00231-7>
8. Al-Kiyumi RM, Foh CH, Vural S, Chatzimisios P, Tafazolli R (2018) Fuzzy logic-based routing algorithm for lifetime enhancement in heterogeneous wireless sensor networks. *IEEE Trans Green Commun Netw* 2:517–532. <https://doi.org/10.1109/TGCN.2018.2799868>
9. Kulkarni RV, Förster A, Venayagamoorthy GK (2011) Computational intelligence in wireless sensor networks: a survey. *IEEE Commun Surv Tutor* 13:68–96. <https://doi.org/10.1109/SURV.2011.040310.00002>
10. Agrawal D, Pandey S (2018) FUCA: Fuzzy-based unequal clustering algorithm to prolong the lifetime of wireless sensor networks. *Int J Commun Syst* 31:34–48. <https://doi.org/10.1002/dac.3448>
11. Heinzelman WR, Chandrakasan A, Balakrishnan H (2000) Energy-efficient communication protocol for wireless microsensor networks. In: *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*. IEEE Computer Society, p 10
12. Hamidzadeh J, Ghomanjani MH (2018) An unequal cluster-radius approach based on node density in clustering for wireless sensor networks. *Wirel Pers Commun* 101:1619–1637. <https://doi.org/10.1007/s11277-018-5779-1>
13. Kim JH, Chauhdary SH, Yang WC, Kim DS, Park MS (2008) PRODUCE: a probability-driven unequal clustering mechanism for wireless sensor networks. In: *22nd International Conference on Advanced Information Networking and Applications-Workshops (aina workshops 2008)*. IEEE, pp 928–933
14. Yu J, Qi Y, Wang G (2011) An energy-driven unequal clustering protocol for heterogeneous wireless sensor networks. *J Control Theory Appl* 9:133–139. <https://doi.org/10.1007/s11768-011-0232-y>
15. Lee S, Choe H, Park B, Song Y, Kim C (2011) LUCA: an energy-efficient unequal clustering algorithm using location information for wireless sensor networks. *Wirel Pers Commun* 56:715–731. <https://doi.org/10.1007/s11277-009-9842-9>
16. Yu J, Qi Y, Wang G, Guo Q, Gu X (2011) An energy-aware distributed unequal clustering protocol for wireless sensor networks. *Int J Distrib Sens Netw* 7:202145. <https://doi.org/10.1155/2011/202145>
17. Kim J, Park S, Han Y, Chung T (2008) CHEF: cluster head election mechanism using fuzzy logic in wireless sensor networks. In: *Proceedings of 10th International Conference on Advanced Communication Technology*, pp 654–659
18. Gajjar S, Sarkar M, Dasgupta K (2014) Cluster head selection protocol using fuzzy logic for wireless sensor networks. *Int J Comput Appl* 97:38–43. <https://doi.org/10.5120/17022-7310>
19. Mehra PS, Doja MN, Alam B (2020) Fuzzy based enhanced cluster head selection (FBECS) for WSN. *J King Saud Univ Sci* 32:390–401. <https://doi.org/10.1016/J.JKSUS.2018.04.031>

20. Bagci H, Yazici A (2013) An energy aware fuzzy approach to unequal clustering in wireless sensor networks. *Appl Soft Comput* 13:1741–1749. <https://doi.org/10.1016/J.ASOC.2012.12.029>
21. Logambigai R, Kannan A (2016) Fuzzy logic based unequal clustering for wireless sensor networks. *Wirel Netw* 22:945–957. <https://doi.org/10.1007/s11276-015-1013-1>
22. Mao S, Zhao C, Zhou Z, Ye Y (2013) An improved fuzzy unequal clustering algorithm for wireless sensor network. *Mob Netw Appl* 18:206–214. <https://doi.org/10.1007/s11036-012-0356-4>
23. Baranidharan B, Santhi B (2016) DUCF: Distributed load balancing unequal clustering in wireless sensor networks using fuzzy approach. *Appl Soft Comput* 40:495–506. <https://doi.org/10.1016/J.ASOC.2015.11.044>
24. Sert SA, Bagci H, Yazici A (2015) MOFCA: multi-objective fuzzy clustering algorithm for wireless sensor networks. *Appl Soft Comput* 30:151–165. <https://doi.org/10.1016/J.ASOC.2014.11.063>
25. Mirzaie M, Mazinani SM (2017) MCFL: an energy efficient multi-clustering algorithm using fuzzy logic in wireless sensor network. *Wirel Networks*. <https://doi.org/10.1007/s11276-017-1466-5>
26. Tian Y, Zhou Q, Zhang F, Li J (2017) Multi-hop clustering routing algorithm based on fuzzy inference and multi-path tree. *Int J Distrib Sens Netw* 13:155014771770789. <https://doi.org/10.1177/1550147717707897>
27. AlShawi IS, Yan L, Pan W, Luo B (2012) Lifetime enhancement in wireless sensor networks using fuzzy approach and A-star algorithm. In: *IET Conference Publications*
28. Khudair Leabi S, Younis Abdalla T (2015) Energy efficient routing protocol for maximizing lifetime in wireless sensor networks using fuzzy logic. *Int J Adv Comput Sci Appl* 7(10):95–101. <https://doi.org/10.14569/IJACSA.2016.071012>
29. Jiang H, Sun Y, Sun R, Xu H (2013) Fuzzy-logic-based energy optimized routing for wireless sensor networks. *Int J Distrib Sens Netw* 9:216561. <https://doi.org/10.1155/2013/216561>
30. Ortiz AM, Royo F, Olivares T, Castillo JC, Orozco-Barbosa L, Marron PJ (2013) Fuzzy-logic based routing for dense wireless sensor networks. *Telecommunication systems*. Springer, Berlin, pp 2687–2697
31. Haider T, Yusuf M (2009) A fuzzy approach to energy optimized routing for wireless sensor networks. *Int J Disturb Sens J* 9(8):1–8. <https://doi.org/10.1155/2013/216561>
32. Dwivedi A, Sharma A, Mehra PS (2020) Energy-aware routing protocols for wireless sensor network based on fuzzy logic: a 10-years analytical review. *EAI Endorsed Trans Energy Web*. <https://doi.org/10.4108/eai.6-10-2020.166548>
33. Mamdani HE (1977) Application of fuzzy logic to approximate reasoning using linguistic synthesis. *IEEE Trans Comput C* 26:1182–1191. <https://doi.org/10.1109/TC.1977.1674779>
34. Al-Quh MAH, Saroit IA, Mohammed Kotb DA (2016) FEQRP: a fuzzy based energy-efficient and QoS routing protocol over WSNs. *IOSR J Comput Eng* 18:79–89. <https://doi.org/10.9790/0661-1804057989>
35. Mehra PS, Doja MN, Alam B (2016) Enhanced stable period for two level and multilevel heterogeneous model for distant base station in wireless sensor network. *Advances in intelligent systems and computing*. Springer, Berlin, pp 751–759. https://doi.org/10.1007/978-81-322-2517-1_72
36. Smaragdakis G, Matta I, Bestavros A (2004) SEP: a stable election protocol for clustered heterogeneous wireless sensor networks. *Second Int Work Sens Actor Netw Protoc Appl (SANPA)*. <https://doi.org/10.3923/jmcomm.2010.38.42>

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



Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Energetic and exergetic study of dual slope solar distiller coupled with evacuated tube collector under force mode

Aseem Dubey, Samsher, Anil Kumar*

Department of Mechanical Engineering, Delhi Technological University (DTU), Delhi 110042, India

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Dual Slope Solar Distiller
Evacuated Tube Collector
Energy
Exergy
Water Depth

ABSTRACT

In this article, energetic and exergetic analysis of symmetric dual-slope solar still, oriented East-West at a latitude of 28°35' N and integrated with ETC under forced operation is carried out. The performance is evaluated at an optimal flow rate within the vacuum tubes with a viewpoint of maximum heat extraction. The energetic and exergetic efficiencies are increased by ~ 8.0% and ~ 6.0%, respectively, with an increase in flow rate from 0.01 to 0.24 kg/s. With increased water temperature, evaporative fraction exergy increases, ranging 0.2–0.9. The liner, water mass, glass cover, and collector have daily exergy efficiencies of ~ 8.12, ~30.1, ~41, ~18.0%, while overall exergy efficiency and global exergy efficiency are estimated as 4.9 and ~ 7.1%, respectively.

© 2021 Elsevier Ltd. All rights reserved.

Selection and peer-review under responsibility of the scientific committee of the Technology Innovation in Mechanical Engineering-2021.

1. Introduction

Energy and drinking water are the basic needs of human beings. The scarcity of pure water is more prevalent in the areas lying in the arid region's belt. Since the past, various conventional desalinating technologies are in use, but these technologies contribute to greenhouse gas emissions (GHGs) due to the energy generated from conventional fuels [1]. The conventional technologies can be replaced with solar energy-driven technologies, sustainable and eco-friendly, and eliminate the major running costs. Dwivedi and Tiwari [2] carried out an annual assessment of the single and dual slope solar stills and reported that the dual-slope solar still yields more than the single slope during peak summer; the single slope yields more during winter. The low productivity was the main issue with the conventional solar stills operating under passive mode. Thus, to increase the performance, various active methods have been used with various configurations.

The use of vacuum tubes in solar distillation leads to improved performance due to the various merits of flat plate collectors (FPC). Morrison et al. [3] concluded the better performance of the system using the ETC (evacuated tube collector) than the FPC (flat plate collector) for high-temperature operation. Various researchers studied the performance of ETC integrated single slope solar still

in thermosyphon mode and found improvement in the performance, which further depends on the number of tubes, water depth, and environmental conditions [4–8]. However, various dis-favorable conditions were reported with the ETC (evacuated tube collector) under natural circulation compared to the forced mode [9]. Patel et al. [10] studied stepped type solar distiller charged with ETC water heater and found yields ~ 24% higher during summer than uncoupled solar still.

The mass flow rate within the tubes is one of the factors, which influenced the thermal performance. Louise and Simon [11] recommended an optimum flow-rate ranging ~ 0.006–0.015 kg per second for the optimum performance of ETC. In the area of solar desalination, Kumar et al. [12] theoretically investigated an ETC integrated single slope solar distiller under the forced mode, having 0.03 m depth of water, and reported maximum daily yield and efficiency as 3.9 kgm⁻² and 33.8%, respectively, at an optimal flow rate of 0.006 kg/s/tube. Zhang et al. [13] investigated that thermal efficiency, outlet water temperature affected significantly by the flowrate through a flat plate collector using 10 riser pipe and recommended a water flowrate ranging 0.06–0.08 kg/s. Recently, Dubey et al. [14] theoretically reported 0.06 kg/s flowrate as optimal for the optimum performance of dual-slope (DS) solar stills coupled with ET (evacuated tube) collector under force mode.

Exergy is more of a qualitative value of energy and maximum work potential obtainable from the energy concerning the surrounding ambient conditions. As compared to the energy-based

* Corresponding author.

E-mail address: anilkumar76@dtu.ac.in (A. Kumar).

assessment, an exergy-based analysis accurately measures system's performance, which necessitates an exergetic assessment of the modified system, besides the energy approach. Singh et al. [15] performed the study on a PV/T integrated hybrid dual-slope solar distiller in New Delhi zone (latitude 28°35'N) and found annual energetic efficiency as 17.4%, while exergetic efficiency as 2.3%. Ranjan and Kaushik [16] reviewed the solar distillers using the energy and exergy approach. They noticed the exergy efficiency < 5.0% using single effect solar stills, which reached 8.5% for the integrated active solar distiller. The effect of inlet flow-rate on the cascaded solar distiller was studied by Zoori et al. [17]. They reported an increase in the exergy efficiency from 3.14 to 10.5% with decreasing flow ~ 0.003–0.001 l/s.

As per the literature, a dual-slope solar distiller in force mode with 'N' parallel vacuum tubes has not been investigated thoroughly from the energetic and exergetic viewpoint. The concept of exergy, accounting for the effect of water depth for the same climatic and operational parameters of the solar still oriented E-W at 15° and collector at the New Delhi, India (28°35' N latitude and 77°12'E longitude) gathered particular attention to analyze this modified geometry. Therefore, the main motive of the current research is to conduct performance evaluation of ETC augmented dual slope solar distiller under force mode employing energetic and exergetic approach with a view point of the maximum energy extraction from the vacuum tubes at an optimal flow with the variation of water depth.

2. Proposed system

Fig. 1 shows a simplified sketch of an ET (evacuated tube) collector coupled DSS (dual slope solar) still in force circulation. The major elements are glass cover, basin body, ETC, and a pump. The basin is even, black painted to absorb the maximum radiation. Fibre reinforced plastic (FRP) is selected as basin's material, and is oriented in the East-West direction for higher yield during summer.

On the top of the basin, two glass covers, to withstand self-weight, winds, rain, hail, temperature, and impact, are inclined at 15° with 0.04 m thickness. Sealing is used to block the leakage of vapour b/w the cover and basin body. The evacuated collector, composed of several parallel tubes, is oriented due south to receive most of the radiations. Water from the basin is circulated through ETC in a close loop with the help of a DC pump. A valve is also used to avoid reverse flow during night, and regulate the flow-rate. The

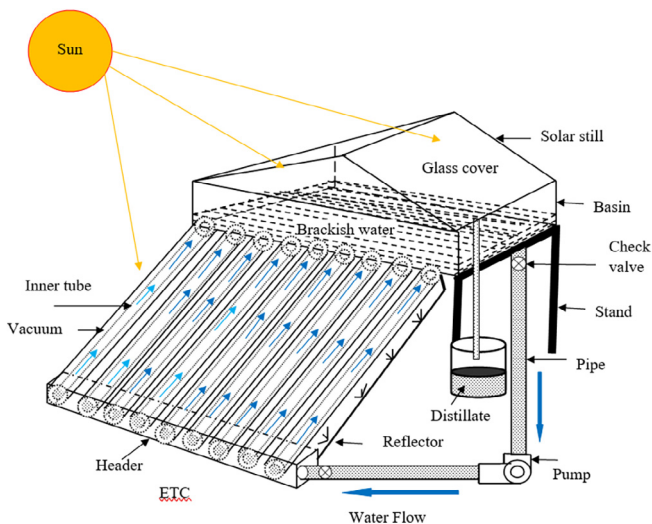


Fig. 1. Diagrammatic depiction of ETC integrated DSS distiller.

design specifications thus selected and the values of various coefficients for the dual-slope solar still, ETC, glass cover, and the pump are used as reported by various researchers [14].

3. Mathematical formulation

The different assumptions and thermophysical relations can be used to establish the different heat balance (HB) equations within the unified system [14].

If $I_c(t)_T$ is the total solar flux on the circumferential area of the vacuum tubes (direct along with intercepted radiation from the diffuse reflector), then the instantaneous heat extraction rate from the ETC is as under;

$$\dot{q}_{uc} = I_c(t)_T \eta_{ic} \quad (1)$$

$$\text{where, } I_c(t)_T = I_c(t)N_c[0.5A_t + (CD - d_t)\gamma\rho_r L_t]$$

3.1. Water temperature at the outlet of the collector and in the basin

The outlet water temperature from the collector (T_{cw}) can be estimated with the help of relations reported [12] and using the equations of basin and glass cover temperatures, and yield as reported [14].

3.2. Energetic and exergetic analysis

The overall instant energy efficiency can be evaluated as;

$$\eta_{i,overall} = \frac{m_{ew,T} L}{(I_c(t)A_{cT} + I_E(t)A_{gE} + I_W(t)A_{gW} + 2.7W_p)3600} \times 100 \quad (2)$$

And daily overall energy efficiency can be estimated using the cumulative value of daily energy output and solar energy on the overall system area.

Exergy, for a thermal system, is a part of energy (\dot{q}) and is evaluated between the source (T_{sw}) and sink (T_a) temperatures as;

$$\dot{Ex} = \dot{q} \left(1 - \frac{T_a}{T_{sw}} \right) \quad (3)$$

$$\text{where } \dot{q} = h_{ew} A_b (T_{sw} - T_g)$$

With an increase in the ratio $\frac{T_a}{T_{sw}}$, the exergy transfer decreases.

The following expression to estimate the exergy of the sun's radiation [$I_s(t)$] is used [16];

$$\dot{Ex}_{Sun} = \left(1 - \frac{4}{3} \times \left(\frac{T_a}{T_{sun}} \right) + \frac{1}{3} \times \left(\frac{T_a}{T_{sun}} \right)^4 \right) I_{sun}(t) \approx 0.933 I_{sun} \quad (4)$$

The instantaneous exergetic efficiency can be evaluated under the general rule given as;

$$\eta_{i,EX} = \frac{\text{Exergy out from the component}}{\text{Exergy Input to component}} = 1 - \frac{\text{Ex}_{destruction} \text{ in the component}}{\text{Exergy input to component}} \quad (5)$$

And for the present design of solar still, the instant overall ($\eta_{i,EX,overall}$) and global exergy (solar still alone) efficiencies can be written as;

$$\eta_{i,EX,overall} = \frac{h_{ew}(T_w - T_g)A_b \times \left[1 - \left(\frac{T_a}{T_w} \right) \right]}{[0.933I_c(t)A_{cT} + 0.933[I_E(t)A_{gE} + I_W(t)A_{gW}] + W_p} \times 100 \quad (6a)$$

$$\eta_{i,EX,global} = \frac{h_{ew}(T_w - T_g)A_b \times \left[1 - \left(\frac{T_a}{T_w} \right) \right]}{[\dot{Ex}_{c,ETC} + 0.933[I_E(t)A_{gE} + I_W(t)A_{gW}] + W_p} \times 100 \quad (6b)$$

The daily overall exergy efficiency can be estimated using the cumulative value of daily exergy output and exergy of solar radiation on the overall system area, accounting for the total exergy input to the solar still alone.

Following the exergy matrix reported by the various researchers [16], the analytical expressions for the various components can be estimated as;

3.2.1. Basin water

Within the basin water, the exergy balance is given hereunder;
(a) input solar exergy for the water mass

$$\alpha_w \dot{E}x_{sun} + \dot{E}x_{bw} + \dot{E}x_{c,etc} \quad (7a)$$

(b) total upward exergy transfer from the water surface ($\dot{E}x_{1wg}$) is written as [18–19];

$$\dot{E}x_{1wg} = \dot{E}x_{ewg} + \dot{E}x_{cwg} + \dot{E}x_{rwg} \quad (7b)$$

(c) day time exergy accumulated in the basin water may be written as;

$$\dot{E}x_{acm} = M_{sw} C_w \left[(T_{sw} - T_{sw,i}) - T_a \ln \frac{T_{sw}}{T_{sw,i}} \right] \quad (7c)$$

The total diurnal exergy accumulated in the water is utilized for the nocturnal distillation.

3.2.2. Basin liner

The input exergy is the part of radiation exergy absorbed by the basin ($\alpha'_b \dot{E}x_{sun}$). Depending on the difference between the temperature ($T_b - T_{sw}$), some is transferred to the basin water ($\dot{E}x_{bw}$), and some to the environment ($\dot{E}x_{ba}$), while the remaining destroyed (\dot{I}_b). The exergy analysis at the basin liner is expressed by Eqn. (8).

(a) exergy absorbed by the liner = $\alpha'_b \dot{E}x_{sun}$ (8a)

(b) exergy transferred (lost) from liner to the ambient;

$$\dot{E}x_{ba} = h_{ba} (T_b - T_a) A_b \left(1 - \frac{T_a}{T_b} \right) \quad (8b)$$

(c) exergy transfer from liner to the basin water is only during the sunshine hours subject to the positive value of ($T_b - T_{sw}$) and can be expressed as;

$$\dot{E}x_{bw} = h_{bw} (T_b - T_{sw}) A_b \left(1 - \frac{T_a}{T_b} \right) \quad (8c)$$

where α'_b is the effective absorptivity of the basin liner.

3.2.3. Glass cover

The glass cover aids the condensation process by rejecting heat to the sink (surroundings). There is an insignificant difference in both east and west glass covers temperature. For this reason, the average of both the surfaces has been considered while carrying out the exergetic evaluation.

(a) exergy absorbed = $\alpha_g \dot{E}x_{sun} + \dot{E}x_{1wg}$ (9a)

Some absorbed exergy is lost in the ambient and the remaining destroyed due to irreversibility (I_g).

(b) external exergy transfer can be assessed as;

$$\dot{E}x_{1ga} = h_{1ga} (T_g - T_a) A_g \left(1 - \frac{T_a}{T_g} \right) \quad (9b)$$

$$h_{1ga} = 5.7 + 3.8V_a$$

where h_{1ga} is the overall heat loss coefficient from the top of the glass cover.

3.2.4. Collector

The exergy associated with ETC, following Jafarkazemi [17], is hereunder;

(a) The exergy input to ETC can be stated as;

$$\dot{E}x_{iETC} = 0.933 I_c(t) A_{cT} + \dot{E}x_{output \text{ from ETC}} \quad (10a)$$

(b) Exergy gain in the ETC tubes can be calculated as;

$$\dot{E}x_{c,ETC} = \dot{m} C_w [(T_{cw} - T_{cwi}) - T_a \ln \frac{T_{cw}}{T_{cwi}}] \quad (10b)$$

4. Results and discussion

The flow rate is optimized by estimating the maximum collector water temp. at the outlet for the system's yield and efficiency with incremental flow rates. For the numerical simulation, the initial water temperature, ambient temperature, condensing cover temperature, and radiation are considered in accordance with the experimental observations reported for a typical day [14].

Fig. 2 illustrates the outcome of numerical simulation for the flow rate combined with the number of tubes and water depth in the basin. Maintaining the water temperature at collector outlet $\sim 98.5^\circ\text{C}$, the effect of flowrate on the system productivity, energy and exergy efficiencies with the number of tubes combined with water depth are depicted to estimate optimal flow rate per tube irrespective of collector size. The yield, overall energy and exergy efficiencies are increased with the flowrate for each combination, reaches the maximum and start decreasing further, with insignificant change. Daily yield, energy efficiency and exergy efficiency obtained vary in the range of 6.18–6.44 kg, 31.4–33.75% and 4.37–4.84%, respectively, with the increase of mass flow rate from 0.01 kg/s to 0.06 kg/s using 10 tubes (i.e. flow rate per tube varies from 0.001–0.006 kg/s). Similar trends are also noticed with the combination of 20 and 30 tubes for a water depth 0.010 m and 0.125 m, respectively, for nearly same water temperature attainable ($\sim 98.5^\circ\text{C}$) at collector outlet for each combination and flow ranging 0.006 to 0.007 kg/s/tube is found optimum, irrespective of the number of tubes. It is found that with an increase in the tubes, the system yields higher but with a significant decrease in efficiencies. At optimal flow rate, maximum yields of 6.644 kg, 6.618 kg and 7.082 kg, while maximum energy efficiencies as $\sim 33.8\%$, 25.4%, 20.9% and exergy efficiencies as $\sim 4.9\%$, 3.6% and 2.99% are estimated at 0.006 kg/s/tube, and using 10, 20 and 30 ETC tubes, respectively. With the increase of tubes from 10 to 30 and at an optimum flow-rate, the yield enhances by $\sim 6.62\%$, while energy and exergy efficiencies reduce by $\sim 38.6\%$. The optimal flow rate for the present configuration is validated and found in the range reported by various researchers. The respective yield, energetic and exergetic efficiencies are increased by 8.0%, 8.0%, and 6.0%, with an enhancement in flow ranging 0.01–0.24 kg/s using 10 vacuum tubes, and found optimum at 0.006 kg/s/tube (i.e. 0.06 kg/s) flowrate.

Fig. 3 shows the change in instant exergy transfer associated with different elements of the system, derived at the optimum flow-rate with 10 tubes and 0.005 m water depth. Exergy transfer from the liner to water is observed contrary post 14:00 hrs, because of a comparatively lower liner temperature than the basin water, due to the supply of preheated water from the collector. The exergy transfer from the convective mode is noted lowest. The evaporative exergy is higher than the other modes of internal exergy transfer as desired. The highest value of exergy transfer through the evaporation mode is observed at 13:00 hrs due to notable differences in the glass cover and water temperatures

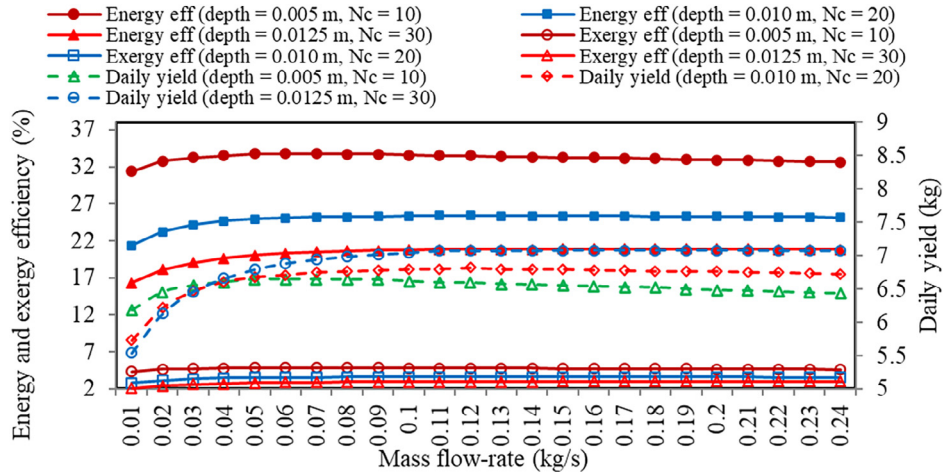


Fig. 2. Effect of flow-rate on the performance.

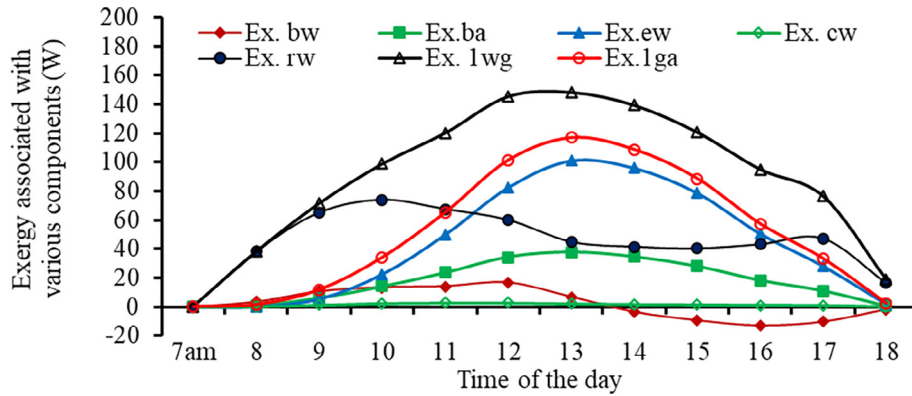


Fig. 3. Variation of instant exergy transfer between various components.

and found ~ 100 W, followed by radiative (40 W) and convective modes.

The Fractional exergy transfer by 03 modes is estimated using the respective fraction of $\dot{E}x_{ewg}$, $\dot{E}x_{cwg}$ and $\dot{E}x_{rwg}$ out of total exergy transfer ($\dot{E}x_{1wg}$), and variation with water temp. is depicted in Fig. 4. The evaporative exergy fraction affects the yield, whereas convective and radiation fractions have a negligible impact on productivity. With an enhancement in water temperature, evaporative fraction exergy increases in the range of 0.2–0.9, whereas

convective and radiative fraction decreases significantly. When the basin water temperature is at a peak, the evaporative exergy fraction exhibits a higher value, whereas convective and radiative fractions exhibit the lowest value.

Fig. 5 shows the overall and global (solar still alone) instant exergetic efficiencies of the system evaluated and are found ranging 0.0–8.4% and 0.0–12.0%, respectively. Thus, it is revealed from the outputs that T_{sw} has a noticeable effect on the increase in evaporative exergy due to a higher difference in temperature b/w the water (T_{sw}) and condensing cover (T_{gi}), i.e. ($T_{sw}-T_{gi}$).

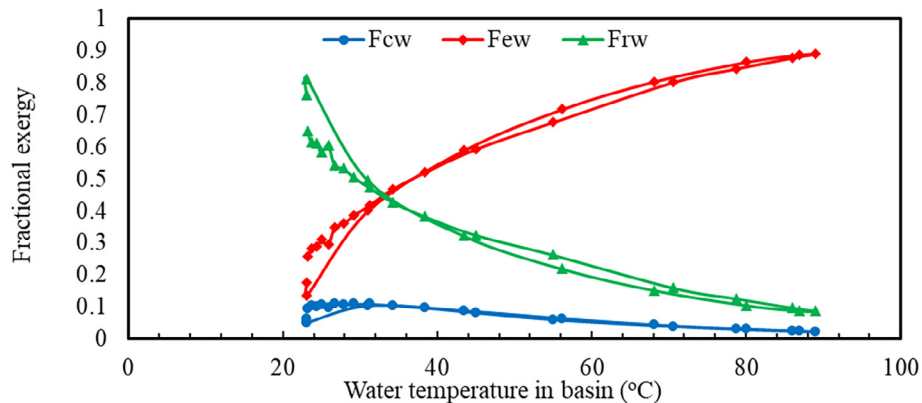


Fig. 4. Variation of fractional exergy transfer in different modes.

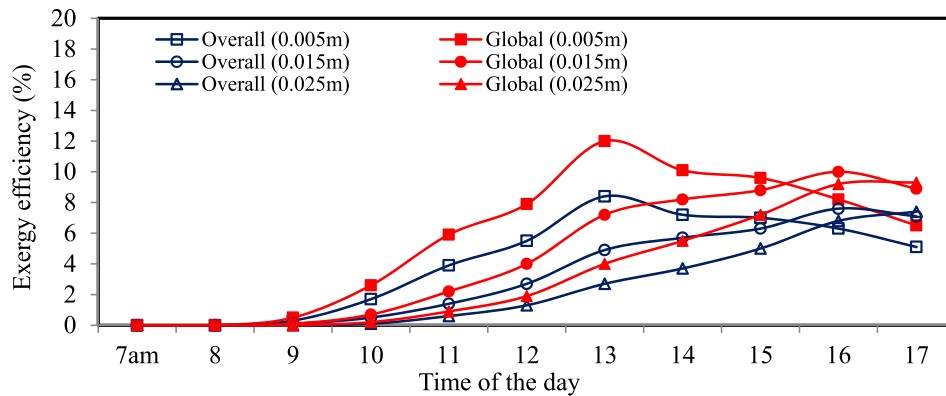


Fig. 5. Effect of water's depth on instant exergetic efficiency using 10 tubes.

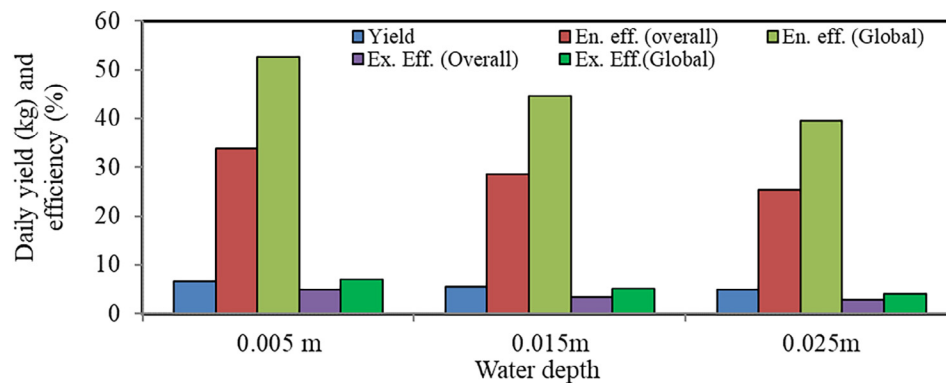


Fig. 6. Effect of water depth on the daily performance of the solar still with 10 tubes.

Fig. 6 depicts the variation of the energy and exergy efficiencies with water depth for the same climatic and operational conditions. Owing to the higher irreversibility of energy quality, the exergy efficiency is lower than the energy efficiency. The daily energetic and exergetic efficiencies are observed to be reduced as depth is increased. The system's daily En. and Ex. Efficiencies are reduced by 33.8–25.2% and 4.9–2.7%, respectively, with an increase in water depth (0.005–0.025 m) owing to a higher thermal inertia impact. The liner, water mass, glass cover, and collector have daily exergy efficiencies of ~8.12, ~30.1, ~41, and ~18%, respectively, while the overall exergy efficiency (complete system) and global exergy efficiency (solar still alone) are estimated as 4.9 and ~7.1%, respectively, at 0.005 m depth of water and at optimal flow condition. The decrease in yield by ~22.0%, overall energy efficiency by ~25.0% and exergy efficiency by ~44.0% is noticed with increment in water level (0.005 m to 0.025 m).

5. Conclusions

Based on the exergetic performance evaluation, the conclusions drawn for the modified geometry of solar distiller are as under;

- I. The system performance is optimal at 0.006 kg/s/tube flow-rate for each combination. The overall energy and exergy efficiencies are estimated as ~33.8 and ~4.9%, respectively, which comparatively decreases with increased tubes.
- II. For the liner, water mass, glass cover, and collector, daily exergy efficiencies are obtained as ~8.12, ~30.1, ~41.0, and ~18.0%, while overall exergy efficiency (complete system) and global exergy efficiency (solar still alone) are estimated as 4.9 and ~7.1%, respectively, at 0.005 m water depth.

- III. The decrease in yield by ~22.0%, overall energy efficiency by ~25.0% and exergy efficiency by ~44.0% is noticed with increased water depth (0.005 m to 0.025 m).
- IV. The fractional evaporative exergy is found in the range of 0.2–0.9, increasing temperature and dominating over other modes above ~35 °C water temperature.

CRediT authorship contribution statement

Aseem Dubey: Methodology, Visualization, Investigation, Writing - original draft. **Samsher:** Visualization. **Anil Kumar:** . . .

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 would like to thank Delhi Technological University, Delhi, for awarding the fellowship for pursuing the Ph.D. in the Mechanical Engineering Department and gratefully acknowledging the Centre for Energy and Environment to support the present work.

References

- [1] K.S. Reddy, H. Sharon, Energy–environment–economic investigations on evacuated active multiple stage series flow solar distillation unit for potable water production, *Energy Convers. Manage.* 151 (2017) 259–285.
- [2] V.K. Dwivedi, G.N. Tiwari, Energy and exergy analysis of single and double slope passive solar still, *Trends Appl. Sci. Res.* 3 (3) (2008) 225–241.

- [3] G.L. Morrison, I. Budihardjo, M. Behnia, Measurement and simulation of flow rate in a water-in-glass evacuated tube solar water heater, *Sol. Energy* 78 (2005) 257–267.
- [4] H.N. Panchal, M. Doshi, K. Thakor, A. Patel, Experimental investigation on coupling evacuated glass tube collector on single slope single basin solar still productivity, *J. Mech. Eng.* 2 (11) (2011) 1–9.
- [5] R.V. Singh, S. Kumar, M.M. Hasan, M.E. Khan, G.N. Tiwari, Performance of a solar still integrated with evacuated tube collector in natural mode, *Desalination* 318 (2013) 25–33.
- [6] K. Sampathkumar, T.V. Arjun, P. Senthikumar, The experimental investigation of a solar still coupled with an evacuated tube collector, *Energ. Source Part A* 35 (2013) 261–270.
- [7] H.N. Panchal, A. Awasthi, Theoretical modelling and experimental analysis of solar still integrated with evacuated tubes, *Heat Mass Transf.* 53 (2017) 1943–1955.
- [8] M. Yari, A.E. Mazareh, A.S. Mehr, A novel cogeneration system for sustainable water and power production by integration of a solar still and PV module, *Desalination* 398 (2016) 1–11.
- [9] A.I. Sato V.L. Scalon A. Padilha Numerical analysis of a modified evacuated tubes solar collector 2012 Santiago de Compostela, Spain 384–389
- [10] J. Patel, B.K. Markam, S. Maiti, Potable water by solar thermal distillation in solar salt works and performance enhancement by integrating with evacuated tubes, *Sol. Energy* 188 (2019) 561–572.
- [11] J.S. Louise, F. Simon, Theoretical flow investigations of an all glass evacuated tubular collector, *Sol. Energy* 81 (6) (2007) 822–828.
- [12] S. Kumar, A. Dubey, G.N. Tiwari, A solar still augmented with an evacuated tube collector in forced mode, *Desalination* 347 (2014) 15–24.
- [13] D. Zhang, J. Li, Z. Gao, L. Wang, J. Nan, Thermal performance investigation of modified flat plate solar collector with dual-function, *Appl. Therm. Eng.* 108 (2016) 1126–1135.
- [14] A. Dubey, S. Kumar, A. Arora, Enviro-energy-exergo-economic analysis of ETC augmented double slope solar still with 'N' parallel tubes under forced mode: Environmental and economic feasibility, *J. Clean. Prod.* 279 (2021) 123859.
- [15] G. Singh, S. Kumar, G.N. Tiwari, Design, fabrication and performance evaluation of a hybrid photovoltaic thermal (PVT) double slope active solar still, *Desalination* 277 (2011) 399–406.
- [16] K.R. Ranjan, S.C. Kaushik, Energy, exergy and thermo-economic analysis of solar distillation systems: A review, *Renew. Sustain. Energy Rev.* 27 (2013) 709–723.
- [17] H.A. Zoori, F.F. Tabrizi, F. Sarhaddi, F. Heshmatnezhad, Comparison between energy and exergy efficiencies in a weir type cascade solar still, *Desalination* 325 (2013) 113–121.

Further Reading

- [1] F. Jafarkazemi, E. Ahmadi, H. Abdi, Energy and exergy efficiency of heat pipe evacuated tube solar collectors, *Int. J. Therm. Sci.* 20 (1) (2016) 327–335.



International Journal of Research in Engineering and Innovation (IJREI)

journal home page: <http://www.ijrei.com>

ISSN (Online): 2456-6934



Enhancements in mechanical properties of dissimilar materials using friction stir welding (FSW) - A review

R.S Mishra, Shivani Jha

Department of Mechanical Engineering, Delhi Technological University Delhi, India

Abstract

The various light weighted material which reduces the overall weight of a component, increases the mechanical properties like tensile strength, hardness, fatigue strength etc. plays a important role in the emerging world of technology. New developments in friction stir welding contributes to the major acceptance in the research field. Among the various metals, the research is focused on the aluminum metal matrix composite which replaces the conventional welding of iron and steel. Moreover, the joining of dissimilar metals aluminum and other alloys is a need of the physical world. In the present paper the main emphasis is on the enhancement of mechanical properties in the FSW of the dissimilar aluminum joints.

©2020 ijrei.com. All rights reserved

Keywords: Mechanical Properties Enhancement, FSW, Dissimilar metals, Green Technology

1. Introduction

To meet the ever demanding requirement of industry in joining of various materials welding technology now a day comes into rescue to resolves the demand. As when compared to other joining techniques such as adhesive and mechanical fasteners welding again comes on top. The basic requirement of any joint is to achieve satisfactory physical, mechanical and tribological properties which are ideally superior to the base materials. Development regarding improvement in joint quality has been addressed worldwide by various researchers [1]. Although, various welding technique leads to formation of defects such as cracks, voids, and inter-metallic compounds in the joints. Thus, a better welding process needs to be employed so as to decrease or eliminate these persisting problems by employing joining technique which is either in a solid state or it is in a semi solid state. Friction stir welding (FSW) is one of such contemporary solid state joining process which makes a high strength joint by transforming the metal into a plastic state at a certain temperature necessarily below the melting point, and then under high forging pressure the mechanically stirs of two metals form a high-strength welded joint [2–5]. However, keeping in mind the recent industrial requirement of light weight metals or composites there are still many challenges that has to be addressed for joining of

such metals or composites.

1.1 Aluminum alloys

Many aluminum alloys are strong by virtue of precipitation hardening through natural or artificial ageing from the solution-treated condition. The heat associated with welding changes the microstructure of the material. The effect of welding is to cause a drop in hardness from HV_{max} towards HV_{min} as the peak temperature experienced increases. This is because precipitates will coarsen and reduce in number density in regions remote from the heat source, and will re-enter solution when the peak temperature is sufficiently high. Some re-precipitation may occur during the cooling part of the thermal cycle, resulting in a hardness value beyond HV. The ultimate result is the continuous line with a minimum in hardness somewhere in the heat-affected zone, due to the competing effects of dissolution and re-precipitation. But in contrast to age hardenable AA 6082, where a minimum hardness occurs in the HAZ, FSW of non-hardenable AA 5082 results in uniform hardness across the weld. This general scenario may be complicated by the effects of deformation in FSW as described for the specific example of AA 2219. AA 2219 is a copper precipitation-strengthened alloy containing about 6.3 wt% Cu, which because of its strength and

toughness at low temperatures, is used for containing liquified gases for rockets of various kinds. It is frequently supplied in the T87 condition, meaning that it has been solution treated, cold-worked (10% reduction in rolling) and artificially aged. It can be welded using arc processes but this results in a reduction in the cross-weld strength because the proof strength of the fusion zone decreases to about 140 MPa compared with the 370 MPa of the plate. The former can be increased to between 220–275 MPa using pulsed or pulsed electron beam welding techniques because this promotes finer grains in the fusion zone. Friction stir welding does not seem to have an advantage over arc welding with respect to the strength of the fusion zone. It was observed that the TMAZ is somewhat softer than the fusion zone because the latter dynamically recrystallizes into a fine grain structure. It is the coarsening of the Al_2Cu precipitates in the TMAZ that is partly responsible for its softening. Some transmission electron micrographs across the weld; these show clearly the huge changes due to the heat from the process. The formation of coarse precipitates at the grain boundaries, and their associated precipitate-free zones, are common detrimental features in the microstructure. Aluminum being one of the light weight material which is most commonly used in industry needs severe attention as welding of dissimilar aluminum alloys is difficult. Fusion processes of such material can result in significant loss of strength in the joint due to the intense heat generation because of thermally activated softening mechanisms. Dissimilarity in welding can be viewed in various aspects and can be categorized accordingly such as similar base metals but different thicknesses or shape, welding of similar metals with different alloy compositions, welding of dissimilar metals with some compatibility on the phase diagram as in the case of aluminium and copper, welding of incompatible dissimilar metals, such as between magnesium and steel, welding of metal and ceramic which are considered to form metallic bonds between each other. Keeping in mind the need of energy efficiency, environmental friendliness, and versatile technology for joining of dissimilar metals FSW is considered to be the most significant development in recent decades. This process offers a number of advantages over conventional joining processes. The few of them includes (a) absence of expensive consumables such as a cover gas or flux; (b) ease of automation of the machinery involved; (c) low distortion of the work-piece; and (d) good mechanical properties of the resultant joint [12]. The fact that FSW welds in precipitation hardened alloys lead to a weak zone is not surprising given that the majority of strengthening in most strong alloys comes from precipitates. There is some evidence that manipulation of pin profiles and FSW parameters may help improve slightly, the hardness in the central region or indeed, in the HAZ. Theoretical work has also been done to see if cryogenic cooling after the tool pass can help retain alloying elements in solution after the peak temperature is experienced, so that the alloy can then naturally age and not develop the coarsened microstructures typical of slow cooling from the peak temperature. However, computations indicate that the advantage in so doing is likely to be minimal. It was observed that the grain size increases with increase in peak temperature caused by increase in rotational speed. Here grain size is related to peak temperature by assuming static grain-growth of dynamically recrystallized grains, during the cooling of

the thermal cycle. The precipitate free zones form near the grain boundaries because grain boundaries act as sinks for nearby dislocations, reducing nucleation sites for precipitates and also as precipitation sites, effectively reducing the solute content around them. As grain size increase, assuming constant width of PFZs, their volume fraction decreases with increase peak temperature. Not all alloys of aluminum are precipitation hardened. In the 2000 series alloys, the strength depends more on grain size (d), which has been expressed in terms of the Zener-Holloman parameter

$$\log d = a + b \log Z$$

Where a and b are empirical constants based on data from extrusion experiments and the hardness is then related to d using a form typical of the Hall–Petch type equation:

$\text{HV} = \text{HV}_0 + c/\sqrt{d}$ where c is a constant. In the cast Al–Si alloys, friction stir welding breaks up the large silicon particles in the nugget and the TMAZ, Fig. 26; as a consequence, the fracture is located in the base plate during cross-weld tensile tests because in this case, it is the coarse silicon particles which control failure. FSW can also heal casting defects such as porosity. Corrosion studies indicate that the weld zones produced by friction stir welding have comparable environmentally assisted cracking susceptibility as the unaffected parent.

1.2 Magnesium alloys

Magnesium alloys, normally produced by casting, may find significant applications in the automotive and aerospace industries with rapid growth particularly in die-cast vehicle components because of their better mass-equivalent properties. They are used for light-weight parts which operate at high speeds. The motivation for using FSW for magnesium alloys is that arc welding results in large volumes of non-toxic fumes. On the other hand, solid-state FSW does not result in solute loss by evaporation or segregation during solidification, resulting in homogeneous distribution of solutes in the weld.

Also, many magnesium alloys in the cast condition contain porosity which can be healed during FSW. The hardness and strength can be retained after friction stir welding. There is no significant precipitation hardening in the alloy studied (AZ31, $\approx \text{Mg-3Al-1Zn}$ wt% wrought) and the net variation in hardness over the entire joint was within the range 45–65 HV, with the lower value corresponding to the base plate. In the same system, a higher starting hardness of 70 HV leads to a substantially lower hardness in the nugget (50–60 HV); the variations in hardness appear to be consistent with measured variations in grain size in accordance with the form of the Hall–Petch relationship. The grains in both the nugget and TMAZ tend to be in a recrystallized form, and tend to be finer when the net heat input is smaller (for example at higher welding speeds). In Mg–Zr alloys with Zr-containing particles, FSW leads to a considerable refinement of the grain structure and sound welds can be produced in thin sheets over a wide range of welding conditions for sheets thicker than about 3 mm, the welds contained defects associated with an inability to supply sufficient heat during welding. There are, however, contradictory results showing

successful welds in 6 mm thick Mg–Zn–Y–Zr plates so it is unlikely that these results are generic to magnesium alloys.

1.3 Copper alloys

Copper which has much higher thermal diffusivity than steel cannot easily be welded by conventional fusion welding techniques. Heat input required for copper is much higher than conventional. FSW because of the greater dissipation of heat through the work-piece. Recently, FSW has been successfully used to weld 50 mm thick copper canisters for containment of nuclear waste. FSW in copper alloys have all the typical zones found in other materials: the nugget, TMAZ, HAZ and base structure. The nugget has equiaxed recrystallised small grains and its hardness may be higher or lower than the base material depending on the grain-size of the base metal. When 4 mm thick copper plates with average grain size of 210 μm were welded at high rpm (1250) and low welding speed (1.01 mm/s), nugget had lower hardness (60–90 HV), compared to base metal (105–110 HV). Even though grain size decreased from 210 to 100 μm , hardness decreased slightly due to reduction in dislocation density relative to base metal. Similar decrease in dislocation density in the nugget zone compared to parent metal has been observed for AA 7075 and AA 6061.

On the other hand, when 2 mm thick copper plates with average grain size of 30 μm were welded at 1000 rpm and 0.5 mm/s low welding speed, nugget (128–136 HV) was harder than the base metal (106–111 HV) due to reduction in average grain size to 11 μm . Flores et al. have also shown that as-cast AA 7073 showed that weld nugget was harder than base metal while the 50 % cold-rolled alloy showed reduced hardness in the nugget.

1.4 Titanium alloys

By far the most dominant of titanium alloys is Ti–6Al–4V, which in its commercial condition has a mixed microstructure consisting of hexagonal-close packed α and body-centred cubic β phases, which is the stable phase at high temperatures. This alloy, which accounts for about half of all the titanium that is produced, is popular because of its strength (1100 MPa), creep resistance at 300°C, fatigue resistance and cutability. Friction stir welding must clearly disrupt the base microstructure both through the thermal and deformation components of the process, but the consequences of this on performance during fabrication and service need investigation. General investigations on fatigue performance indicate that the crack growth rate in the HAZ can be higher or lower than the base material depending on specimen geometry, microstructure and residual stress levels. Several Experiments have also been conducted by several investigators, on a fully β -titanium alloy in thin sheet form, primarily to prove that the crystallographic texture observed corresponds to one generated by shear deformation, consistent with similar observations in aluminium alloys. Pure titanium in its hexagonal close-packed α -form is interesting because there is also a tendency for deformation by mechanical twinning during friction stir welding. The nugget region of an FSW joint is found to contain a large density of dislocations and mechanical twins, with transmission microscopy showing an elongated fine-structure,

but the overall grain shape seem to remain equiaxed on the scale of optical microscopy. It is speculated that recrystallisation must have occurred during welding but was followed by a small amount of plastic deformation. The HAZ simply revealed grain growth, a consequential lower hardness, and hence was the location of fracture in cross-weld tests. There was no clearly defined TMAZ as is typical in aluminium FS-welds.

1.5 Steels

The friction-stir welding of steels has not progressed as rapidly as for aluminium for important reasons. First, the material from which the tool is made has to survive much more strenuous conditions because of the strength of steel. Second, there are also numerous ways in which steel can be satisfactorily and reliably welded.

Third, the consequences of phase transformations accompanying FSW have not been studied in sufficient depth. Finally, the variety of steels available is much larger than for any other alloy system, requiring considerable experiments to optimize the weld for a required set of properties. Early optimism that FSW will become a commercially attractive method for the fabrication of ships, pipes, trucks, railway wagons and hot plate has not yet come to fruition.

That the application of FSW to steels is premature is emphasized by the fact that with few exceptions, only elementary mechanical properties have been characterized; most reports are limited to simple bend, tensile and hardness tests. For serious structural applications of the type proposed above it would be necessary to assess fracture toughness and other complex properties in greater depth. There are indications that elongation suffers following FSW. A typical temperature profile behind a friction-stir weld on steel. The maximum temperature reached is less than 1200°C and the time t_8-5 taken to cool over the range 800–500°C is about 11 sec. Therefore, the metallurgical transformations expected on the basis of cooling rates alone are not expected to be remarkably different from ordinary welds.

Various researches show that the yield and ultimate tensile strengths has been improved up to 100% as compared to the base parent metal of the joints. In the FSW, the materials do not go into molten state and then it does not solidify. This is why aluminum which is practically difficult to be welded using fusion joining techniques is weldable in this case. Also, the joint achieved in FSW of such aluminum metal are defect free [2, 5, 6–10]. FSW not only finds its application in case of soft materials but it can be employed in variety of harder and dissimilar materials. A vast majority of research has been carried out of material ranging from low to intermediate melting points, i.e. Mg-alloys and Cu-alloys and its process efficiency has been determined. Furthermore, tests have also been done on high strength structural materials with high melting points, i.e. Fe, Ti and Ni alloys, dissimilar alloys, metal matrix composites, polymers, etc. [11–13]. Metal matrix composites (MMCs) due to their excellent mechanical, physical and tribological properties are of great interest in recent decades. They possess characteristics such as lightweight, high strength, high stiffness, wear resistance, and creep resistance, high electrical and thermal conductivity [7]. Friction stir welding offers ease of handling, precise external process control and high

levels of repeatability thus creating very homogeneous welds. FSW need not any preparation of the sample and pollution created during the welding process is also very less.

Sahlot et al.[14] investigated the dissimilar lap joint of CuCrZr alloy and 316L stainless steel using friction stir welding. The thickness of the CuCrZr alloy is 6mm and 316L stainless steel is 3mm. Along the weld cross section, the higher load bearing ability was achieved throughout the joint due to the strong mechanical interlocking feature as well as indicating the good mechanical bonding between Cu and Fe. As the formation of pronounced hooking occurs at the Cu/ steel alloy interface, the mechanical interlocking property can be enhanced. To enhance the weld strength, the FSW parameters such as improving material flow, heat transfer, traverse speed, rotation speed and tool geometry can be optimized.

Moradi et al [15] examined the texture evolution and microstructure of friction stir welded dissimilar AA2024 and AA6061 alloys. The AA2024 were adjusted on advancing side as well as AA6061 on the retreating side on the bed of the machine. The thickness of both the sheets are 6mm. The fine equiaxed grain structure is observed in the stirred zone on the retreating and the advancing sides both in contingency of the static as well as dynamic recrystallisation. Due to the difference of the temperature on the retreating side and advancing side and initial size of precipitates, the higher volume of fraction of precipitates appears on the retreating side in the stirred zone. On the advancing side, there is less overall texture intensity on the contrary the texture intensity is increased on the retreating side stack up against with initial sheets. The initial elements completely eliminated on the both sides. Moreover, some strong shear textures observed owing to severe shear deformation during the FSW.

Infante et al.[16] studied the fatigue behaviour of dissimilar joints using FSW. The investigation is performed within Lightrain project and the objective is to improve the life cycle value of the passenger railway car. In the study, the two samples are taken. First is AA6082-T6 and AA5754-H111, and AA6082-T6 the thickness of the alloys is same i.e. 2 mm. The Lap joint specimens is tested on a constant amplitude loading in accordance with a stress ratio $R=0.1$. The tested specimen of fatigue analysis results in the comprehensive metallographic characterization of the welded zone. Moreover, the fatigue test results also show the hardness distribution at the welded zone. The base metal AA5754 and AA6082 have the higher fatigue strength than the similar and dissimilar joints as there is a hook defect in the weld joint. Improvement in fatigue performance is observed at lower applied stress ranges, the fatigue performance results in the dissimilar AA6082 and AA5754 FSW weld specimens shows a shallower S-n curve as compared with the AA6082-AA6082 FSW weld specimens.

Zandsalami et al. [17] analyzed the mechanical properties of the dissimilar 6061 aluminum alloy and 430 stainless steel. The thickness of the base metals taken as 5mm. The microstructure of the joints is examined by the Energy dispersive X-ray, Scanning electron microscopes and optical microscopy. Moreover, mechanical properties are evaluated by tensile and microhardness test. The best microstructure is obtained at a rotational speed of 900 r/min, tool offset of zero and a traverse speed of 120 mm/min.

The most significant factor is tool offset in accordance with the weld quality. A composite structure has been shown in the stir zone of the weld joint which consist of the dispatched steel particles presented in aluminium.

The best joint quality is obtained at an offset of zero, includes the serrated nature as well as the mechanical locking of the dissimilar weld joint. At the values above and below the zero offsets, the formation of weld defects such as voids and microcracks decreased the tensile strength of the weld joint.

Ahmed et al. [18] investigated the similar and dissimilar friction stir welding of AA7075 and AA5083. The type of joint is the butt joint. The friction stir welding is done at a rotational speed of 300 rpm and several traverse welding speed of 50, 100, 150 and 200 mm/min. With the use of electron backscattered technique, the crystallographic textures and microstructures are observed. The tensile and microhardness test is done to examine the mechanical properties. As the welding speed increases from 50 mm/min to 200 mm/min, results into a reduced grain from $6\mu\text{m}$ to $2\mu\text{m}$ size of similar AA7075 as well as in case of AA5083 from $9\mu\text{m}$ to $3\mu\text{m}$. In case of dissimilar welding, there is no significant with the average grain size of $4\mu\text{m}$ in the two cases of welding speed of 50mm/min and 200 mm/min. In the Nugget zone, the crystallographic texture reflects simple shear texture irrespective of the effect of the welding speed in similar and dissimilar weld joints. In the similar weld joints, the hardness profile reflects the typical behaviour with the reduction in hardness in nugget zone of AA7075. In the nugget joint of similar AA5083 weld joints, there is a increase in the hardness number. In the dissimilar weld joint, it is observed that there is a smooth transition in the hardness among the two hardness alloys. The experiments showed that the ultimate tensile strength examined the values in between the 245MPa and 267 MPa with efficiency of joint ranges from 77% and 87% in accordance with the strength AA5083BM. The dissimilar weld joint shows a brittle and ductile fractographic features such as grain boundary cleavage, facets decohesion and dimples.

Mehta et al. [19] analyzed the conventional and cool assisted FSW of AA6061 and AZ31B alloys. The thickness of 6mm is used for the base materials. This process of welding joint is analyzed by visual inspection, scanning electron micrographs, optical macro plus microscopy, energy dispersive X-ray spectroscopy, X-ray diffractions, microhardness indentation and tensile testing. In the nugget zone, it is observed the presence of onion rings comprises of various phases as Mg in an aluminum matrix and Al in Mg matrix. Moreover, there are intermetallic compounds such as $\text{Al}_{13}\text{Mg}_2$ and $\text{Al}_{12}\text{Mg}_{17}$. A diffusion layer has been observed on the aluminum side. Moreover, there is no presence of diffusion layer on the Mg side. The tensile strength is improved by cool assisted welding process as there is decrement in the intermetallic compounds along the weld bead. There is a highest hardness peak are analyzed in the nugget zone when the welding is done by conventional method.

Celik and Cakir [20] investigated the mechanical and microstructural properties of Al-Cu butt joint by the friction stir welding. The parameters are taken at different tool traverse speed (20, 30, 50 mm/min) and tool rotational speed ranging from 630 rpm, 1330 rpm and 2440 rpm with four various tool position (0, 1, 1.5, 2 mm). The microstructure are observed by the optical

microscope and SEM with EDS. X-Ray diffraction is to determine the intermetallic phases that is presented in weld zone. Along the side of Fine Cu particles, high tensile strength is observed.

Husain Mehdi et al. [21-24], investigated the effect of friction stir processing on TIG-welded joints with different fillers were used to improve the mechanical properties of TIG-welded joints, the FSP tool pin rotates on an already welded joint by TIG welding to lower the welding load and improve the weld quality by adjusting the processing parameters of friction stir processing. After analyzing the mechanical properties of TIG + FSP-welded joint, computational fluid dynamics-based numerical model was developed to predict the temperature distribution and material flow during TIG + FSP of dissimilar aluminum alloys AA6061 and AA7075 by ANSYS fluent software.

Ghaffarpour et al. [25] investigated the microstructure and mechanical properties of dissimilar aluminum sheets I.e. 5083-H12 and 6061-T6 welded by friction stir welding. The optimization of FSW parameters by DOE and RSM techniques. A very little difference is observed between the measured and predicted strength of the components. Dissimilar weld joint alloys 5083-H12 and 6061-T6 has the lower hardness than that of both BMs of material in the stir zone. In case of dissimilar alloys, HAZ shows the lower hardness in comparison with other zones of welding. The HAZ of AA6061-T6 is observed with minimum hardness. The outcomes of tensile test as well as hardness test are same with results of LDH tests. By enhancing the rotational speed results in decrement in hardness in stir zone.

Rec et al [26] analyzed the effect of process parameters upon the mechanical properties and microstructures of dissimilar weldments AA7075-T651 and AA5083-H111 alloys. The various parameters (tool pin design, tool rotational speed and configuration of joined alloys) are taken. According to the study, there is influence of alloy placement and tool rotational speed on the formation of weld. In accordance with the configuration, the AA5083-H111 alloy is on the advancing side and the AA7075-T651 is on the retreating side. Moreover, higher mixing of both materials is obtained at high rotational speeds. Despite of this, more welding defects such as voids, porosity and wormholes were observed in the stir zone of the weld joint. There is increase in tool rotational speed results into the decrement of the mechanical properties irrespective of the configuration and pin design. The higher weld efficiency and tensile strength is achieved by the use of triflate pin. There is no effect of configuration on the mechanical properties. The best defect free weld is obtained when triflate pin with the configuration (AA5083 is on the advancing side and AA7075 is on the retreating side) with a tool rotational speed of 280 rpm.

Mehta et al [27] obtained the effect of tilt angle on the microstructural and mechanical properties of dissimilar FSW of as electrolytic tough pitch copper and aluminum 6061-T651.

In this experiment, the tool tilt angle ranges from 0° to 4° with the regular interval of 1°. Moreover, the various parameters such as welding speed, workpiece material position, tool pin offset and tool rotational speed are kept constant. The various examinations such as microstructure analysis, scanning electron microscopy, macro hardness test, tensile test and energy dispersive x-ray spectrographic test to investigate the weld joint properties. The

results of various tests are show that the defect free weld at the tilt angles of 2°, 3°, 4°.The highest tensile strength and macro hardness is obtained at 4° in the nugget zone. At the copper side, thermo mechanically affected zone confirmed the weakest zone.

Ratnam et al [28] optimize to enhance the mechanical properties of dissimilar AA2024 and AA6061 alloys of 6mm by FSW. The chosen three levels are welding speed, tool rotational speed and tool tilt angle. The orthogonal array is taken as L27 and the results are obtained by Taguchi's ANOVA. For the tensile strength, the most significant factor is tool rotational speed and the least is welding speed. The best tensile strength is obtained at welding speed 11mm/min, tool rotational speed 1340 rpm and tool tilt angel of 2°. In accordance with the hardness, the most and least significant factor are tool rotational speed and welding speed respectively. At the welding speed of 11 mm/min, tool rotational speed of 2000 rpm and tool tilt angle of 3°, the optimum hardness is achieved. A defect free weld joint is achieved by twin-pin tool. Husain Mehdi et al. [29-31], In tungsten inert gas welding (TIG), micro-cracks, porosity, coarse grain structure and high residual stress distribution were found due to persisting thermal conditions. The TIG welded joint is processed using friction stir processing with input process parameters to avoid these defects. The tensile test results shows that the hybrid TIG + FSP welded joint had higher tensile strength than TIG welded joint with filler ER4043, whereas the increment in the micro-hardness of TIG + FSP welded joint was observed. The grain size also decreases when tool pin rotates on TIG welding with different processing parameters. It was found that the maximum tensile stress, % elongation and micro-hardness at nugget zone for TIG + FSP welded joint.

Pourali et al [32] analyzed effect of welding parameters on the formation of intermetallic compounds in aluminum and steel FSW. Due to the major difference in steels and aluminum, there is formation of thick brittle intermetallic compounds at the weld joint. The dissimilar material thickness is taken as 2mm and Al1100 and St 37 low carbon steel were lap welded by FSW. The welding parameters were carried are rotational speeds (315 and 400 rpm) and welding speeds (50 and 63 mm/min). According to the EDS analysis, a thick layer of Fe-rich IMCs is obtained in weld joint interfaces up to 93µm as it does not show any effect on the joint strength. Moreover, welding defects such as voids results in the detrimental condition in the weld strength. Lower welding time as well as lower welding speed results in fine mechanical fixing and increase heat input is obtained with the high rotational speeds which confirms the good mechanical mixing and metallurgical bond. At the high rotational speed and lower welding speed, the tensile strength is optimized. At the welding speed of 50mm/min and tool rotational speed of 400 rpm results in the maximum shear tensile load is 1925 N. Among the all tensile specimens, the failure happens in the Al-side nugget zones.

Rogriduez et al [33] investigated the micro structural and mechanical properties of AA6061 and AA7050 aluminum alloys. Except the rotational speeds, the other parameters are kept constant. As the tool rotational speed was varied, the microstructure shows the presence of bands of mixed and unmixed elements that represents the extent of material mixing. Increase in the tool rotational speed results in the enhancing the

material intermixing as well as joint strength. From the scanning electron microscopy, it is evident that failure occurs in the stir zone at low tool rotational speed owing to improper material intermixing.

2. Conclusions

This paper contributes the review towards the recent research in friction stir welding of dissimilar welding for enhancing the mechanical properties of materials. The following conclusions were drawn:

- By increasing the welding speed, the mechanical properties such as tensile strength increases
- By increasing tool rotational speed the welding, the mechanical properties like tensile strength decreases.
- The best defect free weld is obtained when triflute pin with the configuration (AA5083 is on the advancing side and AA7075 is on the retreating side) with a tool rotational speed of 280 rpm.
- The results of various tests are show that the defect free weld at the tilt angles of 2°, 3°, 4°. The highest tensile strength and macro hardness is obtained at 4° in the nugget zone.
- The tensile strength is improved by cool assisted welding process as there is decrement in the inter-metallic compounds along the weld bead.
- The best microstructure and joint quality is achieved at tool offset zero. The defect such as voids, micro cracks arises which decreases the tensile strength of the weld joint.

References

- [1] Shah, L. H., Othman, N. H., & Gerlich, A. (2018, April 3). Review of research progress on aluminum–magnesium dissimilar friction stir welding. Science and Technology of Welding and Joining. Taylor and Francis Ltd.
- [2] Kumar, N., Yuan, W., & Mishra, R. S. R. S. (2015). Friction Stir Welding of Dissimilar Alloys and Materials. Friction Stir Welding of Dissimilar Alloys and Materials (pp. 1–126). Elsevier Inc.
- [3] Jamshidi Aval, H. (2015). Microstructure and residual stress distributions in friction stir welding of dissimilar aluminum alloys. Materials and Design, 87, 405–413.
- [4] Aydin, H., Bayram, A., Uğuz, A., & Akay, K. S. (2009). Tensile properties of friction stir welded joints of 2024 aluminum alloys in different heat-treated-state. Materials and Design, 30(6), 2211–2221.
- [5] Wang, B., Lei, B. B., Zhu, J. X., Feng, Q., Wang, L., & Wu, D. (2015). EBSD study on microstructure and texture of friction stir welded AA5052-O and AA6061-T6 dissimilar joint. Materials and Design, 87, 593–599.
- [6] Guo, J. F., Gougeon, P., & Chen, X. G. (2012). Characterization of welded joints produced by FSW in AA 1100-B 4C metal matrix composites. Science and Technology of Welding and Joining, 17(2), 85–91.
- [7] Salih, O. S., Ou, H., Sun, W., & McCartney, D. G. (2015). A review of friction stir welding of aluminum matrix composites. Materials and Design, 86, 61–71.
- [8] Venkateswarlu, D., Nageswara rao, P., Mahapatra, M. M., Harsha, S. P., & Mandal, N. R. (2015). Processing and Optimization of Dissimilar Friction Stir Welding of AA 2219 and AA 7039 Alloys. Journal of Materials Engineering and Performance, 24(12), 4809–4824.
- [9] Hu, Z. L., Wang, X. S., & Yuan, S. J. (2012). Quantitative investigation of the tensile plastic deformation characteristic and microstructure for friction stir welded 2024 aluminum alloy. Materials Characterization, 73, 114–123.
- [10] Guo, J., Gougeon, P., & Chen, X. G. (2012). Microstructure evolution and mechanical properties of dissimilar friction stir welded joints between AA1100-B 4C MMC and AA6063 alloy. Materials Science and Engineering A, 553, 149–156.
- [11] Padhy, G. K., Wu, C. S., & Gao, S. (2015, November 1). Auxiliary energy assisted friction stir welding – Status review. Science and Technology of Welding and Joining. Maney Publishing.
- [12] Dialami, N., Chiumenti, M., Cervera, M., & Agelet de Saracibar, C. (2017). Challenges in Thermo-mechanical Analysis of Friction Stir Welding Processes. Archives of Computational Methods in Engineering, 24(1), 189–225.
- [13] Peel, M., Steuwer, A., Preuss, M., & Withers, P. J. (2003). Microstructure, mechanical properties and residual stresses as a function of welding speed in aluminum AA5083 friction stir welds. Acta Materialia, 51(16), 4791–4801.
- [14] Sahlot, P., Nene, S. S., Frank, M., Mishra, R. S., & Arora, A. (2018). Towards attaining dissimilar lap joint of CuCrZr alloy and 316L stainless steel using friction stir welding. Science and Technology of Welding and Joining, 23(8), 715–720.
- [15] Moradi, M. M., Jamshidi Aval, H., Jamaati, R., Amirkhani, S., & Ji, S. (2018). Microstructure and texture evolution of friction stir welded dissimilar aluminum alloys: AA2024 and AA6061. Journal of Manufacturing Processes, 32, 1–10.
- [16] Infante, V., Braga, D. F. O., Duarte, F., Moreira, P. M. G., De Freitas, M., & De Castro, P. M. S. T. (2016). Study of the fatigue behaviour of dissimilar aluminum joints produced by friction stir welding. In International Journal of Fatigue (Vol. 82, pp. 310–316). Elsevier Ltd.
- [17] Zandsalimi, S., Heidarzadeh, A., & Saeid, T. (2019). Dissimilar friction-stir welding of 430 stainless steel and 6061 aluminum alloy: Microstructure and mechanical properties of the joints. Proceedings of the Institution of Mechanical Engineers, Part L: Journal of Materials: Design and Applications, 233(9), 1791–1801.
- [18] Ahmed, M. M. Z., Ataya, S., El-Sayed Seleman, M. M., Ammar, H. R., & Ahmed, E. (2017). Friction stir welding of similar and dissimilar AA7075 and AA5083. Journal of Materials Processing Technology, 242, 77–91.
- [19] Mehta, K. P., Carlone, P., Astarita, A., Scherillo, F., Rubino, F., & Vora, P. (2019). Conventional and cooling assisted friction stir welding of AA6061 and AZ31B alloys. Materials Science and Engineering A, 759, 252–261.
- [20] Celik, S., & Cakir, R. (2016). Effect of friction stir welding parameters on the mechanical and microstructure properties of the Al-Cu butt joint. Metals, 6(6).
- [21] Husain Mehdi, R.S. Mishra, Influence of Friction Stir Processing on Weld Temperature Distribution and Mechanical Properties of TIG-Welded Joint of AA6061 and AA7075. Transactions of the Indian Institute of Metals, (2020). <https://doi.org/10.1007/s12666-020-01994-w>.
- [22] Husain Mehdi, R.S. Mishra, Effect of friction stir processing on mechanical properties and heat transfer of TIG welded joint of AA6061 and AA7075, Defense Technology (2020), <https://doi.org/10.1016/j.dt.2020.04.014>.
- [23] Husain Mehdi, R.S. Mishra, Investigation of mechanical properties and heat transfer of welded joint of AA6061 and AA7075 using TIG+FSP welding approach, Journal of Advanced Joining Processes Volume 1, 2020, <https://doi.org/10.1016/j.jajp.2020.100003>.
- [24] Husain Mehdi, R.S. Mishra (2019), Analysis of Material Flow and Heat Transfer in Reverse Dual Rotation Friction Stir Welding: A Review, International Journal of Steel Structure, vol-19, issue 2, pp 422-434.
- [25] Ghaffarpour, M., Kolahgar, S., Dariani, B. M., & Dehghani, K. (2013). Evaluation of dissimilar welds of 5083-H12 and 6061-T6 produced by friction stir welding. Metallurgical and Materials Transactions A: Physical Metallurgy and Materials Science, 44(8), 3697–3707.
- [26] Kalembe-Rec, I., Kopyściński, M., Miara, D., & Krasnowski, K. (2018). Effect of process parameters on mechanical properties of friction stir welded dissimilar 7075-T651 and 5083-H111 aluminum alloys. International Journal of Advanced Manufacturing Technology, 97(5–8), 2767–2779.
- [27] Mehta, K. P., & Badheka, V. J. (2016). Effects of tilt angle on the properties of dissimilar friction stir welding copper to aluminum. Materials and Manufacturing Processes, 31(3), 255–263.
- [28] Ratnam, C., Sudheer Kumar, B., & Sunil Ratna Kumar, K. (2018). Optimization of friction stir welding parameters to improve the mechanical properties of dissimilar AA2024 and AA6061 aluminium alloys. International Journal of Mechanical and Production Engineering Research and Development, 8(6), 937–944.
- [29] Husain Mehdi, R.S. Mishra (2016), Mechanical Properties and Microstructure Studies in Friction Stir Welding (FSW) Joints of Dissimilar Alloy A Review, Journal of Achievements of Materials and Manufacturing Engineering vol-77, issue 1, pp 31-40.

- [30] Husain Mehdi, R.S.Mishra (2019), Study of the influence of friction stir processing on tungsten inert gas welding of different aluminum alloy, SN Applied Sciences, 1 (7) 712. <https://doi.org/10.1007/s42452-019-0712-0>.
- [31] Husain Mehdi, R.S. Mishra (2020), Effect of Friction Stir Processing on Microstructure and Mechanical Properties of TIG Welded Joint of AA6061 and AA7075, Metallography, Microstructure, and Analysis (Springer), <https://doi.org/10.1007/s13632-020-00640-7>
- [32] Pourali, M., Abdollah-zadeh, A., Saeid, T., & Kargar, F. (2017). Influence of welding parameters on intermetallic compounds formation in dissimilar steel/aluminum friction stir welds. Journal Of Alloys And Compounds, 715, 1-8.
- [33] Rodriguez, R. I., Jordon, J. B., Allison, P. G., Rushing, T., & Garcia, L. (2015). Microstructure and mechanical properties of dissimilar friction stir welding of 6061-to-7050 aluminum alloys. Materials and Design, 83, 60–65.

Cite this article as: R.S. Mishra, Shivani Jha, Enhancements in mechanical properties of dissimilar materials using friction stir welding (FSW) - A review, International Journal of Research in Engineering and Innovation Vol-4, Issue-3 (2020), 154-160. <https://doi.org/10.36037/IJREI.2020.4306>.

Evaluating Deep Neural Network Ensembles by Majority Voting cum Meta-Learning scheme

Anmol Jain, Aishwary Kumar, Seba Susan^{[0000-0002-6709-6591]*}

Department of Information Technology,
Delhi Technological University,
Bawana Road, Delhi, India-110042
seba_406@yahoo.in

Abstract. Deep Neural Networks (DNNs) are prone to overfitting and hence have high variance. Overfitted networks do not perform well for a new data instance. So instead of using a single DNN as classifier we propose an ensemble of seven independent DNN learners by varying only the input to these DNNs keeping their architecture and intrinsic properties same. To induce variety in the training input, for each of the seven DNNs, one-seventh of the data is deleted and replenished by bootstrap sampling from the remaining samples. We have proposed a novel technique for combining the prediction of the DNN learners in the ensemble. Our method is called *pre-filtering by majority voting coupled with stacked meta-learner* which performs a two-step confidence check for the predictions before assigning the final class labels. All the algorithms in this paper have been tested on five benchmark datasets namely, Human Activity Recognition (HAR), Gas sensor array drift, Isolet, Spambase and Internet advertisements. Our ensemble approach achieves higher accuracy than a single DNN and the average individual accuracies of DNNs in the ensemble, as well as the baseline approaches of plurality voting and meta-learning.

Keywords: Deep neural network (DNN), Ensemble, Majority voting, Meta-learning, Bootstrap sampling.

1 Introduction

Deep Neural Network (DNN) has multiple hidden layers and each hidden layer has hundreds or thousands of activation units present in it [1]. When we use DNN as the classifier, issues of computational expense and overfitting of data may crop up. DNNs usually exhibit high variance for small real-world datasets. Because of high variance, when a novel data instance is fed, the model does not perform well. So instead of a single DNN, we propose to use an ensemble of multiple DNNs [2], each of them making independent errors, and we can combine their predictions in some manner to get a better model. Using the DNN ensemble instead of a single DNN also minimizes, to some extent, the problem of convergence to local minima due to gradient descent optimization [3, 4]. Alternative solutions to usage of ensembles for inducing variety in

learning include the global optimization of network weights using evolutionary algorithms such as Particle Swarm Optimization [5].

Suppose we have n number of independent DNN learners and let m_j be the output of the j^{th} learner. The combined variance of the net ensemble output y can be written as

$$\text{Var}(y) = \text{Var} \left(\sum_j \frac{1}{n} * m_j \right) \quad (1)$$

which can be rewritten as in (2) and (3).

$$\text{Var}(y) = \frac{1}{n^2} * n * \text{Var}(m_j) \quad (2)$$

$$\text{Var}(y) = \frac{1}{n} * \text{Var}(m_j) \quad (3)$$

i.e. the variance of an ensemble gets reduced by a factor of n assuming that the n learners are uncorrelated. The entire concept of getting a good ensemble is to get independent learners instead of individual good learners. We can get independent learners in an ensemble by sampling the training set with replacement, also called bootstrap aggregation or bagging [6]. We create a subsample of size s from the initial dataset of size n . Sampling is done in such a way that the subsample of size s is identically and independently distributed (IID) and can be considered as representative for the whole sample. It is also possible to have different training subsets by using feature selection method, selecting only a subset of attributes to train each learner [7], resulting in diverse and independent learners. In some cases, where the dataset is small, some random noise could be introduced (such as gaussian noise) in the training data, to get independent learners, that reduces the generalization error [8]. Adding noise to the training data contributes to a regularization factor that reduces the variance. It is also possible to have DNNs with different architectures and varying hyper-parameters such as different number of hidden layers and activation units in each layer [9], batch size, number of epochs, activation functions etc. In our paper, we investigate DNN ensembles with induced variety in training input, and explore various techniques of combining the DNN outputs in an effective manner. The organization of this paper is as follows. Basic ensemble concepts are revisited in section 2, the proposed DNN ensemble and learning methodology is presented in section 3, the experimentation and the results are discussed in section 4, and the final conclusions are given in section 5.

2 Combining the results of independent learners

Once we have a set of independent learners, the next step is to combine their predictions to get better predictions than the individual trained models. There are various approaches for fusing the outcomes of the n learners in an ensemble. We review several of the popular decision-fusion techniques next.

(i) *Model averaging or unweighted voting.* In this approach, the predicted probabilities of a class, from all the independent learners, are summed up and the final prediction is made by taking the maximum of all the probabilistic sums of predicted classes [10]. This works fine when we have a good estimation of probabilities. However, this method does not incorporate inter-classifier and intra-classifier biases.

(ii) *Weighted model averaging or weighted voting.* When combining the predictions from the learners, there are some learners that are more significant as compared to the other learners. In that case, we allot weights to predictions corresponding to the significance-level of each independent learner [11]. The optimal weights can be assigned by using gradient descent optimization procedure or using grid search. Simply, we can also assign weights proportional to accuracies of individual DNNs (Eq. (4)) or inversely proportion to the variances contributed by them (Eq. (5)).

$$weight \propto accuracy \quad (4)$$

$$weight \propto \frac{1}{variance} \quad (5)$$

However, some results also suggest that choosing optimized weights results in loss of generalization because of overfitting.

(iii) *Plurality voting.* In this approach each learner makes a prediction for a particular class label and the final candidate is chosen which has got the maximum votes [12]. Suppose we have three class labels c_1 , c_2 and c_3 and out of n learners n_1 learners predict c_1 , n_2 predict c_2 and n_3 predict c_3 . So, the final prediction is given as $\text{argmax}(n_1, n_2, n_3)$.

(iv) *Majority voting.* Slightly distinct from plurality voting due the incorporation of a threshold for the maximum votes obtained, here, every learner makes individual predictions and the candidate for final prediction is the one which gets more than half of the total votes [13]. If none of the class labels qualifies the criteria then that test instance is considered as an error.

(v) *Meta-learner classifier.* The outcomes from all the base learners in the ensemble are treated as level 1 predictions; these predictions are fed as input to a meta-learner classifier [14] (also called level 2 learner). The outcomes from level 2 are treated as the final outcomes for a particular test instance. A most common instance in machine learning is using a meta-learner classifier for learning the hidden state activations (feature-vectors) of independent learners, after fusing the features by concatenation [15]. In ensemble learning, we fuse the predicted outputs of independent learners rather than their hidden state activations. Meta-learners include Neural Networks or fully connected dense layers [16], AdaBoost [17] and XGBoost [18].

3 Proposed ensemble approach

Our homogenous ensemble comprises of seven identical DNNs having two hidden layers with 1200 and 800 activations units, respectively. We followed a novel approach to diversify DNN learning by deleting $1/7^{\text{th}}$ of the training data and replenishing the lost samples by randomly replicating the remaining samples, a concept known as bootstrap sampling [19]. The exercise is repeated for all seven DNNs in the ensemble (each time a different $1/7^{\text{th}}$ segment is deleted) to ensure variety in the training input, as shown in Fig. 1 (a). Different fusion strategies for combining the DNN outputs were tried in order to reduce the overall variance of the ensemble and to get better results in the final prediction. These are described next in the order from (i) to (iii).

(i) In the first experiment, we implemented plurality voting. We took the maximum of the probabilities predicted by a DNN for each class label and assigned the class label corresponding to that probability value to the test data. Output class labels from each of the DNNs were polled. We store the frequency of each class label as the result of polling. Finally, the test instance is allotted the class label with the maximum frequency.

$$Class = \text{argmax}(\text{fre}[y_k]) \quad (6)$$

where the array $\text{fre}[y_k]$ stores the count of class label y_k and $Class$ is the final class label predicted by our ensemble for the test instance.

(ii) In the second experiment, we used a meta-learner XGBoost for learning the DNN predictions and predicting the ensemble output. The resampled training set was given as input to the DNNs in the ensemble and their corresponding outputs were combined (level 1 prediction) as a feature-vector that was then used to train the meta-learner classifier XGBoost. Then for the test set, we first got level 1 prediction and then for the final ensemble output we passed it to the meta-learner.

(iii) The third experiment involves the proposed *pre-filtering by majority voting coupled with a stacked meta-learner* approach. When using the polling-based method to track count of each class label and determining the class label having maximum count, there could be cases when multiple class labels have the same frequency and also get the same share of maximum votes. In those cases, the polling-based method usually selects any one of those class labels as the final prediction. To make decisions in such cases, we first perform filtering of level 1 predictions using majority voting and then use the stacked meta-learner for the cases with no clear majority. The predictions having count of class label greater than or equal to $n-1$, where n is the ensemble size, were left as it is and the rest were filtered out and fed as input to the meta-learner. Addition of a filtering stage improves the performance of the meta-learner which is now trained on difficult cases only. If at least $n-1$ learners out of n predicted the same class for a given instance, we were highly confident of the prediction. The remaining filtered instances were passed to a meta-learner for further surety, which is a better approach for resolving conflicts of maximum votes. The training process for the meta-learner is shown in Fig. 1 (b). The test sample is presented to all the DNNs in the ensemble, and the DNN outputs are subject to majority voting. If the maximum number of votes is less than $n-1$, the DNN predictions are given as an input feature vector to the trained meta-learner for predicting the class label. The process flow for the test instance is

shown in Fig. 2. We also experimented by varying the number of DNNs in our ensemble from one to eight with step-size of one. Accuracies were recorded for each ensemble and the optimal ensemble size was determined. The graphs in Fig. 3 indicate that a choice of seven DNNs is optimal for our experiments.

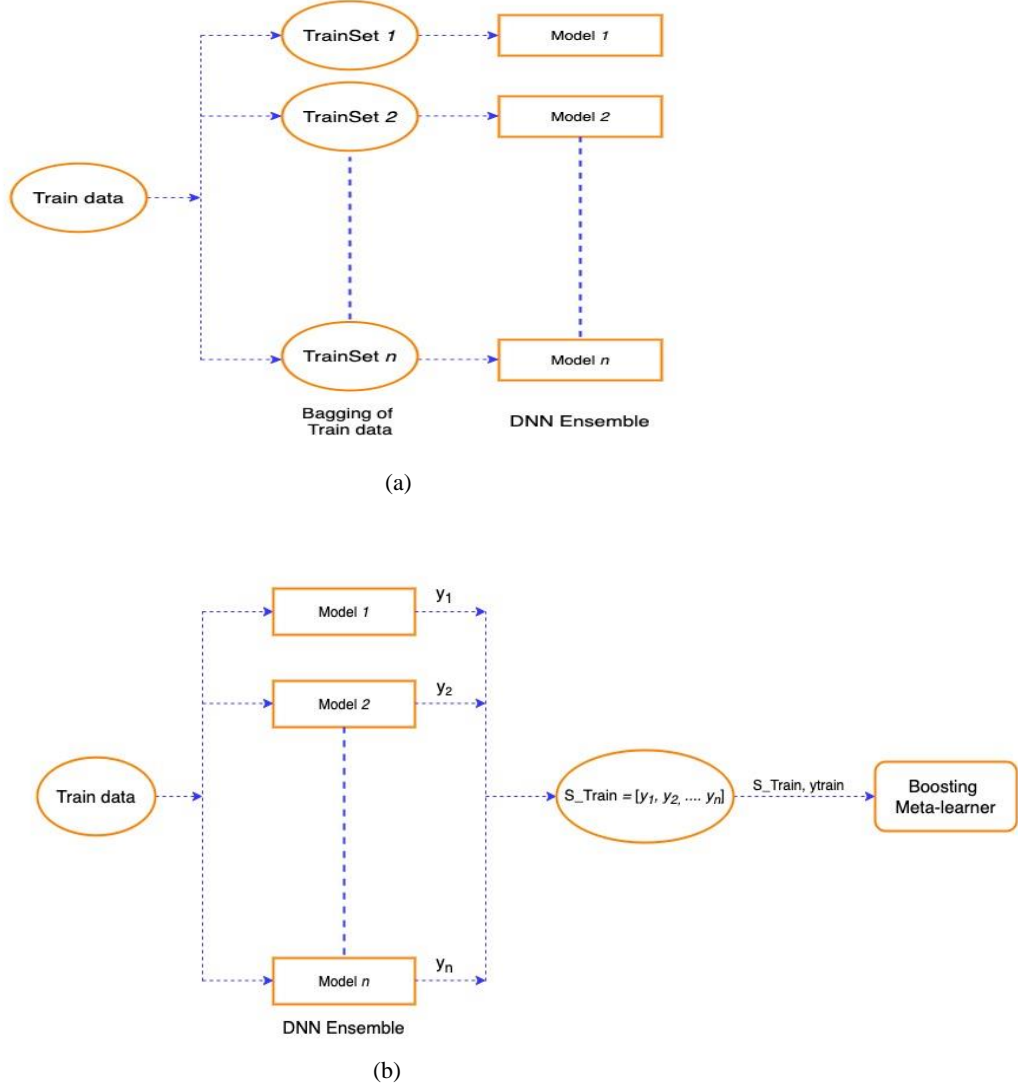


Fig. 1. DNN ensemble (a) training of individual DNNs (b) training the meta-learner for combining outputs of DNN ensemble

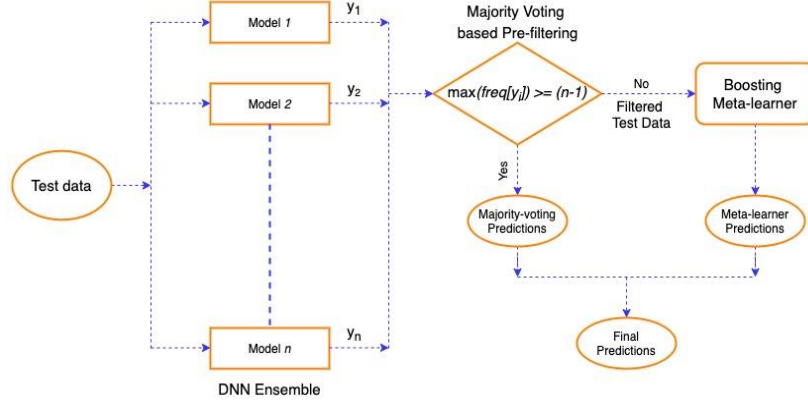


Fig. 2. Process flow for obtaining final prediction for a test sample.

4 Results

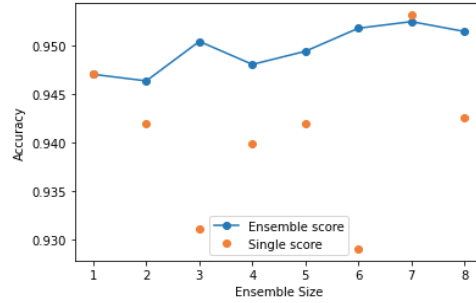
The software implementation was performed in Python 3.7 on an Intel dual-core processor. We conducted our experiments on five datasets namely, Human Activity Recognition (HAR) [20] and Gas sensor array drift, Isolet, Internet advertisements, and Spambase available in the UCI machine learning repository [21]. The HAR dataset is split into training and testing sets at the source. For the rest of the datasets, 80:20 split is used with the number of epochs set to 25. We choose an odd number of DNNs in our ensemble ($=7$). The optimal ensemble size was observed to be seven as demonstrated in the graphs in Fig. 3, drawn for the HAR dataset. The classification results are shown in Tables 1 (highest scores) and 2 (all fusion strategies). As observed from Table 2, all the fusion strategies of plurality voting, meta-learning and the proposed *pre-filtering by majority voting coupled with stacked meta-learner* gave accuracies that were better than the accuracy of the individual DNN and mean accuracy of all DNNs shown in Table 1. Our observations from Table 2 are: 1) Pre-filtering by majority voting increases the performance of meta-learning and overall gives a consistent performance 2) Meta-learning by itself is the second-best performer outperforming the proposed method in only one case out of five. 3) Plurality voting by itself does not perform as well. In the case of Spambase, the performance is unaffected by the choice of fusion strategy.

Table 1. Performance of DNN ensemble versus individual and mean DNN accuracies

Name of dataset	Individual Accuracy	Mean accuracy	(Highest) Ensemble Accuracy (refer Table 2)
Human activity recognition (HAR)	93.9%	94.5%	95.3%
Gas sensor array drift	96.0%	97.2%	98.1%
Isolet	94.5%	94.0%	96.2%
Internet advertisements	97.2%	97.5%	98.1%
Spambase	94.2%	93.6%	95.1%

Table 2. Accuracy of decision fusion strategies (highest accuracy highlighted in bold)

Name of dataset	Plurality voting (maximum votes)	Meta-learning with XGBoost	Majority voting cum Meta-learning (proposed)
Human activity recognition (HAR)	95.0%	95.0%	95.2%
Gas sensor array drift	97.9%	98.1%	98.1%
Isolet	96.0%	96.2%	96.1%
Internet advertisements	97.7%	97.7%	98.1%
Spambase	94.6%	94.6%	94.6%

**Fig. 3.** Variation of accuracy of HAR dataset using proposed method with the size of ensemble (average accuracy of DNNs for each ensemble shown in orange dots).

5 Conclusion

In this paper we propose a homogeneous ensemble learning approach using DNNs. The training input was diversified for the seven DNNs by careful sampling. The fusion strategies of plurality voting, meta-learning and the proposed *pre-filtering by majority voting coupled with stacked meta-learner* result in better accuracies as compared to the individual accuracies of DNNs and their mean accuracies, for the five datasets. The proposed fusion method improved the results of meta-learning in most of the cases.

References

1. CireşAn, Dan, Ueli Meier, Jonathan Masci, and Jürgen Schmidhuber. "Multi-column deep neural network for traffic sign classification." *Neural networks* 32 (2012): 333-338.s
2. Hong, Shenda, Meng Wu, Yuxi Zhou, Qingyun Wang, Junyuan Shang, Hongyan Li, and Junqing Xie. "ENCASE: An ENsemble CLASSifiEr for ECG classification using expert features and deep neural networks." In *2017 Computing in Cardiology (CinC)*, pp. 1-4. IEEE, 2017.

3. Gori, Marco, and Alberto Tesi. "On the problem of local minima in backpropagation." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 1 (1992): 76-86.
4. Susan, Seba, Rohit Ranjan, Udyant Taluja, Shivang Rai, and Pranav Agarwal. "Neural net optimization by weight-entropy monitoring." In *Computational intelligence: theories, applications and future directions-volume II*, pp. 201-213. Springer, Singapore, 2019.
5. Susan, Seba, Rohit Ranjan, Udyant Taluja, Shivang Rai, and Pranav Agarwal. "Global-best optimization of ANN trained by PSO using the non-extensive cross-entropy with Gaussian gain." *Soft Computing* (2020): 1-13.
6. Breiman, Leo. "Bagging predictors." *Machine learning* 24, no. 2 (1996): 123-140.
7. Ho, Tin Kam. "The random subspace method for constructing decision forests." *IEEE transactions on pattern analysis and machine intelligence* 20, no. 8 (1998): 832-844.
8. Raviv, Yuval, and Nathan Intrator. "Bootstrapping with noise: An effective regularization technique." *Connection Science* 8, no. 3-4 (1996): 355-372.
9. Partridge, Derek. "Network generalization differences quantified." *Neural Networks* 9, no. 2 (1996): 263-271.
10. Davidson, Ian, and Wei Fan. "When efficient model averaging out-performs boosting and bagging." In *European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 478-486. Springer, Berlin, Heidelberg, 2006.
11. Sollich, Peter, and Anders Krogh. "Learning with ensembles: How overfitting can be useful." In *Advances in neural information processing systems*, pp. 190-196. 1996.
12. Mu, Xiaoyan, Paul Watta, and Mohamad H. Hassoun. "Analysis of a plurality voting-based combination of classifiers." *Neural processing letters* 29, no. 2 (2009): 89-107.
13. Ruta, Dymitr, and Bogdan Gabrys. "Classifier selection for majority voting." *Information fusion* 6, no. 1 (2005): 63-81.
14. Wolpert, David H. "Stacked generalization." *Neural networks* 5, no. 2 (1992): 241-259.
15. Susan, Seba, and Jatin Malhotra. "Learning Interpretable Hidden State Structures for Handwritten Numeral Recognition." In *2020 4th International Conference on Computational Intelligence and Networks (CINE)*, pp. 1-6. IEEE, 2020.
16. Madisetty, Sreekanth, and Maunendra Sankar Desarkar. "A neural network-based ensemble approach for spam detection in Twitter." *IEEE Transactions on Computational Social Systems* 5, no. 4 (2018): 973-984.
17. Freund, Yoav, Robert Schapire, and Naoki Abe. "A short introduction to boosting." *Journal-Japanese Society For Artificial Intelligence* 14, no. 771-780 (1999): 1612.
18. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785-794. ACM, 2016.
19. DiCiccio, Thomas J., and Bradley Efron. "Bootstrap confidence intervals." *Statistical science* (1996): 189-212.
20. Anguita, D., A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz. "A Public Domain Dataset for Human Activity Recognition using Smartphones." In *21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, pp. 437-442. CIACO, 2013.
21. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.



Evaluating the Effect of Process Parameters on FSP of Al5083 Alloy Using ANSYS

Shourya Sahdev, Himanshu Kumar, Ravi Butola*, Ranganath M. Singari

Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India

Corresponding Author Email: ravibutola33855@gmail.com

<https://doi.org/10.18280/acsm.450203>

ABSTRACT

Received: 4 January 2021

Accepted: 11 March 2021

Keywords:

friction stir processing, process parameters, aluminium 5083, numerical modelling

Friction stir processing (FSP), compared to other solid-state processing methods, is a one-step process that attains refinement and homogeneity in its Microstructure. The complex configuration of various kinds of welds in FSP and their 3-D (three-dimensional) nature makes it tough to develop an overall system of ruling equations for theoretically analysing the functioning of the friction stir processed materials. The experimental trials are usually expensive and time-consuming. These hurdles can be overcome often by doing numerical analysis. The mechanical and microstructural characteristics of the Stir-Zone can be precisely supervised by enhancing the parameters of tool design, material properties, parameters of friction stir processing, and active heating and cooling. In this study, the significance of process parameters during FSP of Aluminium 5083 and the role of numerical analysis using ANSYS Workbench in the prediction of material behaviour have been discussed.

1. INTRODUCTION

The Welding Institute UK in 1991 invented a solid-state metal welding process Friction Stir Welding (FSW) [1]. In this method, a non-depleting rotating tool (of a material tougher than the workpiece material) is plunged into the abutting edges of the workpiece, followed by the translation of the rotating tool relative to the workpiece to form a weld next to the joint line as shown in (Figure 1). This consequently results in the Severe Plastic Deformation (SPD) and dynamic recrystallization in the weld region at elevated temperatures lower than the MP (melting point) of the workpiece material [2, 3]. FSW is widely adopted for joining hard-to-weld metallic alloys in different industry fields and is considered to be more effective than conventional fusion welding [4, 5]. Friction Stir Welding can weld aluminium alloys or different metal alloys which are reckoned to be non-weldable by regular methods because of porosity in the fusion zone and poor microstructural solidification. Some alloys can also be resistant welded but it is considered to be costly because of surface preparation hence it is not a viable option. Also, the loss in mechanical characteristics of the base material is sufficiently lower than that obtained by conventional methods. In contrast to fusion welding, FSW has reduced deformation and gives rise to fine equiaxed recrystallized grains and good mechanical traits in the welded workpiece. FSW is considered to be an environment-friendly process as smoke, arc glares, and fumes are not produced during the process [6]. Microstructural evolution during the FSP changes the grain boundary character, granular size, texture and breakup and redistribution of dispersoids. FSW leads to the formation of distinct microstructural zones and each zone imparts different mechanical properties. The various zones formed during the FSW process are Thermo-mechanically Affected Zone (TMAZ), Heat Affected Zone (HAZ) and Stir Zone (SZ) [2, 7, 8]. The stir zone consists of fully recrystallized fine-grained

material and corresponds to the position of the tool pin during the joining process. The TMAZ is created on either part of the SZ and the temperature, microstructural changes and strain are lower in this area concerning the SZ. Proximate to the THAZ, the HAZ is created. HAZ is common to all joining techniques and corresponds to the area subjected to a thermal cycle but the material in this region is not deformed during the welding process. Besides, the direction of rotation of the tool also affects the microstructural properties of the workpiece. The forward-moving side is the region in which the course of tool rotation and the course of the translational tool motion is the same. The solid material commences altering into a semi-solid one in this region. The semi-solid material retreated and cooled on the retreating side. The direction of translational motion of the tool is opposite to the direction of tool rotation on this side. The mechanical features of the SZ are improved in comparison to the base metal due to the uniformly distributed particles and the homogenous microstructure present in the stir zone [9]. A better surface finish was observed in the SZ in comparison to the base metal. FSW led to the removal of surface defects like voids and cracks in the stir zone, the hardness values in the SZ were noted to be more uniform in comparison to the base metal and the ultimate tensile strength and yield strength also showed an improved value in the SZ. The process parameters (such as tool traverse speed, plunge depth, tool rotational speed, and tool tilt angle) considered during the process highly influence microstructural properties and the temperature values recorded in the SZ [10].

Considering the influence of FSW on the microstructural characteristics of the workpiece and the corresponding refinement in the mechanical characteristics of the SZ, a new metal processing mechanism established on the elemental fundamentals of the FSW process was developed named the Friction Stir Processing (FSP) [11]. FSP is a very effective solid-state processing method that is used to provide a localized modification and alter the microstructural properties

of metallic materials. FSP is widely used for surface applications as it leads to an improvement in tensile strength and hardness of the material. It also refines the grain structure and thus improving the mechanical and wear properties of the metal [12-14].

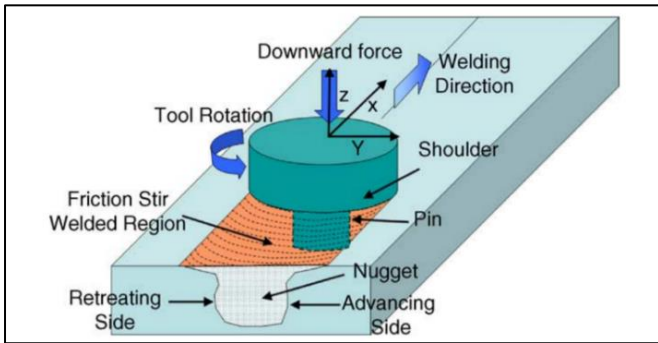


Figure 1. Schematic diagram of FSW [15]

Al-5083 (Al-Mg alloy) has been widely used in aerospace and transportation industries because of its good formability, high strength to weight ratio, excellent corrosion resistance and low density. As one of the solution-strengthened alloys, Al-5083 can only be strengthened by work hardening and micro alloying [15]. Therefore, FSP is a notable method used in the fabrication of Al-5083 for various surface applications. Though there is a notable amount of literature on the development of surface composites of Al-5083, there is a lack of literature on the role of numerical simulation in the material development process. Based on the above issues, two aspects need to be discussed in depth: the role of computational simulation in the material development process and the effect of various parameters on the workpiece material.

The objectives of this particular investigation are to analyse the experimental data collected corresponding to processing parameter values, to determine a relationship, to obtain the desired properties of the material. In this work, a third-dimensional (3-D) thermo-mechanical framework of the FSW of Aluminium 5083, an aluminium alloy is developed using the help of the Finite Element Method with ANSYS 18.1 software to comprehend and validate the role of process parameters in FSP. Four sets of process parameters were selected and a parametric analysis was conducted to ascertain

the influence of speed of rotation, the translational speed, and plunge depth on the thermal field around the Aluminium 5083 alloy during the FSW process.

2. FRICTION STIR PROCESSING

Mishra et al. developed a new metal processing technique based on the FSW process in 2002 [16]. It was initially proposed as a new technique for developing surface composites that were effective in increasing the microhardness of the material. It is a bulk processing technique. FSP has found various other applications such as to improve the malleability of materials, repair of casting defects, development of surface composite materials, modification of welded joints, etc. [17]. FSP is a comparatively new method of Super Plastic Deformation (SPD) in contrast to the other methods of SDP such as multi-directional forging (MAF), accumulative roll-bonding (ARB), high-pressure torsion (HPT), and equal channel angular pressing (ECAP) [18]. It is also faster than other solid-state processing methods. Z.Y. Ma et al. produced 7075Al alloy plates that were subjected to Super Plastic Deformation using FSP [19]. 7075Al alloys with a high-quality grain length were produced and led to substantially better superplastic ductility, decrease in flow stress, decreased optimum temperature, and a change to greater optimum strain rates. FSP has successfully processed materials such as AA2519, AA7075, AA5083 aluminium alloys, Stainless steel, AZ61 magnesium alloy, and nickel-aluminium bronze. FSP is considered to have several advantages as compared to the other SPD processes that have been stated as follows [20]:

- (1) FSP is more efficient in homogenizing powder metallurgy processed aluminium alloys.
- (2) It can modify the microstructure of the metal matrix composites and thus the joint area impart great metallurgical properties.
- (3) Helps to eliminate casting defects and is used for property enhancement in casting Al alloys.
- (4) It is able to break up or dissolve the second phase particles which bring about a considerable improvement in properties.
- (5) No surface cleaning is required before the process.

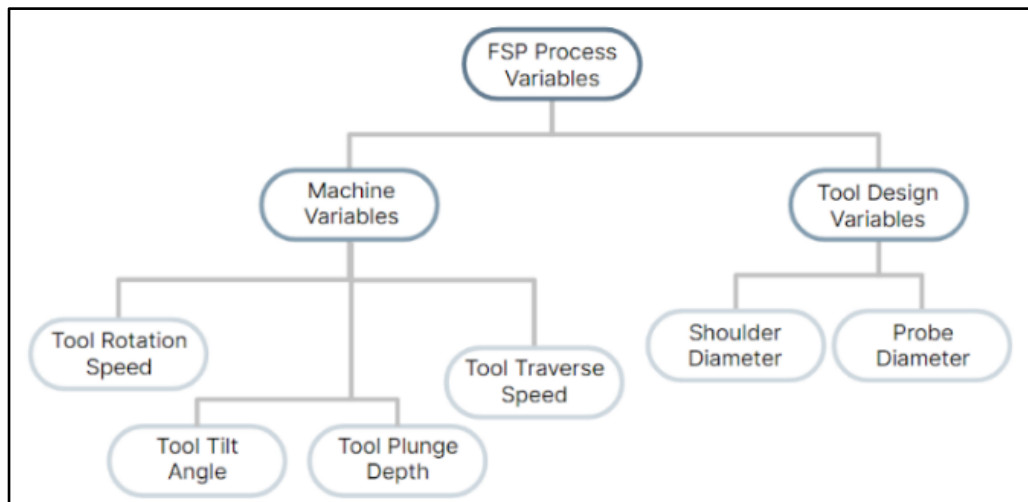


Figure 2. Effect of FSP process variables

The consequent microstructural change in metallic materials after FSP are highly influenced by various factors which include processing parameters, welding parameters, joint design, and tool geometry (Figure 2). Despite the fact that friction stir processing is a noble technique it is still not sufficiently mature for solving practical applications [21]. Therefore, mathematical physics models are used for parameter optimisation of the process and hence helps us formulate a suitable process. Some recent studies have suggested that the thermal fields during the FS process have a critical role in determining the resulting material properties and microstructure of the material [22]. Some of the important aspects of FSP have been discussed below.

2.1 Processing parameters

The tool geometry and welding parameters put a massive effect on the distribution of temperature and material flow behaviour and thus causing a change in the microstructure of the material [23]. It is of very much importance to find the ideal values for process parameters as they control the biggest component of the heating source. Low heat input causes an increase in the grain refinement, but a high heat input is needed to plasticise or soften up the material. Smaller grain size is observed, while the defects increase for a low value of rotational speed and a high value of traverse speed and vice versa [24]. Therefore, traverse and rotational speed needs to be optimised to attain a SZ with no defects and reduced grain size. Some of the major processing parameters have been discussed below.

2.1.1 Tool rotation rate

More dissolution of soluble particles and greater fragmentation of insoluble particles have been observed for an increment in tool rotation rates [25]. An increment in tool rotation rate improves the resultant grain size and leads to a substantial rise in the temperature of the stir zone. Tool rotation speed is also important for the stirring and unification of the workpiece material. Direction of tool rotation, affects the microstructural evolution of the material [26]. The microstructure of the material is not found to be symmetric w.r.t. the traverse length after a single pass of the tool [27], hence multiple passes of the tool are preferred as they lead to a more homogenous processed region.

2.1.2 Tool traverse speed

The tool traverse speed is accountable for moving the material from the initial to the rear portion of the workpiece. A low tool traverse speed causes a rise in the grain size of the FS processed zone. It has a negligible effect on the tool wear. An increase in traverse speed of the tool increases the microhardness of FS processed surface composites because of the higher distribution of reinforcement particles [28].

2.1.3 Tilt angle

The FSP tool tilt angle makes sure that the tool shoulder holds the stirred material and transfers it towards the back of the tool shoulder. A rise in the tool tilt angle brings about an increase in processing temperature [29]. A low value of the tilt angle is responsible for the defects in the processed zone while a higher value of the tilt angle increases the particle size and grain size of the processed zone [30].

2.1.4 Tool insertion depth

The tool insertion depth is responsible for maintaining contact with the molten metal and therefore ensure the production of a defect-free processed zone. It also helps to reduce the tool wear. But a high insertion depth can lead to an increase in the width of the plasticised region and a decrease in the hardness of the processed region [31]. An increase in tool insertion depth also leads to an increase in an excessive flash and hence a concave processed-zone is produced in such cases.

2.2 Microstructural evolution

FSP leads to refinement of grains due to dynamic recrystallization and therefore the enhancement of some mechanical properties [32]. FSP results in a significant improvement in microstructural refinement, an increase in density, and homogeneity of the processed zone. Some studies show that the fatigue strength decreases with decrease in ductility, and ultimate tensile strength but there is an improvement in the fatigue life of the processed material in comparison to the base metal [33]. Micrographs showed that differences in the size of inter-metallic bonding were found in FS processed materials which indicated that fracture is initiated due to the breaking of the brittle intermetallic bonding. Due to heterogeneous plastic deformation, the evolution of various grain particles may be different during FS Process [34]. No variation in structure is observed even after multiple passes, but it leads to the homogenisation of the microstructure. FSP is majorly used for the introduction of certain reinforcement particles in a material leading to the production of a surface composite that results in an improvement in properties of the material. According to the latest research, the process parameters contribute a lot towards microhardness of surface composites of FS processed materials, confirming that the most influential process parameters are rotational speed, reinforcement type and tool profile [35]. It was concluded that B₄C should be considered as a better reinforcement material in comparison to SiC and RHA for improving the microhardness value of the composites, the value increased by 1.5 to 1.6 times from that of the workpiece material. The effects of FSP on cast aluminium alloy were studied by Ma et al. [36]. It was concluded that FSP led to the fine break-up and redistribution of constituent particles thus homogenizing the cast microstructure and completely eliminating porosity. For homogenising the particle distribution in surface composites, FPS can be effectively used [37].

2.3 Surface composites

FSP is a versatile process used for the production of surface composites [38]. Engineering applications that involve surface interactions, Surface composites are ideal materials for those applications. The various types of surface composites under development are in-situ composites, micro composites, hybrid composites, and nano-composites [39-42]. Surface composites are used to improve surface properties like abrasion resistance, corrosion resistance, hardness, fatigue life, formability, strength, and ductility [43]. Al, Mg, Steel, and Ti-based alloys are used for surface composites. Butola et al. [44] successfully fabricated the SAM (Self-Assembled Monolayer) technique of Al-B₄C nano-ceramic surface composite by implementing FSP. The mechanical properties of fabricated nano-ceramic surface composites on evaluation after one pass exhibited

higher hardness, ultimate tensile strength, and finer grain structure in comparison to the base metal (BM). The fabrication of AA7075-T6 reinforced with SiC and Aloe vera ash, using FSP showed an improvement in wear and mechanical properties of all fabricated composite in comparison to the base metal [45]. Microhardness improved with the introduction of reinforcement.

3. MATERIAL PROPERTIES

H13 tool steel material was preferred as the tool material for the present study as it is recommended for the manufacturing of friction stir welding [46]. The material (Al 5083) was selected for the workpiece. Aluminium 5083 is highly resistant to attack by industrial chemical environments and seawater. It is also known for remarkable performance in extreme environments. The FSP of Al5083 is highly desirable to introduce surface modification and surface hardening and it is considered to have potential applications for manufacturing hybrid alloys. Vaira Vignesh et al. compared the intergranular corrosion of the Al5083 specimen with that of its FS Processed specimen [47] by applying the nitric acid mass loss test. The intergranular corrosion of the FS Processed specimen was lower than the base metal. R. Vaira Vignesh et al. also concluded that the FSPed specimens of Al5083 have better wear resistance as compared to the base metal. A. Yazdipour et al. studied the effect of cooling rate on the FSPed specimens of Al5083 [48]. It was observed that the cooling rate affects the final grain size and the grain growth of the metal. There is no effect of multiple passes on the microstructural and mechanical properties of the stir zone of the FSPed specimens of Al5083 [49]. Recent studies have shown that the inclusion of vibration during the FSP process leads to an improvement in mechanical and microstructural properties such as grain refinement, strength, microhardness and formability of the FSPed specimens of Al5083 [50, 51].

4. NUMERICAL ANALYSIS OF FSP

A three-dimensional (3-D) thermomechanical model of the FSP of Aluminium 5083 is developed with the help of Finite Element Method using the ANSYS 18.1 software in order to understand and validate the role of process parameters in FSW. The ANSYS software was selected on the basis of its superiority in terms of simulating mechanical properties, temperature distributions, heat transfer, and deformation [52]. The mathematical model developed for the Friction Stir Processing helps in simulating the temperature and the stress values, that are generated in the heat-affected zone. The major process parameters that govern the value of thermomechanical stress generated during the motion of the tool are its plunge depth, translational speed, rotational speed, and tool angle.

4.1 Finite Element Model (FEM) description

The model was simulated using transient structural analysis in ANSYS Mechanical APDL and user-defined functions were implemented to account for an extra DOF for temperature. To deal with distortion of the mesh and high calculation time, features of the code that are built-in such as Augmented Lagrangian and mass scaling method were used. The workpiece dimensions were taken to be $100 \times 50 \times 5$ mm

and the tool diameter was taken to be 20mm with a thickness of 5mm. The dwell time was taken to be 5 seconds. To study the effects of process parameters on the temperature distribution behaviour w.r.t displacement of the tool this model was used. The ANSYS model is shown in Figure 3 which depicts the direction of tool rotation, direction of tool translation and the constraints applied to the FEM model.

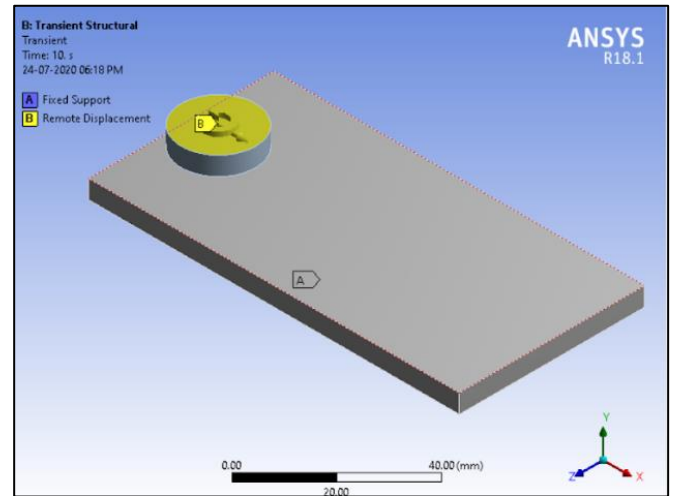


Figure 3. ANSYS model

4.2 Mesh properties

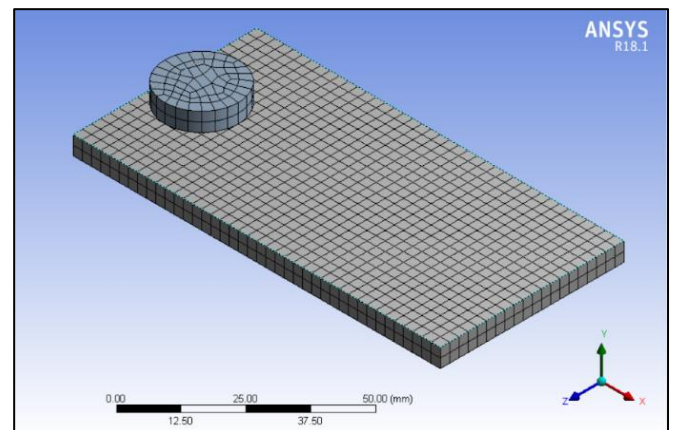


Figure 4. Meshed ANSYS FSP model

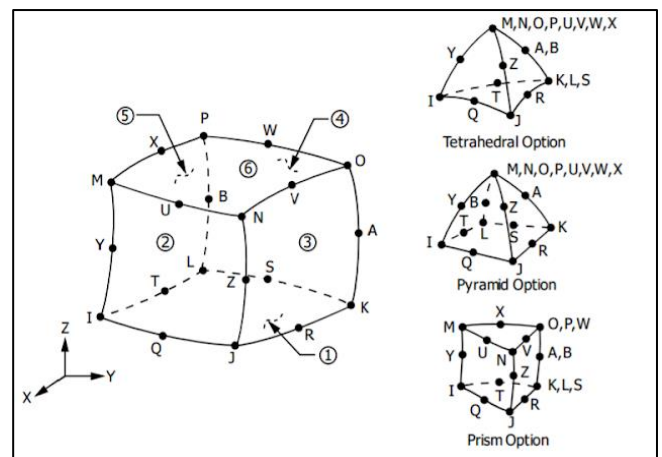


Figure 5. Types of Mesh in ANSYS

The mesh size was selected to be 2.5mm to ensure less computational time and maintain the result accuracy (Figure 4). The type of mesh used in this analysis was SOLID226 (Figure 5). SOLID226 is a 20-Node 3D Coupled-Field Solid that helps to enable an extra DOF of temperature for structural analysis at the contact surface, and thus account for the temperature change along with geometry. This type of mesh was selected because its structural capabilities include plasticity, elasticity, hyper elasticity, large strain, viscoelasticity, stress stiffening effects, viscoelasticity, creep, and large deflection.

4.3 Experimental procedure

Three steps namely preheating, plunging, and traversing were defined for simulation, and data was collected according to the process parameter values as presented in Table 1.

Table 1. Process parameter values

Experiment No.	Rotational Speed	Translational Speed	Plunge Depth
1.	200 rpm	60 mm/min	0.2
2.	500 rpm	60 mm/min	0.2
3.	200 rpm	40 mm/min	0.2
4.	200 rpm	60 mm/min	0.5

5. EFFECT OF PROCESS PARAMETERS

A complex material movement and plastic deformation can be attained by FSW/FSP. Welding parameters, joint design and tool geometry exert a large impact on the material flow behaviour and temperature distribution. The resultant microstructure and material properties are dependent on the temperature in the SZ. A lower temperature in the SZ (stir zone) leads to an increase in grain size. But to plasticize the material adequate heat input is required. Therefore, the correct value of process parameters needs to be selected so that it can obtain the material with desired properties. The results of the experimental analysis 1 (Figure 6) were compared with the other three analysis to conclude the effect of change of a specific process parameter. The Figures 6-11 depicts the rise in temperature value of the workpiece during FSP. The colour contour chart helps to denote the temperature value associated with a specific colour. Hence, the colour schematic helps to predict and compare the temperature profile that will be obtained on performing FSP using the specific set of process parameters on Al-5083 alloy. The significance of the process parameters considered for this analysis have been discussed below:

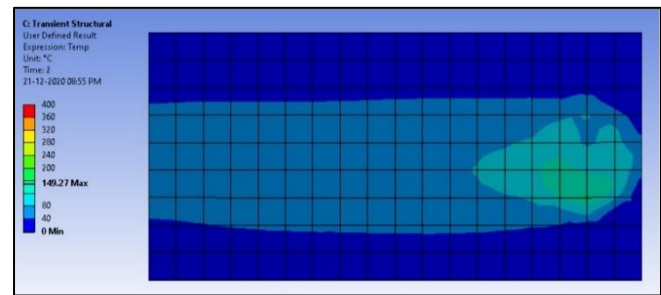


Figure 6. Experimental analysis 1 temperature results

5.1 Rotational speed

The rotation of the tool ends-up in mixing and stirring of material near the rotating pin. For higher the rotation of the tool, higher were the temperatures, due to an increased amount of frictional heating, which resulted in a more severe mixing and stirring of the material. A significant increase in temperature value was observed on increasing the rotational speed from 200rpm. The movement of the tool results in material being stirred and mixed around the spinning pin. Due to greater frictional heating and higher rates of tool rotation an increase in temperature and a more severe mixing and stirring of the given material is observed. A large increase in the temperature value was observed on increasing the rotational speed from 200rpm (Figure 6) to 500rpm (Figure 7). It also led to an increase in the overall heated area on the surface of the workpiece. A maximum temperature = 389.04°C was noted in the stir zone.

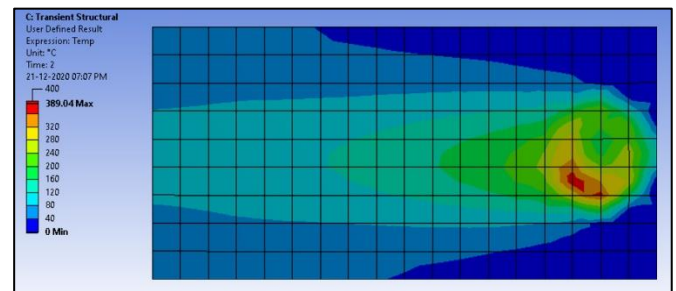


Figure 7. Experimental analysis 2 temperature results

5.2 Translational speed

The translational speed of the tool shifts the plasticized material from the initial position to the rear part of the workpiece. A substantial increase in the temperature value was observed on decreasing the translational speed of the tool from 60 mm/s (Figure 6) to 40 mm/s (Figure 8) with a subtle increase in the overall heated area of the workpiece surface. A maximum temperature of 194.61°C was noted in the stir zone.

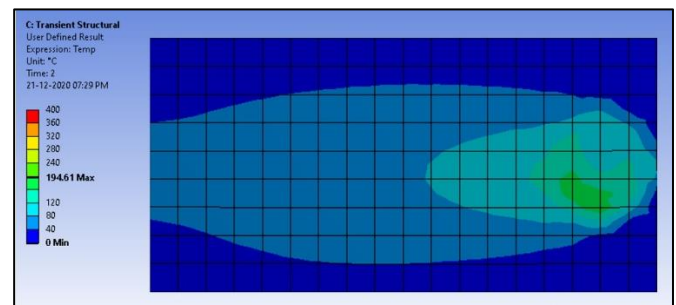


Figure 8. Experimental analysis 3 temperature results

5.3 Plunge depth

A rise in the depth of the plunge reduces mechanical prosperities and contributes to a faulty weld [14]. By raising the plunge depth from 0.2 mm (Figure 6) to 0.5 mm (Figure 9), a rise in the temperature value and a raise in the total heating area of the workpiece surface was observed. A maximum temperature of 328.8°C was noted in the stir zone.

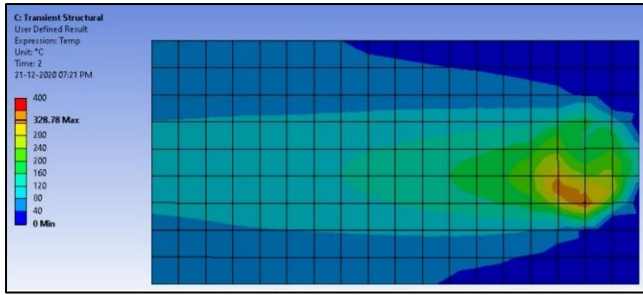


Figure 9. Experimental analysis 4 temperature results

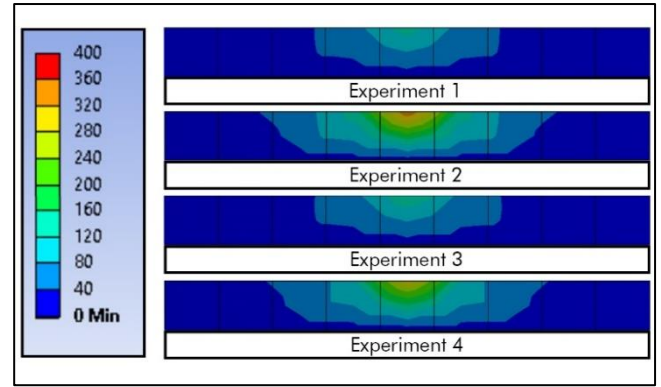


Figure 11. Temperature distribution results

6. ANALYSIS OF STRESS DISTRIBUTION

The difference in the frictional stress distribution result was studied, for the 4 considered experimental analysis. The highest value of frictional stress was observed in experiment 3, that had a higher value of tool traverse speed. The frictional stress is found to have a uniform value throughout the contact surface. Frictional stress increases proportionally to the relative velocity w.r.t. the surface of the tool in contact with the workpiece. The values of maximum frictional stress achieved in the experimental analysis have been discussed below in Table 2.

Table 2. Frictional stress results

Experiment No.	Maximum Value of Frictional Stress
1.	42.5 MPa
2.	72.1 MPa
3.	43.5 MPa
4.	99 MPa

7. ANALYSIS OF TEMPERATURE DISTRIBUTION

The difference in the temperature distribution results (along the thickness of the workpiece) were studied for the 4 considered experimental analysis. The highest value of temperature (389.04°C) was observed in the experimental analysis 2. The area associated with the heat-affected zone increased with an increase in the tool rotational speed. The surface at the top of the workpiece that is in direct contact with the tool is subjected to a higher level of distortion, hence the temperature is highest at the topmost surface [53]. The temperature gradually decreases as we go away from the contact point of tool and surface.

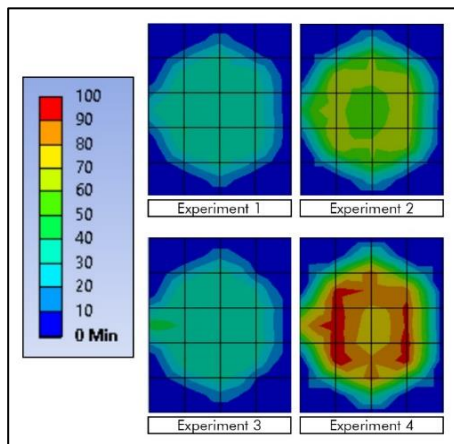


Figure 10. Frictional stress distribution results (on the surface)

8. CONCLUSIONS

- (1) No variation in structure is observed even after multiple passes but the homogenisation of the microstructure is still achieved.
- (2) The variation in temperature distribution due to the direction of tool rotation, can be seen throughout the surface.
- (3) Numerical analysis of the process using ANSYS helps to accurately predict the temperature distribution in the processing region w.r.t a change in tool translational speed, tool rotational speed and tool plunge depth.
- (4) The numerical analysis of the process is not sufficient to predict the mechanical properties of the processed material but helps to calculate a range of appropriate process parameter values that can help achieve the desired temperature range during the friction stir process.

REFERENCES

- [1] Thomas, W.M., Nicholas, E.D., Needham, J.C., Murch, M.G., Temple-Smith, P., Dawes, C.J. (1995). U.S. Patent No. 5,460,317. Washington, DC: U.S. Patent and Trademark Office.
- [2] Sato, Y.S., Kokawa, H., Enomoto, M., Jogan, S. (1999). Microstructural evolution of 6063 aluminum during friction-stir welding. *Metallurgical and Materials Transactions A*, 30(9): 2429-2437. <https://doi.org/10.1007/s11661-999-0251-1>
- [3] Murr, L.E., Flores, R.D., Flores, O.V., McClure, J.C., Liu, G., Brown, D. (1998). Friction-stir welding: Microstructural characterization. *Material Research Innovations*, 1(4): 211-223. <https://doi.org/10.1007/s100190050043>
- [4] Rhodes, C.G., Mahoney, M.W., Bingel, W.H., Spurling, R.A., Bampton, C.C. (1997). Effects of friction stir welding on microstructure of 7075 aluminum. *Scripta materialia*, 36(1): 69-75. [https://doi.org/10.1016/S1359-6462\(96\)00344-2](https://doi.org/10.1016/S1359-6462(96)00344-2)
- [5] Liu, G., Murr, L.E., Niou, C.S., McClure, J.C., Vega, F.R. (1997). Microstructural aspects of the friction-stir welding of 6061-T6 aluminum. *Scripta Materialia*, 37(3): 355-361. [https://doi.org/10.1016/S1359-6462\(97\)00093-6](https://doi.org/10.1016/S1359-6462(97)00093-6)
- [6] Swarnkar, A., Kumar, R., Suri, A., Saha, A. (2016). A review on Friction Stir Welding: An environment friendly welding technique. In 2016 IEEE Region 10

- Humanitarian Technology Conference (R10-HTC) IEEE, 1-4. <https://doi.org/10.1109/R10-HTC.2016.7906807>
- [7] Guerra, M., Schmidt, C., McClure, J.C., Murr, L.E., Nunes, A.C. (2002). Flow patterns during friction stir welding. *Materials Characterization*, 49(2): 95-101. [https://doi.org/10.1016/S1044-5803\(02\)00362-5](https://doi.org/10.1016/S1044-5803(02)00362-5)
 - [8] Storjohann, D., Barabash, O.M., David, S.A., Sklad, P.S., Bloom, E.E., Babu, S.S. (2005). Fusion and friction stir welding of aluminum-metal-matrix composites. *Metallurgical and Materials Transactions A*, 36(11): 3237-3247. <https://doi.org/10.1007/s11661-005-0093-4>
 - [9] Lee, W.B., Yeon, Y.M., Jung, S.B. (2003). The improvement of mechanical properties of friction-stir-welded A356 Al alloy. *Materials Science and Engineering: A*, 355(1-2): 154-159. [https://doi.org/10.1016/S0921-5093\(03\)00053-4](https://doi.org/10.1016/S0921-5093(03)00053-4)
 - [10] Pashazadeh, H., Teimournezhad, J., Masoumi, A. (2014). Numerical investigation on the mechanical, thermal, metallurgical and material flow characteristics in friction stir welding of copper sheets with experimental verification. *Materials & Design*, 55: 619-632. <https://doi.org/10.1016/j.matdes.2013.09.028>
 - [11] Ma, Z.Y. (2008). Friction stir processing technology: a review. *Metallurgical and materials Transactions A*, 39(3): 642-658. <https://doi.org/10.1007/s11661-007-9459-0>
 - [12] Srivastava, A.K., Maurya, N.K., Maurya, M., Dwivedi, S.P., Saxena, A. (2020). Effect of multiple passes on microstructural and mechanical properties of surface composite Al 2024/SiC produced by friction stir processing. *Annales de Chimie-Science des Matériaux*, 44(6): 421-426. <https://doi.org/10.18280/acsm.440608>
 - [13] Maurya, M., Kumar, S., Maurya, N.K. (2020). Composites prepared via friction stir processing technique: A review. *Revue des Composites et des Matériaux Avancés-Journal of Composite and Advanced Materials*, 30(3-4): 143-151. <https://doi.org/10.18280/rcma.303-404>
 - [14] Dwivedi, S.P., Srivastava, A.K., Maurya, N.K., Sahu, R. (2020). Microstructure and mechanical behaviour of Al/SiC/Agro-Waste RHA hybrid metal matrix composite. *Revue des Composites et des Matériaux Avancés-Journal of Composite and Advanced Materials*, 30(1): 43-47 <https://doi.org/10.18280/rcma.300107>
 - [15] Chen, Y., Ding, H., Li, J., Cai, Z., Zhao, J., Yang, W. (2016). Influence of multi-pass friction stir processing on the microstructure and mechanical properties of Al-5083 alloy. *Materials Science and Engineering: A*, 650: 281-289. <http://dx.doi.org/10.1016/j.msea.2015.10.057>
 - [16] Mishra, R.S., Ma, Z.Y., Charit, I. (2003). Friction stir processing: a novel technique for fabrication of surface composite. *Materials Science and Engineering: A*, 341(1-2): 307-310. [https://doi.org/10.1016/S0921-5093\(02\)00199-5](https://doi.org/10.1016/S0921-5093(02)00199-5)
 - [17] Węglowski, M.S. (2018). Friction stir processing—state of the art. *Archives of civil and Mechanical Engineering*, 18: 114-129. <https://doi.org/10.1016/j.acme.2017.06.002>
 - [18] Patel, V.V., Badheka, V., Kumar, A. (2016). Friction stir processing as a novel technique to achieve superplasticity in aluminum alloys: process variables, variants, and applications. *Metallography, Microstructure, and Analysis*, 5(4): 278-293. <https://doi.org/10.1007/s13632-016-0285-x>
 - [19] Ma, Z.Y., Mishra, R.S., Mahoney, M.W. (2002). Superplastic deformation behaviour of friction stir processed 7075Al alloy. *Acta Materialia*, 50(17): 4419-4430. [https://doi.org/10.1016/S1359-6454\(02\)00278-1](https://doi.org/10.1016/S1359-6454(02)00278-1)
 - [20] Mishra, R.S., Ma, Z.Y. (2005). Friction stir welding and processing. *Materials Science and Engineering: R: reports*, 50(1-2): 1-78. <https://doi.org/10.1016/j.mser.2005.07.001>
 - [21] Li, K., Liu, X., Zhao, Y. (2019). Research status and prospect of friction stir processing technology. *Coatings*, 9(2): 129. <https://doi.org/10.3390/coatings9020129>
 - [22] Darras, B.M., Omar, M.A., Khraisheh, M.K. (2007). Experimental thermal analysis of friction stir processing. In *Materials science forum*. Trans Tech Publications Ltd, 539: 3801-3806. <https://doi.org/10.4028/www.scientific.net/MSF.539-543.3801>
 - [23] Sidhu, M.S., Chatha, S.S. (2012). Friction stir welding—process and its variables: A review. *International Journal of Emerging Technology and Advanced Engineering*, 2(12): 275-279.
 - [24] Carlone, P., Palazzo, G.S. (2013). Influence of process parameters on microstructure and mechanical properties in AA2024-T3 friction stir welding. *Metallography, Microstructure, and Analysis*, 2(4): 213-222. <https://doi.org/10.1007/s13632-013-0078-4>
 - [25] Pasebani, S., Charit, I., Mishra, R.S. (2015). Effect of tool rotation rate on constituent particles in a friction stir processed 2024Al alloy. *Materials Letters*, 160: 64-67. <https://doi.org/10.1016/j.matlet.2015.07.074>
 - [26] Nascimento, F., Santos, T., Vilaça, P., Miranda, R.M., Quintino, L. (2009). Microstructural modification and ductility enhancement of surfaces modified by FSP in aluminium alloys. *Materials Science and Engineering: A*, 506(1-2): 16-22. <https://doi.org/10.1016/j.msea.2009.01.008>
 - [27] Pashazadeh, H., Teimournezhad, J., Masoumi, A. (2014). Numerical investigation on the mechanical, thermal, metallurgical and material flow characteristics in friction stir welding of copper sheets with experimental verification. *Materials & Design*, 55: 619-632. <https://doi.org/10.1016/j.matdes.2013.09.028>
 - [28] Ramezani, N.M., Davoodi, B., Aberoumand, M., Hajideh, M.R. (2019). Assessment of tool wear and mechanical properties of Al 7075 nanocomposite in friction stir processing (FSP). *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, 41(4): 182. <https://doi.org/10.1007/s40430-019-1683-1>
 - [29] Abbasi, M., Bagheri, B., Keivani, R. (2015). Thermal analysis of friction stir welding process and investigation into affective parameters using simulation. *Journal of Mechanical Science and Technology*, 29(2): 861-866. <https://doi.org/10.1007/s12206-015-0149-3>
 - [30] Vigneshkumar, M., Padmanaban, G., Balasubramanian, V. (2019). Influence of tool tilt angle on the formation of friction stir processing zone in cast magnesium alloy ZK60/SiCp surface composites. *Metallography, Microstructure, and Analysis*, 8(1): 58-66. <https://doi.org/10.1007/s13632-018-0507-5>
 - [31] Zhao, Y.Q., Liu, H.J., Chen, S.X., Lin, Z., Hou, J.C. (2014). Effects of sleeve plunge depth on microstructures and mechanical properties of friction spot welded alclad 7B04-T74 aluminum alloy. *Materials & Design* (1980-2015), 62, 40-46. <https://doi.org/10.1016/j.matdes.2014.05.012>

- [32] Chaudhary, A., Dev, A.K., Goel, A., Butola, R., Ranganath, M.S. (2018). The mechanical properties of different alloys in friction stir processing: a review. *Materials Today: Proceedings*, 5(2): 5553-5562. <https://doi.org/10.1016/j.matpr.2017.12.146>
- [33] Gope, P.C., Kumar, H., Purohit, H., Dayal, M. (2019). S–N curves for fatigue life estimation of friction stir welded 19501 aluminum alloy T-joint. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 233(2): 664-674. <https://doi.org/10.1177/0954406218760056>
- [34] Su, J.Q., Nelson, T.W., Sterling, C.J. (2005). Microstructure evolution during FSW/FSP of high strength aluminum alloys. *Materials Science and Engineering: A*, 405(1-2): 277-286. <https://doi.org/10.1016/j.msea.2005.06.009>
- [35] Butola, R., Ranganath, M.S., Murtaza, Q. (2019). Fabrication and optimization of AA7075 matrix surface composites using Taguchi technique via friction stir processing (FSP). *Engineering Research Express*, 1(2): 025015.
- [36] Ma, Z.Y., Sharma, S.R., Mishra, R.S. (2006). Effect of friction stir processing on the microstructure of cast A356 aluminum. *Materials Science and Engineering: A*, 433(1-2): 269-278. <https://doi.org/10.1016/j.msea.2006.06.099>
- [37] Bauri, R., Yadav, D., Suhas, G. (2011). Effect of friction stir processing (FSP) on microstructure and properties of Al–TiC in situ composite. *Materials Science and Engineering: A*, 528(13-14): 4732-4739. <https://doi.org/10.1016/j.msea.2011.02.085>
- [38] Sharma, V., Prakash, U., Kumar, B.M. (2015). Surface composites by friction stir processing: A review. *Journal of Materials Processing Technology*, 224: 117-134. <https://doi.org/10.1016/j.jmatprotec.2015.04.019>
- [39] Azizieh, M., Mazaheri, M., Balak, Z., Kafashan, H., Kim, H.S. (2018). Fabrication of Mg/Al12Mg17 in-situ surface nanocomposite via friction stir processing. *Materials Science and Engineering: A*, 712: 655-662. <https://doi.org/10.1016/j.msea.2017.12.030>
- [40] Zahmatkesh, B., Enayati, M.H. (2010). A novel approach for development of surface nanocomposite by friction stir processing. *Materials Science and Engineering: A*, 527(24-25): 6734-6740. <https://doi.org/10.1016/j.msea.2010.07.024>
- [41] Sharma, A., Sharma, V.M., Mewar, S., Pal, S.K., Paul, J. (2018). Friction stir processing of Al6061-SiC-graphite hybrid surface composites. *Materials and Manufacturing Processes*, 33(7): 795-804. <https://doi.org/10.1080/10426914.2017.1401726>
- [42] Singh, S., Pal, K. (2017). Influence of surface morphology and UFG on damping and mechanical properties of composite reinforced with spinel MgAl2O4-SiC core-shell microcomposites. *Materials Characterization*, 123: 244-255. <https://doi.org/10.1016/j.matchar.2016.11.042>
- [43] Butola, R., Tyagi, L., Kem, L., Ranganath, M.S., Murtaza, Q. (2020). Mechanical and wear properties of aluminium alloy composites: A review. *Manufacturing Engineering*, 369-391. https://doi.org/10.1007/978-981-15-4619-8_28
- [44] Butola, R., Murtaza, Q., Singari, R.M. (2020). Formation of self-assembled monolayer and characterization of AA7075-T6/B4C nano-ceramic surface composite using friction stir processing. *Surface Topography: Metrology and Properties*, 8(2): 025030.
- [45] Tyagi, L., Butola, R., Jha, A.K. (2020). Mechanical and tribological properties of AA7075-T6 metal matrix composite reinforced with ceramic particles and aloevera ash via Friction stir processing. *Materials Research Express*, 7(6): 066526.
- [46] Butola, R., Murtaza, Q., Singari, R.M. (2019). CNC turning and simulation of residual stress measurement on H13 tool steel. In *Advances in Computational Methods in Manufacturing*, Springer, Singapore, 337-348.
- [47] Vaira Vignesh, R., Padmanaban, R., Datta, M. (2018). Influence of FSP on the microstructure, microhardness, intergranular corrosion susceptibility and wear resistance of AA5083 alloy. *Tribology-Materials, Surfaces & Interfaces*, 12(3): 157-169. <https://doi.org/10.1080/17515831.2018.1483295>
- [48] Yazdipour, A., Dehghani, K. (2009). Modeling the microstructural evolution and effect of cooling rate on the nanograins formed during the friction stir processing of Al5083. *Materials Science and Engineering: A*, 527(1-2): 192-197. <https://doi.org/10.1016/j.msea.2009.08.040>
- [49] Chen, Y., Ding, H., Li, J., Cai, Z., Zhao, J., Yang, W. (2016). Influence of multi-pass friction stir processing on the microstructure and mechanical properties of Al-5083 alloy. *Materials Science and Engineering: A*, 650: 281-289. <https://doi.org/10.1016/j.msea.2015.10.057>
- [50] Bagheri, B., Abbasi, M. (2019). Analysis of microstructure and mechanical properties of friction stir vibration welded (FSVW) 5083 aluminum alloy joints: Experimental and simulation. *Journal of Welding and Joining*, 37(3): 243-253. <https://doi.org/10.5781/JWJ.2019.37.3.8>
- [51] Bagheri, B., Rizi, A.A.M., Abbasi, M., Givi, M. (2019). Friction stir spot vibration welding: improving the microstructure and mechanical properties of Al5083 joint. *Metallography, Microstructure, and Analysis*, 8(5): 713-725. <https://doi.org/10.1007/s13632-019-00563-y>
- [52] Meyghani, B., Awang, M.B., Emamian, S.S., Mohd Nor, M.K.B., Pedapati, S.R. (2017). A comparison of different finite element methods in the thermal analysis of friction stir welding (FSW). *Metals*, 7(10): 450. <https://doi.org/10.3390/met7100450>
- [53] Asadi, P., Mahdavinnejad, R.A., Tutunchilar, S. (2011). Simulation and experimental investigation of FSP of AZ91 magnesium alloy. *Materials Science and Engineering: A*, 528(21): 6469-6477. <https://doi.org/10.1016/j.msea.2011.05.035>

Evaluation of Moth-Flame Optimization, Genetic and Simulated Annealing tuned PID controller for Steering Control of Autonomous Underwater Vehicle

Sudarshan K. Valluru

Center for Control of Dynamical Systems
and Computation

Dept. of Electrical Engineering
Delhi Technological University
Delhi-110042, India
sudarshan_valluru@dce.ac.in

Karan Sehgal

Center for Control of Dynamical Systems
and Computation

Dept. of Electrical Engineering
Delhi Technological University
Delhi-110042, India
karansehgal_2k18ee089@dtu.ac.in

Hitesh Thareja

Center for Control of Dynamical Systems
and Computation

Dept. of Electrical Engineering
Delhi Technological University
Delhi-110042, India
hiteshthareja_2k18ee084@dtu.ac.in

Abstract—This paper describes an optimal bio-inspired PID controller for accurate steering management of the Autonomous Underwater Vehicle system (AUV). To achieve precise control performance, a PID controller is designed, and its gain parameters K_p , K_i , K_d are tuned by applying Simulated Annealing (SA), Genetic Algorithm (GA) and Moth-Flame Optimization Algorithm (MFO). The experimental response corresponding to the unit step and square input waveform for these proposed nature-inspired optimization algorithms were obtained. The response characteristics like overshoot, rise time, settling time and performances index ITAE were calculated and compared. The experimental results show that MFO-PID is highly efficient, followed by GA and SA, respectively.

Keywords—Moth-flame, GA, SA, AUV, PID, Optimization

I. INTRODUCTION

AUV, which stands for an autonomous underwater vehicle, can perform several operations in shallow and deep-sea environments. They have been successfully applied in various fields including military operations, commercial and research purposes etc. Fitted with electronic subsystems, they allow the robot to steer efficiently in harsh surroundings while undergoing the assigned tasks without any human input. This coherent nature of the AUV is due to its six degrees of freedom (DoF). The device's robust interconnection is shown in Fig.1. and the symbolic notations of position and velocity terms of AUV [1] are shown in Table. I. Despite these adroit networks of subsystems, due to the presence of natural and environmental disturbances such as tidal waves and ocean currents, etc., control of such vehicles becomes an arduous task.

For achieving a more potent control, researchers have utilised several intelligent control methods for AUV control [2], [3]. However, controllers like PD, PID have proved to showcase a more straightforward controlling approach. Moreover, such schemes possess more uncomplicated application in implementation from the linear regime's computational point of view [4]. In contrast, however, PID controllers are suffering from tedious computations during the changes in system parameters because of the occurrence of natural perturbations. Many papers are also available in the literature that erudite a controller's application and design, synthesized using a PID control process for the robust steering control of the AUV system [5], [6]. Such controllers are also in demand to control multiple other systems like the trajectory tracking and control of TRMS and Ball beam systems [7], [8].

This paper implements a PID based controller, which is optimally tuned by a bio-inspired meta-heuristic optimization approach referred to as the Moth-Flame Optimization (MFO), Genetic Algorithm (GA) and Simulated Annealing (SA) for the robust control of an AUV. It is found that the performance MFO-PID is better as compared to GA-PID and SA-PID.

This paper is arranged as mathematical modelling of AUV system is explained in section II, followed by section III & IV, which include synthesis of PID based controller applying MFO. Section V gives the experimental responses of the AUV. Eventually, section VI provides the conclusion.

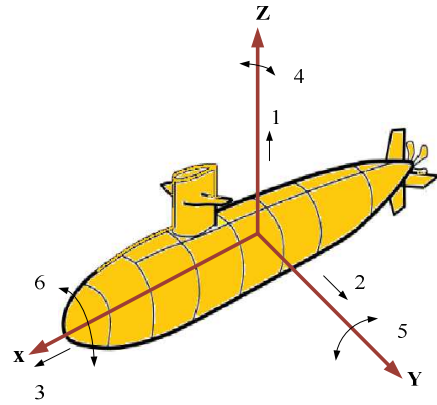


Fig1. Six Degrees of Freedom of an AUV

TABLE I. NOTATIONS FOR AUV MOTION

Direction Of Motion	Moment And Force	Earth-Fixed Frame (Position)	Body-Fixed Frame(velocity)
Surge (Motion along $X - axis$)	X	x	u
Sway Motion along $Y - axis$	Y	y	v
Heave (Motion along $Z - axis$)	Z	z	w
Roll (Rotation along $X - axis$)	K	ϕ	p
Pitch (Rotation along $Y - axis$)	M	θ	q
Yaw (Rotation along $Z - axis$)	N	ψ	r

II. AUV MODELLING

To formulate a PID controller for the steering management of an autonomous underwater system, we would first require its general transfer function which gives information regarding the yaw angle, and its relation with the deflection parameter. For this, the mathematical modelling of an AUV [9] is done. This is carried out by considering two different frames of reference, namely earth-fixed and body-fixed frame. System coordinates of this prototype are expounded using three mutually perpendicular axes starting from a random point. North and East correspond to x and y -axis, respectively. Increasing depth conform with the z -axis. The position vector η and velocity vector v can be described by the equations (1) and (2) respectively.

$$v = [u \ v \ w \ p \ q \ r]^T \quad (1)$$

$$\eta = [x \ y \ z \ \phi \ \theta \ \psi]^T \quad (2)$$

In pure steering plane, simplification of these equations is carried out by considering the origin of the body-fixed frame[10] to concur to the center of gravity:

$$m(\dot{v} + u_0 r) = \sum Y \quad (3)$$

$$I_z \dot{r} = \sum N \quad (4)$$

Surge speed (u_0), is fixed at 0.75 m/s. Assuming the pitch angle and roll to be small:

$$\Psi = \frac{\sin \phi}{\cos \theta} q + \frac{\cos \phi}{\cos \theta} r \approx r \quad (5)$$

In matrix form, the equations (1) to (5) are rewritten as (6):

$$\begin{bmatrix} m - Y_{\dot{v}} & -Y_r & 0 \\ -N_{\dot{v}} & I_{zz} - N_r & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \dot{v} \\ \dot{r} \\ \dot{\Psi} \end{bmatrix} + \begin{bmatrix} -Y_v & -Y_r + m v_0 & 0 \\ -N_v & -N_r & 0 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} v \\ r \\ \Psi \end{bmatrix} = \begin{bmatrix} Y_{\delta i} \\ N_{\delta i} \\ 0 \end{bmatrix} \delta_r \quad (6)$$

By substituting the values of vehicle parameter of AUV[11], [12] dimensions, the hydrodynamic coefficient for u_0 at 0.75 m/s and by applying the state space approach to get equations (7) to (9)

$$\dot{x}(t) = Cx(t) + Du(t) \quad (7)$$

$$C = \begin{bmatrix} -0.114 & -0.2647 & 0 \\ 0.0225 & -0.2331 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad (8)$$

$$D = [0.0211 \ -0.0258 \ 0]^T$$

$$U = \delta_r$$

The relation between yaw (ψ) and rudder deflection (δ_r) in terms of the transfer function is acquired as:

$$\frac{\Psi(s)}{\delta_r(s)} = \frac{-0.0258s - 0.0024}{s^3 + 0.3445s^2 + 0.319s} \quad (9)$$

Now that the system's requisite transfer function has been derived, we can synthesize the proposed compensator.

III. DESIGN OF PID CONTROLLER FOR AUV

PID controller, is a conventional control scheme, is applied extensively for precise control of various systems. Controllers based on PID have been successfully implemented in the design and control of multiple systems[13]–[15]. The PID displays sufficient stability margins, optimum time responses, better system characteristic properties such as low overshoot, and lesser settling time. It consists of proportional, integral and derivative gain parameters which are functions of error between the desired set point and actual system output. The general unity feedback characteristic equation of AUV with PID control laws are written as equations (10), (11) and (12).

$$c(t) = k_p e(t) + k_i D^{-1} e(t) + k_d D e(t) \quad (10)$$

$$U(s) = \frac{c(s)}{e(s)} = k_p + k_i s^{-1} + k_d s \quad (11)$$

$$1 + G(s)U(s) = 0 \quad (12)$$

Now for finding the accurate values of the operational gains (k_p, k_i, k_d), the characteristic equation is optimized based on the Integral Time Absolute Error (ITAE) performance index, i.e., integral of time multiplied by absolute error to minimize the error signal.

$$ITAE = \int_0^t t |e(t)| dt \quad (13)$$

PID controller gains are then tuned using MFO, GA and SA to observe their performances for comparison. The block diagram of PID Controller based optimization of AUV using MFO/GA/SA by taking ITAE as the cost function is shown in Fig.2.

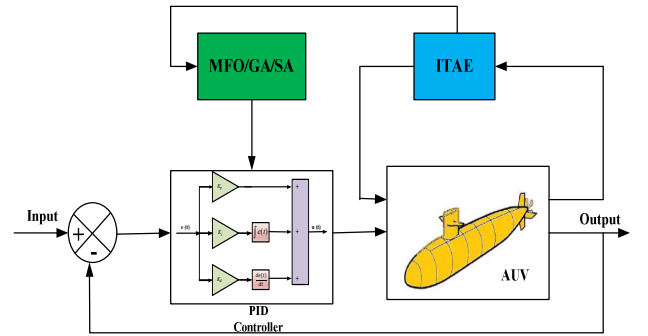


Fig.2. Block Diagram of unity feedback optimal PID control

IV. TUNING OF PID CONTROLLER USING SA, GA AND MFO

PID controllers are commonly used in all dynamical systems, but it requires a monotonous tuning of control actions to avoid sluggishness of the system response. Nature and bio-inspired optimization algorithms can diminish the computational difficulties in the monotonous tuning of PID controllers, thereby the AUV's steering control is with minimum human interventions. Here, the PID controller in AUV's steering loop is tuned by Simulated Annealing, Genetic Algorithm and Moths Flame Optimizer methods.

A. Simulated Annealing Tuned PID Controller

Simulated Annealing is one of the most widely used methods for optimizing control problems of dynamical systems. This algorithmic technique is inspired by the relationship of combinational optimization and quantum and classical or statistical mechanics laws. A simple mechanism of cooling of material is applied in steps till the lowest energy is reached. The state at this lowest energy is the optimized state. Pseudocode for the SA algorithm applied for tuning the gain parameters of the proposed PID controller is given as:

Step 1: The ranges for the 3 gain coefficients of a PID control are fixed in the form of an objective function $f(x) = [K_p, K_i, K_d]^t$
Step 2: Set t_0 as the temperature at the beginning of the process
Step 3: Take s_i as the initial stage of the system
Step 4: Take s_{opt} as the desired optimized state of the system
Step 5: Initially, assign t as t_0 and s_{opt} as s_i
Step 6: **for** $t = 1$ to t_{max} **do**
 Assign s_{opt+1} to adjacent/nearest (s_{opt})
 Change ΔE to $(f(s_{opt+1}) - f(s_{opt}))$
 if $\min(1, e^{-\Delta E/t}) \geq \text{random}(0,1)$ **then**
 Update s_{opt} to s_{opt+1}
 end if
 Assign t as the temperature-schedule(t)
end for
Step 7: Output the final solution of the optimized function

The values for the 3 gain coefficients K_p , K_i and K_d obtained using SA method are -19.9298, -4.9419 and -41.25 respectively. However, the simulated annealing technique poses a critical drawback while working on minimization problems. For instance, any change in the system values that decrease the cost function 'f' will be accepted as desired, but sometimes, changes that increase 'f' might also get counted. This happens with a probability p , known as the transition probability. Due to this disadvantage offered by SA, we use GA to stabilize the system.

B. Genetic Algorithm Tuned PID Controller

Genetic algorithm (GA) is a nature-based optimization technique applied to solve computational problems. It is used on a chromosome population where every chromosome represents a solution that has an associated fitness value to it. This value defines how optimal a solution is. Some arbitrarily generated population is initially taken, followed by the selection process, which is fitness based. The next step involves recombination to develop the next generation. For the above step, parent genes are used to obtain child chromosomes. Several iterative processes continue till the stopping criteria is achieved. Pseudocode for the genetic algorithm is:

Step 1: Designate the ranges for the control parameters of the PID controller to model an objective function as- $f(x) = [K_p, K_i, K_d]^t$

Step 2: Generation of initial source population of M chromosomes randomly.

Step 3: Calculation of the fitness F using Eq. (14)

Step 4: **for** $i = 1:n$

 Take two chromosomes from the current population.

 Crossover is applied to chromosomes with crossover rate x .

 Mutation is applied to the chromosome with mutation rate m to generate a new chromosome.

 Add the above generated chromosome to next generation population.

 Current Population is replaced with next generation population.

end for

Step 5: Finally, output the solution of the optimized global best.

One practical step for applying GA is to evaluate the fitness value for all the chromosomes to reduce the error signal $e(s)$. PID controller is applied to minimize this error. As this fitness value is inversely proportional to the value of the performance index, we can define the chromosomes' fitness as (14).

$$F = \frac{1}{\text{Performance Index}} \quad (14)$$

The tuned optimized parameters K_p , K_i and K_d , using GA are -17.4092, -1.8497 and -38.4368, respectively. The response obtained using GA shows an overshoot of more than 10%. This result also shows that GA does not have a high speed of convergence that is why an alternative algorithm, designed using evolutionary strategies can be deployed to obtain a faster and more efficient performance of the system. The moth flame optimization algorithm (MFO) is one such technique.

C. Moths Flame Optimization Algorithm tuned PID

MFO is a newly developed nature-based solution finding mechanism made from the algorithms inspired by the population search strategy. This bio-inspired optimization algorithm is very flexible, and can easily be implemented in finding the optimal solution of real-world problems. Some applications of MFO are available in literature ranging from the tuning of controllers such as fuzzy-PID, fuzzy-PI, PID, PI. Moth-flame optimization (MFO) technique was introduced by Mirjalili[16]. It initiates by creating moths arbitrarily inside the solution region by a transverse orientation shown in Fig.3.

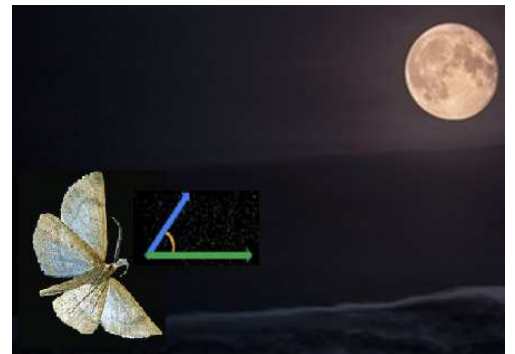


Fig.3 Moth's Transverse Orientation

After this, the fitness function values for every moth are calculated and tagged the most optimum position by flame. Then, depending on the spiral movement function, the moths' positions are modified to attain more acceptable positions sorted by a flame, the new most acceptable positions of individuals are upgraded, and replicating the previous operations.

This process takes place until the total number of iterations have been completed. The MFO algorithm has three main postulates. These are given below:

1. Initializing Population:

It is taken that all moths can fly in all the dimensions. The moth set can be represented as:

$$B = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & \cdots & b_{1,d} \\ b_{2,1} & \cdots & \cdots & \cdots & b_{2,d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{x,1} & b_{x,2} & \cdots & \cdots & b_{x,d} \end{bmatrix} \quad (15)$$

Where x stands for the number of moths and d gives the number of dimensions in the solution region.

Array to store the values of the fitness function is as follows:

$$OB = \begin{bmatrix} OB_1 \\ OB_2 \\ \vdots \\ OB_x \end{bmatrix} \quad (16)$$

The given matrix defines the flames in the d-dimensional region, and their fitness value vector follows it.

$$F = \begin{bmatrix} F_{1,1} & F_{1,2} & \cdots & \cdots & F_{1,d} \\ F_{2,1} & \cdots & \cdots & \cdots & F_{2,d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ F_{x,1} & F_{x,2} & \cdots & \cdots & F_{x,d} \end{bmatrix} \quad (17)$$

$$OF = \begin{bmatrix} OF_1 \\ OF_2 \\ \vdots \\ OF_x \end{bmatrix} \quad (18)$$

The solutions are moths and flames. What distinguishes them is how we treat and update them after every iterative step. Actual searching agents which move throughout the search area are moths. Flames are optimal positions of moths which have been derived until now.

2. Updating Moths' Position:

The optimal global value can be obtained for the optimization problem; this algorithm employs three steps. These are given below:

$$MFO = (I, P, T) \quad (19)$$

where I describe the function, which gives the first moth population randomly

$$I: \phi \rightarrow \{B, OB\}$$

P signifies the moths' movement in the search area

$$P: B \rightarrow B$$

T is the condition for termination

$$T: B \rightarrow \text{true}, \text{false}$$

The equation below, explains the function I , which applies random distribution.

$$M(p, q) = (\text{ub}(p) - \text{lb}(q)) * \text{rand}() + \text{lb}(p) \quad (20)$$

Where lb and ub represent the lower and the upper bounds, respectively.

Three conditions which should be followed while applying a logarithmic spiral are:

- The starting point of the spiral should begin from the moth.
- The ending point of the spiral should be in the flame position.
- Spiral's range fluctuation is not supposed to surpass the search space.

$$S(B_p, F_q) = D_p \cdot e^{bt} \cdot \cos(2\pi t) + F_q \quad (21)$$

Where D_p represents the region within p^{th} moth and q^{th} flame.

$$D_p = |F_q - B_p| \quad (22)$$

B_p describes the logarithmic spiral's shape, and t stands for any arbitrary value within $[r, 1]$. The spiral motion guarantees the balance between exploitation and exploration close to the flame. Where r changes from $[-1, 2]$ in the whole process of iteration, which is called as the convergence constant. The logarithmic spiral shape, as described above, is shown in Fig.4.

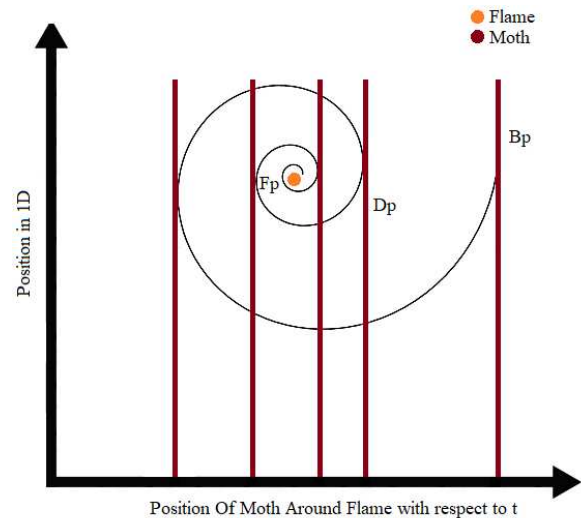


Fig.4. Logarithmic spiral shape

3. Mistakes Update in Flames

Moth positions are updated in n different locations inside the search region, which might reduce the exploitation of the most optimized solutions. Thus, minimizing the flames solves the conflict using an equation given below:

$$\text{flame no.} = \text{round} \left(N - l * \frac{N-1}{T} \right) \quad (23)$$

N represents the total flames

l represents the present iteration

T represents the total iterations

Pseudocode for the MFO algorithm is described as:

Step 1: Assign the fixed ranges for each control parameter of the proposed PID controller in the form of a cost function as

$$f(x) = [K_p, K_i, K_d]^t$$

Step 2: Initialize the Moth-Flame population

Step 3: Initialize position of moth B arbitrarily

Step 4: **for** $p = 1$ **to** n **do**

 Calculate the value of fitness function F

end for

Step 5: **While** iterations \leq total iterations **do**

 Update flame no by using Eq. (23)

$OB = \text{Fitness Func}(B)$;

if iteration == 1

$F = \text{sort}(B)$;

$OF = \text{sort}(OB)$;

else

$F = \text{sort}(B_{t-1}, B_t)$;

$OF = \text{sort}(B_{t-1}, B_t)$;

end

Step 6: **for** $p = 1:n$

for $q = 1:d$

 Update the values of r and t

 Calculate the value D using Eq. (22)

 Update $S(p, q)$ using Eq. (21)

end

end

Step 7: Consequently, the precisely tuned solution of the optimized cost function is output.

The tuned optimally calculated parameters K_p , K_i and K_d , by using MFO are -15.5343, -0.025 and -38.93, respectively.

V. SIMULATION AND RESULTS

For comparison of the optimization performances of Moth-Flame Optimization (MFO), Genetic Algorithm (GA) and Simulated Annealing (SA) method, the output response of AUV, controlled by tuned PID controllers are observed for unit step and unit square wave input. In the square wave input case, the chosen frequency for simulation was fixed at 15 mHz.

The tuned PID controller's step and square wave responses for MFO-PID, GA-PID and SA-PID controllers for AUV steering control are shown in Fig.5. and Fig.6 respectively.

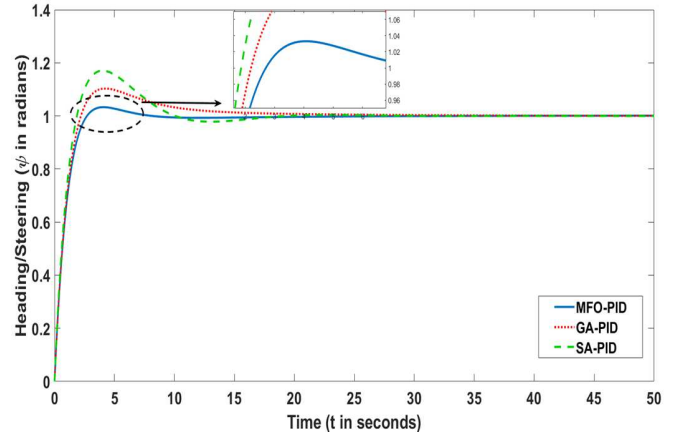


Fig.5. Unit Step Responses by the respective tuned controllers

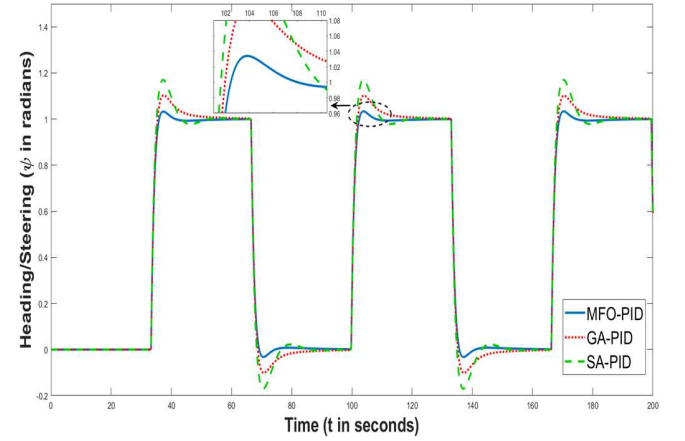


Fig.6. Square Responses by the respective tuned controllers

The precisely optimized solutions of the gain parameters are given in Table II.

TABLE II. GAIN PARAMETERS OF TUNED PID CONTROLLER

Gain Parameters			
PID Parameters	SA	GA	MFO
K_p	-19.9298	-17.4092	-15.5343
K_i	-4.9419	-1.8497	-0.025
K_d	-41.25	-38.4368	-38.93

The PID transient performance characteristics using MFO, GA and SA are given in Table III.

TABLE III. TRANSIENT PERFORMANCE OF TUNED PID CONTROLLER

ITAE			
Tuning Algorithm	SA	GA	MFO
Rise Time (Sec)	1.37	1.56	1.72
%Overshoot	17.1	10.3	3.3
Settling Time (Sec)	14.1	12.4	5.7

The cost comparison amongst MFO, GA and SA are given in Table IV.

TABLE IV. COST COMPARISON

ITAE ERROR			
Tuning Algorithm	SA	GA	MFO
Error(J)	0.6396	0.4371	0.2066

It is observed from Table III that the MFO tuned PID controller has remarkable performance as compared to GA tuned PID and SA tuned PID in terms of lesser overshoot and settling time. It proves that the MFO tuned PID for AUV steering control is optimally and accurately meeting the design objectives. It is also noted from Table IV that the cost and error are drastically reduced in contrast to GA tuned PID and SA tuned PID controllers.

VI. CONCLUSION

To obtain the robust control for an autonomous underwater vehicle, this paper presents the application of a bio-inspired optimization algorithm called Moth-Flame Optimization to optimally tune the gain parameters of a PID based controller for efficient motion stabilization of the AUV system. Tuning of the proposed PID controller is done for an error-based performance index ITAE. Genetic Algorithm and Simulated Annealing Method are also used to compare the system's step and square responses. The response obtained for MFO-PID is clearly, better than GA-PID, followed by SA-PID in case of overshoot, settling time and rise time. The scope of future work can be the application of this algorithm to design more complex controllers for advanced systems.

REFERENCES

- [1] M. A. Abkowitz, *Stability and Motion Control of Ocean Vehicles*. The MIT Press, 1969.
- [2] J. Lorentz *et al.*, "A fuzzy rule-based algorithm to train perceptrons," *IEEE J. Ocean. Eng.*, vol. 19, no. 4, pp. 359–367, Nov. 2020, doi: 10.1016/S0165-0114(03)00242-2.
- [3] A. J. Healey and D. Lienard, "Multivariable Sliding-Mode Control for Autonomous Driving and Steering of Unmanned Underwater Vehicles," *IEEE J. Ocean. Eng.*, vol. 18, no. 3, pp. 327–339, 1993, doi: 10.1109/JOE.1993.236372.
- [4] Astrom. K.J. and Hagglund.T, *PID controllers: theory design and tuning*. North Carolina, USA: Instrument Society of America, Research Triangle Park, 1995.
- [5] S. P. Hou and C. C. Cheah, "Can a simple control scheme work for a formation control of multiple autonomous underwater vehicles?," *IEEE Trans. Control Syst. Technol.*, vol. 19, no. 5, pp. 1090–1101, 2011, doi: 10.1109/TCST.2010.2076388.
- [6] F. Kong, Y. Guo, and W. Lyu, "Dynamics Modeling and Motion Control of an New Unmanned Underwater Vehicle," *IEEE Access*, vol. 8, pp. 30119–30126, 2020, doi: 10.1109/ACCESS.2020.2972336.
- [7] S. . Valluru, M. Singh, Ayush, and A. Dharavath, "Design and Experimental Implementation of Multi-loop LQR, PID, and LQG Controllers for the Trajectory Tracking Control of Twin Rotor MIMO System.," in *Intelligent Communication, Control and Devices.*, K. A. Choudhury S., Mishra R., Mishra R., Ed. Springer Nature Singapore Pte Ltd, 2019, pp. 599–608.
- [8] S. K. Valluru, M. Singh, and S. Singh, "Prototype Design and Analysis of Controllers for One Dimensional Ball and Beam System," in *1st IEEE International Conference on Power Electronics. Intelligent Control and Energy Systems*, 2016, pp.1–6.
- [9] K. P. Valavanis, D. Gracanin, M. Matijasevic, R. Kolluru, and G. A. Demetriou, "Control Architectures for Autonomous Underwater Vehicles," *IEEE Control Systems Magazine*, vol. 17, no. 6, pp. 48–64, 1997.
- [10] J. A. Monroy, E. Campos, and J. A. Torres, "Attitude control of a Micro AUV through an embedded system," *IEEE Lat. Am. Trans.*, vol. 15, no. 4, pp. 603–612, 2017, doi: 10.1109/TLA.2017.7896344.
- [11] B.Jalving, "The NDRE-AUV Flight Control System," *IEEE J. Ocean. Eng.*, vol. 19, no. 4, pp. 497–501, 2007.
- [12] A.Healey and Marco D.B, "Experimental verification of mission planning by autonomous mission execution and data visualization using the NPS AUV II," in *IEEE Symposium on Autonomous Underwater Vehicle Technology*, 1992, pp. 65–72, doi: 10.1109/AUV.1992.225193.
- [13] Sudarshan. K.Valluru, R. Kumar, and R. Kumar, "Design and Implementation of L-PID and IO-PID Controllers for Twin Rotor MIMO System," in *IEEE International Conference on Power Electronics, Control and Automation (ICPECA)*, 2019, pp. 1–5, doi: 10.1109/icpeca47973.2019.8975542.
- [14] S. K.Valluru and M. Singh, "Performance investigations of APSO tuned linear and nonlinear PID controllers for a nonlinear dynamical system," *J. Electr. Syst. Inf. Technol.*, vol. 5, no. 3, pp. 442–452, 2018, doi: 10.1016/j.jesit.2018.02.001.
- [15] S. K. Valluru, M. Singh, M. Singh, and V. Khattar, "Experimental Validation of PID and LQR Control Techniques for Stabilization of Cart Inverted Pendulum System," in *3rd IEEE International conference on Recent Trends in Electronics, Information and Communication Technology*, 2018, pp. 708–712.
- [16] S. Mirjalili, "Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm," *Knowledge-Based Syst.*, vol. 89, pp. 228–249, 2015, doi: 10.1016/j.knosys.2015.07.006.

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/351065515>

Article ID: IJARET_12_04_025 Cite this Article: Kshitij Tripathi, Kunal Kukreja and AK Madan, Evolution in Manufacturing of Grid Stiffened Structures through CAM and Additive Techn...

Article in INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN ENGINEERING & TECHNOLOGY · April 2021

DOI: 10.34218/IJARET.12.4.2021.025

CITATIONS

0

READS

234

3 authors, including:



Kshitij Tripathi

Delhi Technological University

1 PUBLICATION 0 CITATIONS

SEE PROFILE



Kunal Kukreja

Delhi Technological University

1 PUBLICATION 0 CITATIONS

SEE PROFILE



EVOLUTION IN MANUFACTURING OF GRID STIFFENED STRUCTURES THROUGH CAM AND ADDITIVE TECHNIQUES

Kshitij Tripathi, Kunal Kukreja* and Dr. AK Madan

Professor, Department of Mechanical Engineering, Delhi Technological University,
New Delhi, India

*Corresponding Author

ABSTRACT

In this review paper, a comparison is drawn among different manufacturing grid stiffened structures, i.e., isogrid panels manufactured using non-conventional manufacturing techniques such as Fused Deposition Modeling (FDM) over existing conventional methods such as milling; by summarizing existing success analyses. These Isogrids/Grid stiffened structures are primarily building blocks used in aerospace applications such as rocket shells, satellites, space station walls, and other structures requiring additional strengthening such as armour shells and wall structures mega buildings such as stadiums, domes, or even spider webs. A comprehensive analysis is made for various aspects of these advantages and the role of non-conventional techniques is discussed. Moreover, a hypothesis is put forward regarding design optimization and evolution in the future. The isogrids are an upcoming research area that has the capacity to strengthen the aerospace, nuclear and other allied sectors.

Key words: Isogrid, Grid stiffened structures, Mass ratio, Aerospace, Composite Structures, Additive Manufacturing, Rocket Shell.

Cite this Article: Kshitij Tripathi, Kunal Kukreja and AK Madan, Evolution in Manufacturing of Grid Stiffened Structures through CAM and Additive Techniques, *International Journal of Advanced Research in Engineering and Technology (IJARET)*, 12(4), 2021, pp. 217-225.

<http://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=12&IType=4>

1. INTRODUCTION

One of the industries which have gone under tremendous change since inception is the aerospace industry. The aerospace industry [8] is aiming to make effective aircraft, ranging from the passenger aeroplane to rockets aiming for space exploration. Apart from advancements in propulsion systems, the various manufacturers are now addressing a very crucial parameter called the mass ratio, which is simply the ratio of fuel mass to dry mass of the rocket. The higher the mass ratio, the more propellant is required or the dry mass of the rocket should be reduced.

A higher mass ratio is required as it results in a higher delta-v (a measure of the impulse that is needed to perform a manoeuvre), which results in ease in the manoeuvring of the rocket.

To achieve this, the reduction in the dry mass of the rocket is essential as there is a limit to increase the fuel mass. Dry mass reduction can be done by either using lightweight materials or optimizing the geometric design of the structure or even both. In the case of lightweight materials, metals like Aluminium and its alloys are being used. However, the usage of Aluminium and its alloys poses some challenges. The other way is to optimize the geometric design; rocket shells are manufactured either as an isogrid or an orthogrid.

2. THE IMPETUS FOR DESIGN EVOLUTION

It is important to observe that any changes in mass ratios in these aerospace applications will lead to significant financial implications on the operation of any such undertaken project

Any scenarios that might expose the structure to impact loading. It is also important to analyse explicit dynamic behaviour for these structures in the case of any mishappenings. A major scope of development is the multi-layer, multi-metal deposition of the grid stiffeners.

Due to recent strides in the aerospace industry, especially with companies like SpaceX and blue origin, which have introduced the re-usage of rockets for bringing payloads to space. A problem statement has been identified which is fourfold, namely: -

- Assembly time increased in order to accommodate attachment points for payloads; even though they don't affect overall force by a large magnitude, they increase weight and could be incorporated to the shell itself
- After the payload/shuttle has crossed the atmosphere and reached orbit, it is exposed to harsh environments and unbalanced forces which are always delivered as impulse, design adjustments could be made to minimize deformations from these impact loads and explicit dynamic behaviour should be analysed
- Due to the advent of the reusability of the shuttle, rockets and transport vehicles, the financial structure gets a major shake up resulting in redistribution of resources towards more costly, durable, lighter and rarer materials which weren't used previously and a paradigm shift is shown in research towards material and process evolution, resulting in a change in the entirety of the manufacturing line, including process parameters, moulds and tooling.
- Due to the ever changing needs of each transport mission, each rocket needs to be versatile enough to carry a varying amount of unbalanced loads and should be resistant to fatigue and should provide ample thermal resistance as well
- Make rockets with panels so that pieces can be interchanged

3. ISOGRID ARCHITECTURE

An isogrid [9] is a lattice of intersecting ribs forming an array of equilateral triangles. Their manufacturing is based on the sandwich theory, which describes the behaviour of a beam, plate or shell consisting of three layers - 2 face sheets and one core. Traditionally an isogrid structure is machined from a single piece of aluminium stock and consists of a skin with stiffness in a grid-like structure that forms equilateral triangles. Now these equilateral triangles give rise to an isotropic behaviour, leading to their iso-prefix.

The major advantage is that these structures show high stiffness and stability, especially when built from composite materials.



Figure 1 A typical Isogrid Structure [16]

The above figure 1 is showing a typical Isogrid. It consists of conical sub-structures. Isogrids are used wherever there is a need to increase the stiffness of thin-walled structures along with the reduction of weight. We see the use of isogrids in gas turbines engine casings, where thin-engine walls are needed to be reinforced for additional stiffness. It is seen that the hierarchical configuration of isogrids can make these structures more efficient in load-bearing capacity.

4. MANUFACTURING TECHNIQUES

Isogrid structures have been manufactured using different techniques. However, the type of manufacturing technique is heavily dependent on the material being used for the structure itself.

The most common and heavily used material is aluminium, used for the manufacturing of rocket shells, and so on. The most common methods are/were mechanical and chemical milling processes.

In a patent application [11], many different manufacturing techniques were mentioned; one of them being chemical milling machining. One of the methods mentioned is chemical machining. In this, metal removal is achieved by reverse electroplating in which the hydroxide of the metal to be removed is produced; suspended as an emulsion in the electrolytic solution. Another method is using an NC milling machine; which focuses on manufacturing a frusto-conical structure reinforced with isogrid reinforced on its internal structure, formed from a plurality of substantially identical panels. Each panel is manufactured from the metal which will be reinforcing the final structure. The plate to be machined is positioned in an NC milling machine and the triangular pockets are formed. For a good isogrid structure, it is seen during the manufacturing process, the wall thickness of the pocket should be reduced by 1mm. In areas where wall thickness less than 1mm is required, chemical milling is employed. After machining, the panel is rolled or formed into the desired shape and the panels are secured together to form the final structure. The major issue with this method is the use of harmful chemicals to get the desired wall thickness of the pockets, which are hazardous to use. In another method, milling was done twice, first to remove material to the desired pocket wall thickness and second to mill around the periphery of the pocket so formed. The issue with this method is that sufficient wall thickness can't be achieved. Hence in those few pages, it was clear that a method is needed that can bring out the sufficient wall thickness without the use of chemicals.

The above-mentioned processes greatly affected commercial aspects as they are costly, require large amounts of material, and are inefficient. Milling also results in the distortion of the isogrid ribs and this technique is limited to certain isogrid geometries. Hence, a research paper [10] has mentioned the use of abrasive water jet (AWJ) technique, an unconventional method for isogrid fabrication. The potential advantages of this technique are high productivity, no residual stresses, integral machining capability and AWJ is capable of machining a wide

range of isogrid geometries and for a variety of materials. More specifically, the concept of Single-angled jet in a circular tool was utilized for the study because of its versatility and ease of application. Also, before the manufacturing took place, the various parameters of AWJ like Water jet Pressure, Abrasive size, Abrasive material, Water jet flow rate, Stand-off distance etc. were determined for 2 cases; namely, linear cutting using AWJ and milling using the AWJ nozzle.

Since [1], significant progress has been made towards the optimization of isogrid designs and towards improving manufacturing techniques. A major portion of the efforts towards this evolution was put towards the development of composite materials in the 1970s. The use of such materials is suitable for isogrids, due to the stress being heavily distributed along the rib length as the isogrid itself is a composite structure.

Advances have been made and numerous techniques have come out which are suitable for different aspects of design and manufacturing with composite materials. Due research in mould material as well as other parameters.

In different research works, which all were aimed at reducing the drawbacks presented by conventional techniques for composite materials, by taking into focusing on changes mould design to accommodate higher nodal density, change in manufacturing technique to RFW and hybrid tooling and expansion block tooling, to achieve lower costs and higher quality through higher degrees of control of geometry [2] [1] [12].

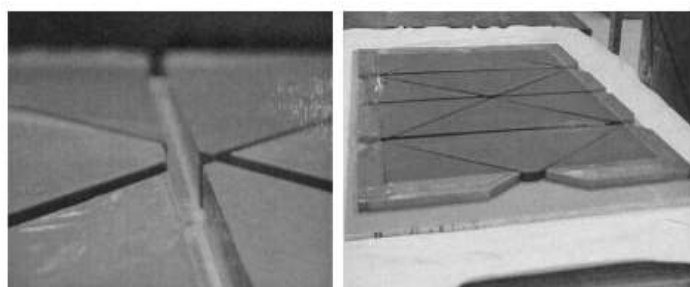


Figure 2 Expansion block tooling [12]

All this evolution has led up to load specific structures called the Advanced Grid Structures focusing on:

- Placement of fibres in a specific pattern or in a specific overlay order for each rib and so on
- Tooling techniques suitable for different geometries
- Mould selection to mitigate thermal expansion and other effects

5. PUSH TOWARDS ADDITIVE MANUFACTURING

A major push has been made toward additive manufacturing techniques due to a decrease in overall material and production costs and high efficiency and accuracy while providing control over parameters like porosity. Also, the ease of prototype manufacturing for testing purposes has made a huge contribution.

In Ming Li et al buckling tests were performed in hierarchical lattices, and tests were performed by 3D printing a plastic model using PVC engineering plastics and showed these structures have greater resistance towards local buckling and have higher global stiffness.

Different methods in additive manufacturing are also considered, like laminated object manufacturing (LOM) or Fused deposition Modeling (FDM) as used [6].

It is seen that there is a big role of 3D printing in the manufacturing of the isogrid structures. In a research work, a lot of focus has been given on the design of hierarchical isogrid lattice panels through additive machining [5]. In that, 3 specimen panels were taken. Each specimen consisted of 2 isogrid layers and the layers' binding together produced lattice structure with T ribs. It was seen that all these structures were 3D printed using the Laminated Object Manufacturing (LOM) technique of 3D printing, which is one of the fastest and most affordable ways to create a 3D prototype [5].

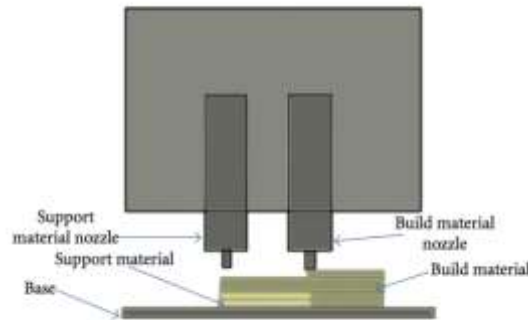


Figure 3 The FDM process [17]

Figure 3 is depicting fusion deposition modelling technique. A lot of focus has been given on the design and characterization of an integrally stiffened structure using additive machining, certain takeaways can be obtained which highlight the role of unconventional machining, especially 3D printing [B] [6]. One major aspect highlighted was that the manufacturing of complex geometries is possible easily in 3D printing without consuming extra time as well as cost. It is also seen that 3D printing eliminates the requirement of assembly as it prints by layered manufacturing, which doesn't require assembly.

6. COMPUTER AIDED MANUFACTURING HURDLES

With the evolution of manufacturing, the algorithms previously running the NC milling and other conventional process machinery were not able to produce complex forms due to the drawbacks of the conventional techniques themselves such as, the workpiece being operated on a uni-directional basis as the work piece itself had enormous cross-sectional area for a 3-axis or a 5-axis milling, or the process itself was operated on a unidirectional basis, such as abrasive waterjet machining (AWJ) as mentioned before. The next problem to be tackled was the programming and the algorithms themselves that weren't advanced or capable enough to rely on learning data sets to optimize path and process parameters. It is well known that the aerospace industry utilizes technology which is way ahead of its time. This industry is considered to be at the pinnacle of manufacturing engineering. However, the computational capabilities weren't advanced enough to such a level, even though they were commercially outsourced with only the state of art products invested due to the push towards the space race and the minimal desired margin of error.

7. ROLE OF UNCONVENTIONAL TECHNIQUES

A comprehensive problem statement was formed after the above analysis for the overall evolution of grid stiffened structures, in particular, isogrid due to its wide-ranging



Figure 4 3D printing for manufacturing Isogrids [3]

Applications in aerospace, and architecture. Figure 3 above is depicting 3D printing technology being used for manufacturing isogrids. The problem statement is from an all-inclusive standpoint taking into account all parameters such as commercial aspects, and material development. It throws light on possible domains of research, namely:

- Due to the emergence of the use of reusable rockets, the constraint for minimizing manufacturing duration, with the same resources, both materialistic and abstract, newer methods of production could be looked at, where focus shifts from maximum production efficiency time-wise, to methods that require a higher amount of time to execute the same operation but give a better quality of product. However, a precarious compromise must be set for each stage of this transition.
- Due to the premise available with non-conventional manufacturing techniques, it is now possible to manufacture grid stiffened structures with lesser amounts of downsides provided (rubber mould with clearance paper reference) by conventional techniques, hence reducing imperfections and caters for design tolerances at the same time.
- The new domain of development brought by these techniques also means that a higher degree of focus can be maintained on design optimization of the structure itself. A possible spectrum of research would be topology optimization of such grid stiffened structures for load specific purposes which may provide higher margins of cost and material saving by distributing loads in novel ways, and the manufacturing of such complex components have now been made possible by these techniques.
- A major advantage of these manufacturing techniques is also the control over extrinsic properties due to a more local approach to the “building block” of the structure, i.e. filaments, fibers, powder chunks etc. as opposed to conventional subtractive techniques, changing the bulk properties of the overall structure and giving rise to more desired deformation behaviors on load application (insert ref)

The research in additive manufacturing technology has expanded exponentially after the development of metallic materials and alloys used in these techniques. The applications of grid stiffened structures, particularly isogrids are heavily impacted by material selection and manufacturing constraints. Lighter and more complex structures can now be produced successfully with tighter tolerances and yet lower factors of safety due to higher amounts of precise load distribution saving millions of dollars’ worth production costs per rocket.

Isogrids are essentially composite structures due to multiple “layers”, namely, the grid and the skin, with multiple sub-layers in hierarchical structures [5]. The structure till now has been constructed mainly out of a single material which can be viewed as a constraint. Different layers

of this “composite” have different amounts of stresses and load distributions when force is applied. With the help of additive manufacturing, different metals, each with different distinctive advantages can be used in these different layers to achieve maximum overall benefit.

8. RESULTS AND DISCUSSION

This work analyses the effect and consequences of using different manufacturing techniques for manufacturing of grid stiffened structure, from three viewpoints, namely:-

Ease of Manufacturing

The reasoning for using a particular manufacturing technique and tooling methods for type and magnitude of loading on the structure.

Taking into account other secondary constraints as a factor, with parameters like temperature, budgetary constraints, aesthetic and applications of the structure itself.

9. CONCLUSION

It is seen that there are many areas, especially the aerospace industry which require isogrids in a large scale in order to increase the mass ratio of the aircrafts along with increase in strength. At the same time, these isogrids find a lot of use in gas turbine casings for increasing the stiffness. It is also visible that wherever hierarchical configuration of isogrids is used, load bearing capacity increases. It is also seen that a large variety of materials can be used for its manufacturing; Aluminum being the most commonly used material. The choice of manufacturing technique will vary from material being used and at the same time, certain parameters are needed to be considered; some of them being ease of manufacturing, quality of product, cost of manufacturing along with production efficiency along with seeing whether the process is hazardous or not. From the above study, unconventional machining has taken the upper hand over the more commonly used milling on the basis of the above parameters. It was seen that AWJ technique and then additive machining has enhanced the quality of the final isogrid fabricated. Also in the future more and more use of unconventional machining will be included for the manufacturing of these structures.

10. FUTURE SCOPE

As seen in the study, a radical shift towards additive manufacturing is being observed and hence, a lot can be done by composite structure manufacturing. By this, instead of using only one suitable material, 2 or more materials with different properties, each fulfilling different purposes of an isogrid, can be used. By doing so, mitigation of the materials' limitations will take place and also, the much-needed increase of the mass ratio will take place due to a reduction in the structural weight along with an increase in resilience. This will help in the lowering in the margin of the factor of safety, hence resulting in a lot of cost-saving per cycle of rocket shell manufacturing. To achieve this weight reduction, topology optimization of structure can be done, in which areas will be removed which sustain minimum stress while making sure stress-induced in other areas/ layers of isogrid do not exceed the elastic limit.

Since there is impact loading on the isogrid structure, analysis of the impact loading on the geometry as well as a study on the explicit dynamics can be done. By doing so, equivalent models can be created which will be useful to study the amount of energy absorbed along with deformation with time for each model. The results will be then compared with each other as well as the pre-existing models, resulting in the optimization of the isogrid structure.

REFERENCES

- [1] Kim, T. D. (2000). Fabrication and testing of thin composite isogrid stiffened panel. *Composite Structures*, 49(1), 21–25. doi:10.1016/s0263-8223(99)00122-1
- [2] Sorrentino, L., Marchetti, M., Bellini, C., Delfini, A., & Del Sette, F. (2017). Manufacture of high performance isogrid structure by Robotic Filament Winding. *Composite Structures*, 164, 43–50. doi:10.1016/j.compstruct.2016.12.061
- [3] Forcellese, A., Pompeo, V. di, Simoncini, M., & Vita, A. (2020). Manufacturing of Isogrid Composite Structures by 3D Printing. *Procedia Manufacturing*, 47, 1096–1100. doi:10.1016/j.promfg.2020.04.123
- [4] Huybrechts, S., & Tsai, S. W. (1996). Analysis and behavior of grid structures. *Composites Science and Technology*, 56(9), 1001–1015. doi:10.1016/0266-3538(96)00063-2
- [5] Li, M., Lai, C., Zheng, Q., Han, B., Wu, H., & Fan, H. (2019). Design and mechanical properties of hierarchical isogrid structures validated by 3D printing technique. *Materials & Design*, 107664. doi:10.1016/j.matdes.2019.107664
- [6] Yang, J. (2015). Design and characterization of an innovative integrally stiffened structure using Additive Manufacturing (Post Graduate). Coventry University.
- [7] Murthy, V. C. A. D., & Santhanakrishnanan, S. (2020). Isogrid lattice structure for armouring applications. *Procedia Manufacturing*, 48, e1–e11. doi:10.1016/j.promfg.2020.05.099
- [8] Rocket Science 101: Lightweight rocket shells – Aerospace Engineering Blog. Aerospace Engineering Blog. (2016). Retrieved from <https://aerospaceengineeringblog.com/rocket-science-101-lightweight-rocket-shells/>.
- [9] Isogrid. En.wikipedia.org. (2020). Retrieved from <https://en.wikipedia.org/wiki/Isogrid>.
- [10] Quest Integrated Inc. (1990). Abrasive-Waterjet Machining of Isogrid Structures (pp. 1,4,6,8). Retrieved from https://www.researchgate.net/publication/235215449_Abrasive-Waterjet_Machining_of_Isogrid_Structures
- [11] Green, R., & Shore, P. (2005). US7631408B2 United States.
- [12] Huybrechts, S. M., Meink, T. E., Wegner, P. M., & Ganley, J. M. (2002). Manufacturing theory for advanced grid stiffened structures. *Composites Part A: Applied Science and Manufacturing*, 33(2), 155–161. doi:10.1016/s1359-835x(01)00113-0
- [13] Vasiliev, V. V., Barynin, V. A., & Razin, A. F. (2012). Anisogrid composite lattice structures – Development and aerospace applications. *Composite Structures*, 94(3), 1117–1127. doi:10.1016/j.compstruct.2011.10.023
- [14] Ananth, Sirija & Whitney, Thomas & Toubia, Elias. (2018). Buckling Stability of Additively Manufactured Isogrid. 10.12783/asc33/26164.
- [15] Yang, Q. & Yang, S. & Lin, X.. (2015). Impact response of stiffened cylindrical shells with/without holes based on equivalent model of isogrid structures. *Computers, Materials and Continua*. 45. 57-74.

- [16] Huybrechts, S., Hahn, S., & Meink, T. (1999). Grid Stiffened Structures: A Survey of Fabrication, Analysis and Design Methods. In International Conference on Composite Materials (p. 10). Paris.
<https://www.iccm-central.org/Proceedings/ICCM12proceedings/site/papers/pap357.pdf>.
- [17] Wong, Kaufui. (2012). K.V. Wong, A.Hernandez, “A Review of Additive Manufacturing,” ISRN Mechanical Engineering, Vol 2012 (2012), Article ID 208760, 10 pages.. ISRN Mechanical Engineering. 2012. 10.5402/2012/208760.
- [18] Akl, W., El-Sabbagh, A., & Baz, A. (2008). Optimization of the static and dynamic characteristics of plates with isogrid stiffeners. *Finite Elements in Analysis and Design*, 44(8), 513–523. doi:10.1016/j.finel.2008.01.015
- [19] Kanou, H., Nabavi, S. M., & Jam, J. E. (2018). Numerical modeling of stresses and buckling loads of isogrid lattice composite structure cylinders. *International Journal of Engineering, Science and Technology*, 5(1), 42. doi:10.4314/ijest.v5i1.4
- [20] Sorrentino, L., Marchetti, M., Bellini, C., Delfini, A., & Albano, M. (2016). Design and manufacturing of an isogrid structure in composite material: Numerical and experimental results. *Composite Structures*, 143, 189–201. doi:10.1016/j.compstruct.2016.02.043
- [21] McDonnell-Douglas Aeronautics Co. Isogrid design handbook. Huntington Beach, CA, 1973

DECLARATION OF INTERESTS

- ☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- ☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: **NA**



Expectation maximization clustering and sequential pattern mining based approach for detecting intrusive transactions in databases

Indu Singh¹ · Rajni Jindal¹

Received: 7 August 2020 / Revised: 9 December 2020 / Accepted: 4 March 2021 /

Published online: 22 May 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Database security is pertinent to every organisation with the onset of increased traffic over large networks especially the internet and increase in usage of cloud based transactions and interactions. Greater exposure of organisations to the cloud implies greater risks for the organisational as well as user data. In this paper, we propose a novel approach towards database intrusion detection systems (DIDS) based on Expectation maximization Clustering and Sequential Pattern Mining (*EMSPM*). This approach unlike any other does not have records and assumes a predetermined policy to be maintained in an organisational database and can operate seamlessly on databases that follow Role Based Access Control as well as on those which do not conform to any such access control and restrictions. This is achieved by focusing on pre-existing logs for the database and using the Expectation maximization clustering algorithm to allot role profiles according to the database user's activities. These clusters and patterns are then processed into an algorithm that prevents generation of unwanted rules followed by prevention of malicious transactions. Assessment into the accuracy of *EMSPM* over sets of synthetically generated transactions yielded propitious results with accuracies over 93%.

Keywords Database intrusion detection · Database security · Expectation maximization clustering · Data dependency mining · Sequential pattern mining · Anomaly detection

1 Introduction

Over the last few years, due to tremendous progress in the availability of Internet, the usage of networked system has expanded exponentially by several folds, which is consequently confronting a graduated extent of infringement of confidentiality, integrity and

✉ Indu Singh
indusingh@dtu.ac.in

Rajni Jindal
rajnijindal@dce.ac.in

¹ Department of Computer Science Engineering, Delhi Technological University, Delhi-110042, India

availability to the databases. The rise in acceptance of databases by technology driven organisations, companies, institutes as the dominant management platform also affects the security quotient to a considerable magnitude.

Thus, data security is more vital than ever. At the same time, the security of data is to be handled for both outsider and insider threats [4]. The CERT Guide to Insider Threats [8] offered specified measures and guidance implemented from executives, managers and other staff in any operational organisation and that too with agile representations. Various measures for cyber frauds, insider threats and theft of information identifications were elucidated with reference to the software life cycle. The effectiveness of the previous security tools was enhanced through rules, configurations and business processes.

An Intrusion Detection System(IDS) [13] provides a framework for monitoring traffic linked with any sort of security violations and generates an alert whenever the presence of any malicious activity is detected. The working of an IDS is established on the hypothesis that due to anomalies in usage patterns, the system becomes sensitive towards numerous types of exploitation evident as masquerading, privilege abuse, legitimate privilege abuse and privilege elevation. Any malicious interruption or violation is centrally collected and forwarded to either the common Admin or the specified Security Information and Event Management System(SIEM).

A dominant obstacle in identification and mitigation of internal threat arises from the very fact that intruders are legit users of the system and together with relevant access rights they are conversant in the organisation of the information schema and therefore the security mechanism that is employed in situ. Therefore, business executive threats will persist for long periods while not being detected and cause extensive damages to information systems. Such complex environments require a high security levels to ensure integrity of information communicated between various entities in an organisation. An intrusion detection system acts as a flexible protection technology for a system's security. Malicious intrusion attacks evolve over time, so it is pertinent for intrusion detection and prevention technologies to adapt along with the level of the threats. Intrusion Detection System (IDS) is an accepted mechanism for coping with most of the future threats. The fundamental motivation behind it is to look for deviations or anomalies within the data resources on the host system.

There exists a number of IDS divided into various families according to their properties [13]. The most common classifications are:

- Network Intrusion Detection Systems (NIDS): which continuously analyse the incoming transactions and traffic within the network.
- Host-based Intrusion Detection Systems (HIDS): which monitors the central file structures of computing systems as well as the incoming network packets.

Other classifications for IDS are based on anomaly detection, and misuse or signature detection.

- Anomaly-based Detection: This detection model is able to detect and adjust to novel attacks which were previously unknown. This method uses machine learning techniques to create a structured model for trusted transactions or behaviour, and then compare new input behavior against this standard model. Although this approach detects previously unknown attacks, it may suffer from false positives and legitimate activity, dissimilar to the ones contained in the model, can accidentally be classified as malicious.

- **Misuse Detection:** This type of model detects possible threats by looking for specific patterns which are previously known to be related to certain threats. Although it is able to detect known attacks it suffers from the fact that it is unable to detect new attacks for which no pattern is available.

In an organisation, the data administrator sets permissions and privileges to enable specific access to the database to authorised users. Role-based access control (RBAC) is such a mechanism to restrict access to the system. RBAC provides the users with different levels of access based on their roles within the organisation. Herein, users are not assigned permissions directly, but rather acquire them through the job functions and roles assigned to them. Instead of managing user access and permissions at the base level, they are consolidated over a network to a set of roles. The Role Based Access Control (RBAC) binds the user to a specific attribute realm as per their roles thus controlling the access to the sensitive parts of the database and ensuring database security at the Role level. This strategy is implemented by extracting sequences [2] and association rules [1] from the valid transactions which are used to verify whether an incoming transaction is legitimate or malicious. The NIST model [49] further laminated the security system by restricting the user activity to specific roles.

Using the Role-based Access Control mechanism [48] the problem of insider attacks is being solved to a great extent. RBAC maps distinct access privileges to users of distinct roles. Conventional techniques at this level employ pattern mining for finding consistent patterns in the valid transactions already stored in the database logs [5, 60]. A similarity score is then generated by evaluating the incoming transaction against the previously mined patterns, subsequently this score is then used to check whether the transaction under consideration is malicious or not, based on a chosen threshold. In our approach we refer to this score as QLAI (Query Log Affinity Index). Our major space of concentration in this paper lies in detection of every type of malicious information transactions. Most of the authentication systems cater to detection of external attacks however, in this paper, we propose a unique approach of classification of transactions which makes use of mined association rules and cluster analysis to observe and stop abuse of internal privileges. In our approach (EMSPM) information dependency rules are obtained by mining user information access patterns using modified Prefixspan and role profiles are generated by agglomeration of the user activity parameters from information logs using Expectation maximization clustering with Gaussian Mixture Model. The new transactions are passed through the anomaly detection algorithm before execution, and an alarm is generated if the transaction is assessed as malicious therefore preventing them from modifying any sensitive data. Our main contribution is summarized as follows:

1. We design a robust DIDS, capable of identifying intrusions in RBAC as well as non-RBAC administered databases, that employs a rule mining component along with a role-independent anomaly detection component to determine whether a query is malicious or not
2. We use data dependency mining to explore read and write rules based on the previous access patterns to determine the legitimacy of the transaction under consideration. Further it uses EM clustering to form role profiles from the transactions available in database logs and subsequently assign the incoming transaction to one of these profiles
3. We propose an approach that reports a substantial increase in the nature of attacks that are captured and gives overall improvement in performance over ancient IDS.

The rest of the papers are composed as follows. An overview of the related work is provided in Section 2. Section 3 describes the proposed approach. Experimental results and

comparison with existing approaches are given in Section 4. Lastly, we state the conclusion in Section 5.

2 Related work

Numerous Intrusion Detection System Models [23, 32, 57] have been introduced for detecting network-based intrusions and intrusions at the OS level but very few models have been developed to recognize database intrusions. The use of database IDSs is necessary for the following purposes:

1. Malicious activities for a database system might be intrusive at different levels of system programs (i.e. Operating System Level) or network level.
2. Intrusion detection system models designed at the Operating System and network-level are ineffective to shield databases from Insider threats. Insider threats are difficult to isolate as the intruders are authorized users of database. [5]

An intrusion is defined as a set of unauthorised activities on a database or a network [22]. In the past few decades, researches and experimentations have accustomed numerous advances in IDS [55, 59, 60] for detecting intrusions at the network level, database level and at the OS level. Research efforts in the identification of abnormal events [12] and frequent patterns illustrate that the detection of abnormal events is based on user behaviour and user Profiling.

According to D.E. Denning [15] a system when examined and inspected for its records in audit for unusual patterns, portrayed possible security violations citing the presence of malicious activity. The model utilized a rule-based pattern matching system to inspect the functionalities such as login and executable programs.

However, at the local or basic level, the system suffered a major drawback due to vulnerabilities since complex functionalities were not dealt with fully. In contrast, Debar et al. [13] illustrated different norms wherein the focus was laid upon the functioning of inner elements along with respective behaviour and principles. The pattern consolidated the suggestion of several evaluation metrics in succession to evaluate the effectiveness of accuracy and thereby employing chance variation procedure. User access patterns of legitimate users are compared to ones stored in database. Bertino and Sandhu [4] considered access control systems for access with certain multilevel security implementations. Cardenas et al. [9] concluded that no matter control systems are pivotal but they are still prone to cyber attacks, thus making it vulnerable to the attackers. In fact, security mechanisms that monitor and find modifications towards the controlled data prior to an attacker damaging a system were well defined and worked upon. Liao et al. [29] proposed that due to alarming increments in security threats and networking, intrusion detection system(IDS) has been decisive and pivotal in the computational world.

2.1 Data dependency mining

Data Dependency mining [21] is used to discover relevant patterns from large Data sets through statistical methods to investigate them by notifying us about the relationship between the attributes. It is proven to be helpful in data rectification and resource filtering. We review a few related papers utilizing this method for intrusion detection.

According to Chen et al. [10] extracting information from databases is known to many researchers and it has been a crucial topic in machine learning and database systems. Data

Mining is used for improving and understanding user behaviour better so as to improve opportunities in various fields. The concept of data dependencies among different attributes of a transaction based on frequent subsequences of database transactions was introduced by Panda et al. [59]. They classified transactions which did not fit with these data dependencies as malevolent but One shortcoming of this method was that proper values of confidence and support parameters were required to be recognised. These problems were solved by Sohrabi et al. [53] by the use of k-optimal rule determination and lift as an interestingness measure.

Hu and Panda [59, 60] highlighted that the data dependency relations are utilized for detecting the database intrusions. Data items are written or read prior to updating. The model by Hu and Panda [59] illustrates techniques for extracting data dependencies and uses Petri nets to simulate normal data. This approach was further extended for mining data dependencies [60] from the database log. The transactions not qualifying to the mined dependencies are treated as malicious ones. Srivastava et al. [55] further modified the above approach by considering the sensitivity of the attributes which accounts for increased performance. The algorithm detected the changes in sensitive attributes with agility.

Hashemi et al. [20] extended the existing dependency mining approaches [10, 43, 59, 60] to enhance intrusion detections. The proposed model mined dependencies existing among data items, hence discovering transactions repudiating them and employing a behaviour similarity test between these and the valid transactions in logs. Malicious transactions are identified as those which don't pass this similarity test, consequently reducing the overall false-positive rate. Through anomaly detection in time series dataset it is capable of detecting those intrusive transactions which despite adhering to the dependencies induces abnormalities in data update patterns, thereby increasing the true positive rate.

Subudhi et al [56] proposed an intrusion detection system (IDSs) which utilised OPTICS clustering along with Ensemble Learning comprising of several aggregation methods like Boosting, Bagging and Stacking for the identification of malicious users in the database. Their approach of intrusion detection was divided into 2 phases namely training and testing. The Training phase preprocessed input dataset features. Thereafter behavioural profiles were generated by using OPTICS clustering. The testing phase checked the belongingness of an incoming transaction with any of the seen profiles.

Sallam et al [46] used anomaly detection algorithm to identify data aggregation and data updation. Their technique employed the usage of normal table reference rate and retrieval of tuples from the former database access logs. Further, the incoming user queries were examined to detect queries that may conclude to surpass the normal rates of data access.

2.2 Sequential pattern mining

Sequential Pattern Mining (SPM) is a data mining subdomain coined by Agrawal et al and is widely used in areas of database security [5, 27, 60] to discover frequently occurring sequences, interesting features and patterns in sequential databases.

Agrawal et al. [1] introduced new algorithms - Apriori and its variations to mine the rules under associations between data items in a database. Apriori Hybrid algorithm provides linear scalability with the size of the database which overcomes the drawback in Apriori algorithm. One major limitation of the algorithm was to settle suitable utility for support and confidence.

Agrawal and Srikant [2] proposed AprioriAll and AprioriSome for extracting sequential patterns in a database. Both the proposed algorithms scale linearly with respect to the size of the database but encounter the similar problem of deciding the support and confidence worths. These limitations were handled when Srikant and Agrawal [54] proposed

generalised sequential pattern mining algorithm which was more efficient than the Apriori algorithm as it considers the distinct signs of the items in real-time applications, with the weight being either the cost or the profit of the item being utilized.

Lan et al. [27] proposed a different method for finding weighted sequential patterns. The maximum weighted upper bound model was considered resulting in higher accuracy and improved efficiency for the subsequences. The major drawback with this model was its inability to handle dynamic addition, removal or updation of sequences.

Rahman et al. [43] developed the notion of pattern mining through an algorithm that pulls out sequences from uncertain databases with weight constraints. The main challenge was to recognize the significant and valid patterns as per their relevance. This drawback was overcome in 2019 when Rahman et al. [42] utilized weight and support constraints to develop patterns in uncertain databases.

Viger et al. [30] proposed a concise and efficient survey to provide the limitations of above stated traditional sequential pattern mining approaches. It also illustrates popular variations of the task of sequential pattern mining and its open source implementations.

2.3 Anomaly detection

Anomaly detection is the method of finding events or observations which differ significantly from the normal trend or patterns. In IDS, anomaly detection is extended to identify malevolent users whose activities significantly drift from a User with normal behaviour.

On the basis of audit trails, DEMIDS [11] created user-profiles which were further utilized to recognize ill-usage and insider exploitation. The proposed approach with the help of a distance measure, inputs information about the data structures and semantics of a particular database schema, which in turn is used to obtain frequent itemsets describing the operation range of users.

Intrusion Detection systems based upon Anomaly Detection are able to restrict role-based intruders [20] by detecting individuals behaving differently than the legitimate users of a particular role, thereby protecting us from insider threats. The number of false-positives an IDS detects is one of the primary points for determining its performance.

Zamanian et. al [62] examined the performance of multiple user profile training procedures to decrease the number of false-positive alarms and presented the notion of symmetry to Grouping profiles, but To minimise the false alerts created, they extended the time window practised for training user profiles.

Ronao et al. [45] introduced another anomaly detection method for Role-Based Access Control databases using principal component analysis and random forest with weighted preference methods. The false-positive and false-negative rates were improved by this approach, furthermore, model building time and execution time were also reduced notably. The major drawback of this technique was confusion among the roles which have similar access permissions.

Yip et al. [61] remarked that more extra inferences can be recognised by analysing data saved in the database. They applied five distinct inference rules instead of simply operating functional dependencies to recognise inferences, also they never encountered any counter-examples to prove the invalidity of the above-stated rules.

To eliminate the risk of internal intruders, Rashid et al. [44] introduced a novel method which employed Hidden Markov Model for detecting anomalies in the normal user behaviour, which led to an improvement in the existing analysis of the system and also safeguarded the system against insider threats.

Kim et al. [25] suggested a novel approach of classifying the user's role and influence by extricating characteristics from SQL queries with the help of CNN-LSTM neural networks which in the anomaly detection technique, automatically extricates essential characteristics of database query and LSTM creates the temporary data of the SQL sequence. (CAP) Class activation map distinguishes the SQL query characteristics that influence the classification further and CNN- LSTM neural networks excel other Modern machine learning techniques by a large extent.

Kamra et. al. (2008) [24] illustrated Dependency and relation analysis of Role-Based Access Control System but the major drawback of this approach was that Query semantics were not taken into account during the generation of quillits

Mazzawi et al [34] coined a two-step new machine-learning algorithm to identify unusual activities. The first step was to check for self-consistency, to ascertain that the user activities are consistent with previous patterns with the help of a probabilistic model. The second step tests for global-consistency, to check the steadiness of user activities with their past actions

Panigrahi et al [40] used a deep learning model which mainly concentrates on exploiting data dependencies, the normal behaviour of users, and data sensitivity of different transactions to identify intrusion. They used various sorts of neural networks in accordance with their robustness of identifying the intrusion on the basis of the type of data such as sequential data and featured data.

2.4 EM clustering

The Expectation-maximization algorithm which attains maximum likeliness estimations of parameters in probabilistic models, was first introduced by Dempster et al [14] in 1977. It is an iterative approach which oscillates within two measures, Expectation (E) and maximization (M) till an expected convergence value is obtained. EM Algorithm uses the "finite Gaussian mixtures" method for clustering data points and evaluates a set of attributes iteratively until convergence is attained.

Bilmes et. al. [6] illustrated the famous application of Expectation-maximization Algorithm, that is, mixture density estimation. One of the properties of this algorithm is that possibility of $LX(\Gamma)$ is non-decreasing on every iteration. Moreover, this algorithm Converges to a local maximum of the possibilities, whenever there exists one or more local maximum of the possible function.

Ordonez et. al. [39] in their study gave preference to EM to cluster data over other algorithms for the following reasons amongst others. EM algorithm has a powerful Statistical basis and this algorithm depends linearly on the size of the database. Moreover, this algorithm is resistant to noisy data and can easily accept large clusters as input.

Do et. al. [16] suggested that True error rates can be determined in a much-principled action by introducing a consistent model of the fault process and MLEs of the characteristics. When the LOS class data, Discretized by applying the (EM) clustering algorithm produced seven clusters. Only 4 main clusters appeared, when clusters with the small number of members were mixed. Through this technique he improved the clustering efficiency, although the time complexity was not reduced at all.

Assaad et. al [3] in their study used Expectation-maximization (VEMDYMIX) to produce a temporary dataset. Their algorithm improved the clustering and evaluation accuracy with the smallest computation expense of time but the space complexity was not reduced at all. Similarly, Kuang et al [26] proposal of network intrusion detection model by combining SVM and TCABC showcases the amplitude of the constant progress and innovation made in the field of intrusion detection.

3 Proposed approach

We propose an approach called EMSPM which is a database intrusion detection system consisting of an initial training phase, rule mining followed by clustering and then classification. The algorithm does not assume any fixed access control policy to be implemented on the organisational database, i.e., it functions both when the policy of the database follows Role Based Access Control (RBAC) as well as those that don't follow RBAC. EMSPM uses existing database logs containing transactional details, to mine read and write (dependency) rules along with Expectation maximization (EM) clustering to generate role profiles by agglomeration of user activity parameters from information logs. The relevant patterns that are extracted by the algorithm are then pre-processed into a form that prevents unwanted rules from being generated. In the first part of the learning phase, generation of sequential patterns succeeded by generation of data dependencies occurs from the pre-processed transaction logs making use of the PrefixSpan [19, 31, 41, 51] algorithm. Classification rules above the predefined confidence value threshold are generated pertaining to the mined data dependencies. In the second part of the learning phase, role profiles (clusters) are generated from defined parameters after analysis of database logs using expectation maximization clustering [6, 14, 37, 50]. It is assumed that only legitimate transaction logs and database logs are provided in this phase. Existing intrusion detection systems are unable to stay robust to dynamic user patterns and do not possess the ability to familiarise themselves with previously unknown attacks. However the approach laid out in this paper has the benefit of lower sensitivity to change in user patterns and its ability to tackle new attacks.

3.1 System architecture

The suggested DIDS operates as an autonomous entity and improves system security without hampering the database management system's functionality. It is designed to detect and interrupt fraudulent transactions, sent by illegitimate users to the DBMS and have been able to bypass the system's preliminary protection. The situation could arise due to a hacked legal account, SQL injection-prone web portal, or user rights violation. In all these cases, the attacker can get access to the database by submitting transactions manually or through an application. EMSPM comprises of two parts, which are:

1. Learning Phase,
2. Detection Phase.

3.2 Learning phase

The learning phase architecture of the proposed Intrusion Detection System is demonstrated in Fig. 1. In this phase, transaction logs are used by the rule generator to remove data dependencies, which are then processed during later phases for compliance computation. Database logs are used to evaluate user access trends in conjunction with conformity estimates, which are then grouped into role profiles. Such clusters of roles and rules produced are then transferred to the detection process. The learning phase comprises of two essential components which are -

1. Rule Mining
2. Expectation maximization (EM) Clustering

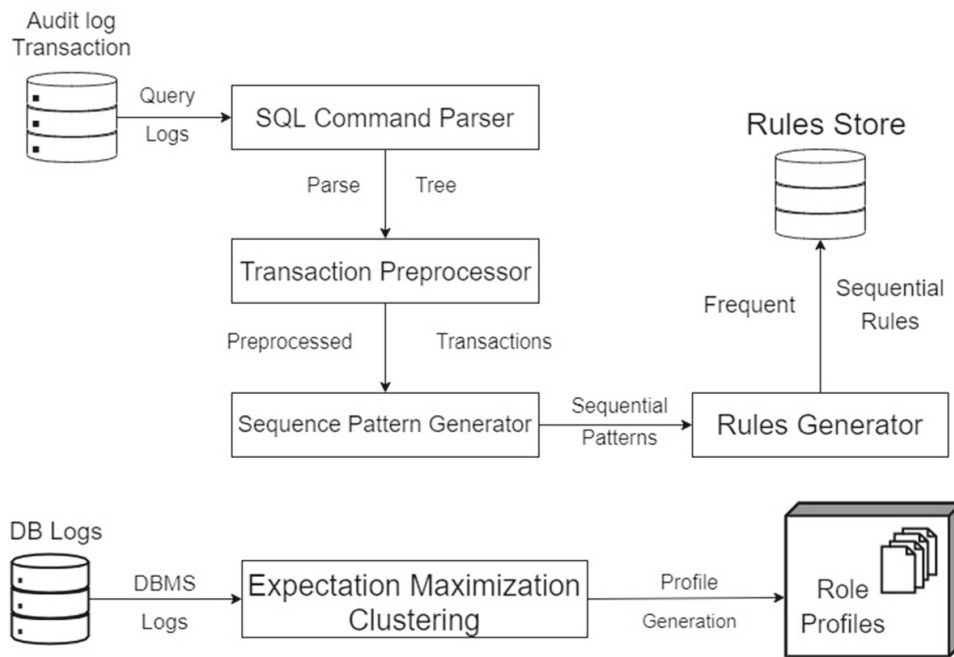


Fig. 1 Learning Phase architecture of proposed EMSPM approach

3.2.1 Rule mining

Rule mining refers to the discovery of patterns that represent the intrinsic and important properties of a database. The relevant patterns that are extracted by the algorithm are then pre-processed into a form that prevents unwanted rules from being generated. In the first part of the learning phase, generation of sequential patterns succeeded by generation of data dependencies occurs from the pre-processed transaction logs making use of the PrefixSpan [19, 31, 41, 51] algorithm. Classification rules that are above the defined confidence value threshold are generated pertaining to the mined data dependencies

SQL command parser This parsing module takes standard SQL command as its input and then parses and analyses them to convert the given input into the form of read and write sequences. The parser assigns a unique transaction ID to every transaction in the database. The commands are executed for tokens taken from the source token list. A parse tree is generated if no errors occur while parsing the input into read and write sequences. The generation of parse tree occurs recursively till the base case is encountered.

For example, If the following transaction is examined:-

accno=select acc_no from customer where name='Mark' and id=4321;

update Account set bal=bal + 4000 where acc_no=accno;

The output after processing would be of the form:-

Tid = < R(id), R(name), R(acc_no), R(acc_no), R(bal), R(bal), W(bal) >

Transaction preprocessor The transaction pre-processor asserts that transactions are defined in a suitable manner. It is important to view each logical transaction as a series of queries (select, add, remove or update). Then each query can be interpreted on item-sets as a series of operations (Read or Write). In order to avoid several representations for the same query, individual items in an item-set are lexicographically sorted.

Sequential patterns Sequence is defined as the order of occurring of transactional operation. Sequential patterns are said to be those items which occur simultaneously, by the same order multiple times in database log.

Definition 1 (Transaction) is an ordered list of operations to a given database instance, wherein the user operation can be represented with $Op \in \{\text{Read}, \text{Write}\}$. The execution of the operations must occur in a discrete manner for the transaction to be completed. Transactions are assigned with their unique identification TID.

Definition 2 (Operation $Op(x)$) represents the relation with the action of the data item x . Operation consists of action to the set $\{\text{read}, \text{write}\}$.

Definition 3 (Sequence (Seq)) refers to the ordered set of read and write operations with respect to time. We define a sequence Seq as

$Seq = \{Op_1(it_1), Op_2(it_2), \dots, Op_n(it_n)\}$, where $Op_i \in \{R, W\}$ and it_i is an attribute of data item. The data dependency sequences that are extracted from transaction logs and can be classified into two categories:

Definition 4 (Read Sequence $RSeq(x)$) is defined for a particular attribute x . We denote read sequence with the form $RSeq(x) = \{R(it_1), R(it_2), \dots, R(it_n), Op(x)\}$ where it_1, it_2, \dots, it_n represents the attributes that need to be read prior to the transaction performing operations $Op \in \{R, W\}$ on attribute x and $R(it_j)$ represents a quantum unit of read operation on it_j .

Definition 5 (Write Sequence $WSeq(x)$) is defined for a particular attribute x . We denote write sequence with the form $WSeq(x) = \{Op(x), W(it_1), W(it_2), \dots, W(it_n)\}$ where it_1, it_2, \dots, it_n represents the attributes that need to write after the transaction performs operations $Op \in \{R, W\}$ on attribute x and $W(it_j)$ represents a quantum unit of write operation on it_j .

Definition 6 (Confidence) provided for a sequence $Seq \in \{RSeq, WSeq\}$, is defined as the ratio of the number of occurrences of an attribute a_i in a given sequence to the total number of occurrences for that attribute a_i . Formally, the definition for Confidence is represented by the given formula:

$$Conf(RSeq(a_i)) = \frac{Count(RSeq(a_i))}{Count(a_i)} \quad (1)$$

Definition 7 (Read Rules(RR)) For each read sequence $\{R(it_1), R(it_2), \dots, R(it_n), Op(x)\}$, generated from sequential pattern mining, we generate read rule with the format $\langle R(it_1), R(it_2), \dots, R(it_n) \rangle \rightarrow Op(x)$, which infer that before x , we need to read it_1, it_2, \dots, it_n

If the rule generated from the read sequence has confidence greater than provided minimum confidence (conf), then the rule is added to the read rules set.

Definition 8 (Write Rules(WR)) For each write sequence $\{Op(x), W(it_1), W(it_2), \dots, W(it_n)\}$, generated from sequential pattern mining, we generate write rule with the format $Op(x) \rightarrow \langle W(it_1), W(it_2), \dots, W(it_n) \rangle$, which infer that after updating x , data items it_1, it_2, \dots, it_n should be updated during the transactions.

Table 1 Transactions for Mining Sequential Patterns

Transaction ID	Transaction
101	xr7, r1, r6, w5, r1, w4
102	r1, r5, w1, r4, r5, w4
103	r1,r6, r2, w4, r7, r3, w6, r1, r6, w2, r3, r5, r2, w5
104	r5, r6, w5, r5, w4
105	r2, w2, r4, r7, w3, r6, w5, r1, w4
106	r6, r5, w5, r3, r4, w4, r3, w7
107	r5, r3, r6, w7
108	r4, w6
109	r2, r5, w6
110	r6, r1, w3, r1, w6, r2, r7, r4, w2

If the rule generated from the write sequence has confidence greater than provided minimum confidence (conf), then the rule is added to the write rules set.

Cardinality of Table 1. displays the read and write items in a transaction. Each of the ten-transactions has been assigned with a unique Transaction ID.

Sequence pattern generator The task of extracting meaningful rules from data consists primarily of mining frequent sequential patterns through the successive algorithms since the transactions have been pre-processed into an ordered set of query templates. For mining frequent patterns we have used PrefixSpan algorithm [31] which mines the complete set of frequent patterns. The procedure to specify all the subsequences of one's transactional patterns and subsequently computing the respective supports to generate frequent patterns has a high computational cost. The improved Prefix-Span (Prefix-projected sequential pattern mining) developed by Han et al. (2001) [19] runs this process with a lower computational cost compared to both Apriori-based GSP [54] and FreeSpan [18] algorithms. The output results in the generation of read write sequence patterns with the minimum support of 4.

Definition 9 (Prefix and postfix) Given two transactional sequences

$a = \{Op_1(it_1), Op_2(it_2), \dots, Op_n(it_n)\}$ and

$b = \{Op_1(it_1), Op_2(it_2), \dots, Op_m(it_m)\} (m < n),$

sequence b is called a prefix of sequence a if and only if $Op_i(it_i) = Op_{i'}(it_{i'}) \forall i \in \{1, 2, \dots, m\}$. Then $\{Op_{m+1}(it_{m+1}), \dots, Op_n(it_n)\}$ is the postfix of sequence a w. r. t. b if b is the prefix of a.

Table 2 Mined Sequential Patterns(Support Threshold)

Sequential Patterns
r6,w5,w4
r7,r6,r1
r7,r6,w4
r7,r6,w5

Definition 10 (Projection) Sequences a and b are two transactional sequences such that b is a subsequence of a . The projection of a onto its prefix b is sequence a' if and only if a' is the longest subsequence of a .

The patterns obtained are converted and stored as a set of read and write rules for classification of the incoming queries.

When all the frequent patterns are kept for rule generation, the issue of data redundancy arises. On the other hand, if the maximal patterns are kept where a sequence is considered to be maximal if all of its super sequences are infrequent [33], then it eliminates data redundancy to an extent by removing overlapping frequent patterns, however inducing bias in the system.

To solve the above mentioned problems, we make a compromise with the two methods by keeping the frequent patterns whose lengths are at least three. In Sequence Pattern Generator, we assign the length of frequent sequential patterns a value of three because the patterns of length one or two carry redundant information [51].

Some frequent patterns of length longer than three were observed to carry duplicate information, however, since we see patterns of length three frequently in practical scenarios we chose to keep the said patterns.

Cardinality of Table 2 displays the number of frequent sequential patterns mined using Table 1. Sequences following the operation of IPPS (Iteratively Pruned Prefix Span) Algorithm. Patterns with support value higher than the specified minimum support are added to the set of Sequential Patterns.

Rule generator The algorithm generates data dependency rules after the formation of frequent sequences. These sequences are transformed to the Read and Write rules.. Read rules are of the form

$$\langle R(it_1), R(it_2), \dots, R(it_n) \rangle \rightarrow Op(x)$$

which provide all the attributes that are read before any operation is performed on attribute it_i . Write rules are of the form

$$Op(x) \rightarrow \langle W(it_1), W(it_2), \dots, W(it_n) \rangle$$

which provide all the attributes that are updated after any operation is performed on attribute it_i . If the confidence of any rule generated from the sequence is greater than provided minimum confidence then the rule is added to the rules set.

Table 3 Read and Write Data Dependency Rules

Sequential Pattern	Data Dependency Rules	Confidence
r7,r6,r1	$r7, r6 \rightarrow r1$	100%
r7,r6,w5	$r7, w5 \rightarrow r6$	100%
r6,w5,w4	$w5, w4 \rightarrow r6$	100%
r6, w5,w4	$w4 \rightarrow w5, r6$	83%
r7,r6,r1	$r1 \rightarrow r7, r6$	80%
r7,r6,w4	$r7, r6 \rightarrow w4$	75%
r7,r6,w5	$r7, r6 \rightarrow w5$	75%
r7,r6,w4	$r6, w4 \rightarrow r7$	60%
r7,r6,w5	$r6, w5 \rightarrow r7$	60%

Cardinality of Table 3. displays the number of data dependency rules generated in the set $\in \{Read, Write\}$ corresponding to the sequential pattern generated from Table 2. Value of confidence for each rule is shown which is maintained above the specified minimum confidence.

The Dependency Rule Miner Algorithm takes as input non-malicious queries fetched from the transactional logs as input and returns a set of read and write rules within a single dictionary.

Algorithm 1 Dependency rule miner.

Data : DB : Initial Database, min_sup : Minimum support, min_conf: minimum confidence, min_len: minimum length of sequential patterns

Result : R: Set of Read and Write rules

```

1 begin
2   Initialisation
3   Dictionary D =  $\emptyset$ ;
4   Seq_Patt  $\leftarrow \pi$ ;
5   pos  $\leftarrow -1$ ;
6   Extract Frequent Sequential Patterns
7   for frequent sequential patterns  $\pi \in Seq\_Patt$  do
8     | IPPS(<>, (i,pos))
9   end
10  Procedure IPPS ( $\beta$ , DB| $\beta$ )
11  if length( $\beta$ ) > min_len then
12    | Seq_Patt  $\leftarrow$  append  $\beta$ 
13  end
14  foreach sequence s with pos in DB| $\beta$  do
15    | foreach item  $\Psi$  in s from pos to length(s) do
16      | if occurrence( $\Psi$ ) ==  $\emptyset$  or occurrence( $\Psi$ ) != start then
17        | | D  $\leftarrow$  generate and store all projections from  $\Psi$ 
18        | end
19      | end
20    | end
21  for projected sequences  $\gamma$  from D in DB| $\beta$  do
22    | if length( $\gamma$ )  $\geq$  min_sup then
23      | | IPPS ( $\beta + \gamma$ , DB| $\gamma$ )
24      | end
25    | end
26  Extract Dependency Rules
27  for frequent sequential patterns  $\pi$  in Seq_Patt do
28    | foreach rule  $\in Rule\_Generator(\pi)$  do
29      | if confidence(rule)  $\geq$  min_conf then
30        | | add rule to R
31        | end
32      | end
33    | end
34 end

```

Dependency Rule Miner Algorithm is then used to generate prefix span rules which are later used in detection phase to classify an incoming transaction. Step 3 initializes the Dictionary storing the projections generated from the formatted sequence tokens. We chose the data structure as a dictionary since it provides us with a highly efficient method to access and store the read and write rules.

Step 4 to 5 initialize the sequential pattern set and the start position of projections to be used. In step 7 to 9, we make a call to the subroutine that for all the sequences it employs the Iterative Pruned Prefix Span (IPPS) algorithm with the starting position of sequence.

In step 11 to 13, we check whether the length of the projected sequential pattern is greater than minimum length threshold (*min_len*). If it is greater then we append the projected pattern to the set of sequential patterns. From steps 14-19, for each read and write sequence projected with the starting position from the projected database, we generate all the possible projections and store the projections in the Dictionary.

In steps 21 to 25, we retrieve the projections stored after processing from the dictionary. In the first call of subroutine, length-1 sequences are generated which works faster as the threshold of support is applied and saves execution time by pruning recursion of other sets. Recursive call is made for the sequence having higher support threshold than certain threshold (*min_sup*) to produce further projection with increasing length of sequence and appending previous projection of sequence.

From lines 26 to 34, sequential patterns are transformed into a set of read and write rules with the function Rule_Generator(). All the possible read and write rules generated having confidence greater than certain threshold (*min_conf*) are collected and other rules are discarded.

Later these R rules are used in Algorithm 3 to compute QLAI (See definition 13 for reference) of various queries within a transaction and classify it as malicious or non-malicious.

3.2.2 Expectation maximization clustering

EMSPM uses Expectation maximization clustering with Gaussian Mixture Models which aims to group the given transactions in database logs into various role profiles. EM Clustering, used as a gaussian mixture model clustering technique, was proposed by Dempster et al. [14]. The scope of EM algorithm's applications and its widespread applicability are evident in the book by McLachlan et al [35]. Neal et al. [37] introduced other variants of the EM algorithm, like "incremental", "sparse" and "winner-take-all" versions which employed joint maximization of the function by other means - which in turn led to maximization of the true likelihood.

EM is a distance based algorithm generally used for estimation of density. EM algorithm with Gaussian mixture model proves to be a robust technique for clustering, especially where the data is determined to be insufficient. Statistical models of the data can be generated using the EM algorithm with minimal computational time requirement (Mitra et al. 2003) [36]. This technique is used to estimate the parameters of probabilistic models to which the algorithm attempts to fit the given data. Starting with random initial parameters, it iteratively performs two main steps. In the first step called expectation step, the expected cluster probabilities are computed while in the second step called maximization step, distribution parameters and their likelihood over given data is computed. The above two steps are repeated until the log-likelihood that measures quality of clustering reaches termination condition which specifies the convergence criteria.

EM clustering was used due to its low computation time and high accuracy. Moreover, other advantages of the EM algorithm as explained by Ordóñez et al. [39] included guaranteed increase in likelihood with each iteration, strong statistical basis, linearity in database size, ability to stay robust to noisy data and capability to handle desired number of clusters (even of high dimensionality) as input. In addition to that, it provides fast convergence, given a good initialisation.

Given a dataset $\{x\}_{i=1}^N$ our goal is to assign every cluster to an instance. We assumed that there are N data points in the dataset and that there are k clusters. We model the index of the cluster as a random variable $z = j$ and a multinomial distribution satisfying $\sum_{j=1}^k \pi_j = 1$ is used to output the probability of the index of the cluster, such that

$$\pi_j = p(z = j), \forall j, j = 1, \dots, k \quad (2)$$

p is a Gaussian distribution, I_j the identity matrix of order j . The parameters that are to be found, i.e. the mean μ_j variance $P_j = \text{diag}(r_1, r_2, \dots, r_j)$ and the distribution function p_j are approximated.

$$\theta = \left\{ \mu_j, \sum_j, \pi_j \right\}_{j=1}^k \quad (3)$$

$$p(x|\theta) = \sum_{z=1}^k (p(x|z, \theta) p(z|\theta) \pi_j) \quad (4)$$

where z is an unknown variable. The total log likelihood of all data is given by

$$l(\theta, D) = \log \prod_{i=1}^N \sum_{j=1}^k \pi_j e^{-\frac{\|x_i - \mu_j\|^2}{2\sigma^2}} \quad (5)$$

We choose parameter values that maximize the likelihood function. Here D denotes the data. Some of the unknowns are assumed to be known, to simplify the optimization, while estimating the others variables and vice versa. For each class, the conditional expectation of $z = j$ given the data and the parameters.

$$\begin{aligned} w_j &= p(z = j|x, \theta) = \frac{p(x|z=j, \theta)}{p(x|\theta)} \\ &= \frac{\pi_j N(x_i|\mu_j, \sum_j)}{\sum_{i=1}^k \pi_j N(x_i|\mu_j, \sum_j)} \end{aligned} \quad (6)$$

Since each point x contributes to w_j in some proportion, for particular x_i we have

$$w_{ij} = \frac{\pi_j N(x_i|\mu_j, \sum_j)}{\sum_{i=1}^k \pi_j N(x_i|\mu_j, \sum_j)} \quad (7)$$

The optimization algorithm is called EM Algorithm and it groups incoming transactions into clusters with similar transactions representing a particular role on the basis of several attributes relating to user-access patterns within transactions.

EM clustering algorithm is used to create transactions role profiles which are later used in detection phase to assign a role profile to an incoming transaction. In steps 3 to 5 cluster means, covariances and mixing coefficients are randomly initialized for each of the K clusters. The aim of EM clustering is to cluster the transactions into K components so as to maximize the log likelihood of the given database computed in step 6. Then E-step and M-step is performed repeatedly until the convergence criteria is satisfied.

In E-step from steps 10 to 11, membership probabilities of each cluster are computed for the given dataset. Each point in the given dataset has a membership probability associated with every cluster, and lies between the values 0 and 1. Membership probabilities are then used to maximize the log likelihood of a given dataset under the current distribution.

From steps 14-17 cluster parameters — means, covariances and mixing coefficients are recalculated using updated membership probabilities.

Algorithm 2 Expectation maximization clustering.

Data : X : user transactions logs , K : number of clusters

Result : $\mu(k)$: means of K clusters,
 $\sum(k)$: covariances of K clusters,
 $\pi(k)$: mixing coefficients of K clusters

```

1 begin
2   Initialization
3    $\mu(k) \leftarrow$  random  $K$  cluster means
4    $\sum(k) \leftarrow$  random  $K$  cluster covariances
5    $\pi(k) \leftarrow$  random  $K$  cluster mixing coefficients
6    $L_0 \leftarrow \sum_{n=1}^N \ln p(x_n | \pi, \mu, \sum)$ 
7   Learning Phase
8   repeat
9     Expectation Step
10    for  $k \leftarrow 1$  to  $K$  do
11       $\gamma_k \leftarrow \frac{\pi(k)p(X/k)}{\sum_{j=1}^K \pi_j p(X/j)}$ 
12    end
13    maximization Step
14    for  $k \leftarrow 1$  to  $K$  do
15       $\mu_k \leftarrow \frac{\sum_{n=1}^N \gamma_k(x_n)x_n}{\sum_{n=1}^N \gamma_k(x_n)}$ 
16       $\sum_k \leftarrow \frac{\sum_{n=1}^N \gamma_k(x_n)(x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^N \gamma_k(x_n)}$ 
17       $\pi_k \leftarrow \frac{1}{N} \sum_{n=1}^N \gamma_k(x_n)$ 
18    end
19     $L_t \leftarrow \sum_{n=1}^N \ln p(x_n | \pi, \mu, \sum)$ 
20  until;
21     $L_t - L_{t-1} < \epsilon$ 
22 end

```

In step 19, updated likelihood is calculated using updated cluster parameters. The above steps are repeated until the algorithm converges i.e improvement of likelihood in current iteration is less than the specified threshold ϵ . After convergence the final cluster parameters represent the role profiles corresponding to the given database of transactions. The algorithm terminates when the criteria in step 21 is satisfied and improvement in log likelihood is below terminating threshold ϵ .

3.3 Detection phase

Once the learning phase of EMSPM is completed, we obtain a set of read and write rules as well as role profile clusters based on user access patterns. The complete detection

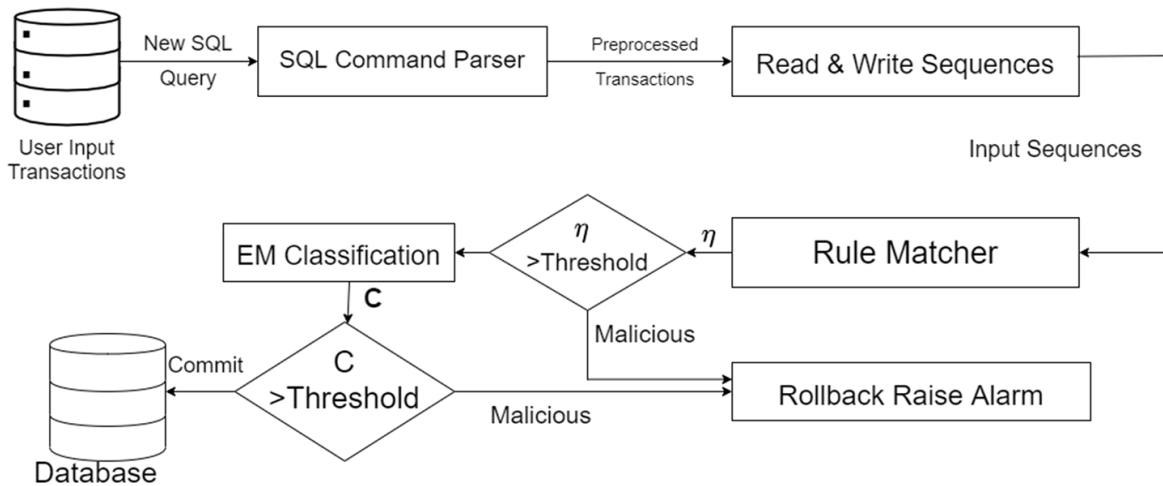


Fig. 2 Detection Phase architecture of proposed EMSPM approach

phase architecture of EMSPM is described in Fig. 2 which is used to classify incoming transactions as malicious or not.

The incoming SQL transaction will be interpreted and parsed against stored rules in the detection process. To classify the transaction as either malicious or non-malicious, the extent of conformity to the mined rules together with membership of the current user profile in the aforementioned role profile clusters are used. They treat compliance as the criterion for identification because in a real world scenario there may be some divergence from the allocated/stored rules.

The detection phase consists of the following modules:

1. Rule Matcher
2. Expectation maximization (EM) Algorithm

3.3.1 Rule matcher

Firstly, new queries pass through the command parser followed by generation of read and write sequences. Then, the Rule Matcher evaluates the read/write sequences of the incoming transaction with the data dependency rules defined during the learning phase for the attributes in the incoming query. The key problem of computation of rule-based similarity decomposes to quantification of the similarity of two sequence based datasets. Some common metrics for evaluating sequence-based similarity include Smith-Waterman [52], Levenshtein [28] and Needleman-Wunsch [38]. In EMSPM, Rule Overlay Count and Rule Overlay Extent determines the extent to which the query sequence matches with the Rules.

Definition 11 (Rule Overlay Count) is generated by the ratio of number of rules that satisfy a query to the total number of rules generated. Given that $0 \leq ROC \leq 1$. We have shown the formal definition of ROC as :-

$$ROC = \frac{Matched(Q, R)}{Total(R)} \quad (8)$$

For example, a sequence given as, $\{R(a), R(b), W(a)\}$

These rules are matched,

$$\langle R(a), R(b), R(c) \rangle \rightarrow W(a)$$

$$< R(a), R(c), R(d), R(b) > \rightarrow < R(e), W(a) >$$

Definition 12 (Rule Overlay Extent ROE) for a specific rule given for a sequence is defined as the degree of similarity with the input sequence and a set of rules. ROE for a rule and sequence is the maximum value generated by the weighted combination of Level-0-Similarity and Level-1-Similarity in the ratio of 3:7 respectively. ROE can measure coherence between the rules and sequence in range [0,1] where higher value of ROE makes the rule high compliance with the query.

$$ROE = \frac{3 * LOS + 7 * L1S}{10} \quad (9)$$

For example, we show the case for the rule generated $\{R(a), R(b), R(c)\} \rightarrow W(a)$, and a sequence

$$\{R(a), R(b), W(a)\}$$

Level-0-Similarity LOS The count of identical attributes between the rule and the sequence. We formally show the definition of LOS as

$$LOS = \frac{\text{Count of Identical Attributes}}{\text{Total Attributes}} \quad (10)$$

For the above example

$$LOS = \frac{(2)^2}{(3)^2} = \frac{4}{9}$$

Level-1-Similarity L1S Ratio of the length of longest common subsequence multiplied by 2 to the total length of the rule and sequence. We have shown the formal definition of L1S.

$$L1S = \frac{2 * LCS(R, Seq)}{(len(R) + len(Seq))} \quad (11)$$

For the above example,

$$L1S = \frac{2 \cdot 3}{(4+3)} = \frac{6}{7}$$

So for this example, the ROE will be calculated as follows:

$$ROE = \frac{3 \cdot \frac{4}{9} + 7 \cdot \frac{6}{7}}{10} = \frac{11}{15}$$

Definition 13 (Query Log Affinity Index QLAI) QLAI is defined as the benchmark for determining whether the incoming Transaction is malicious or not. QLAI takes into account the metrics, Rule Overlay Count(ROC) and Rule Overlay Extent(ROE). We formally show the definition of QLAI as :-

$$QLAI = \alpha \cdot ROC + \beta \cdot ROE \quad (12)$$

where α, β are the weights assigned to the respective metrics.

Steps 2 to 10 of the Detection Phase Algorithm initializes the Dictionary to store the data items required to match in the rule matching phase, AttrCount to count the unique attributes in a sequence obtained from the new query. AttrTotal gives us the total attributes inside a given rule. Match gives us the total number of rules matching a given sequence. ROC, ROE, LCS are initialized empty with LOS and L1S as empty lists. LOS and L1S are taken as list data types as we are appending the values and maximum of the values are taken into account.

Algorithm 3 Detection phase.

Data : R: Rules Store
RP: Role Profiles

Input : TRN : Transaction for New Queries
 η : Malicious Rule Threshold
 (α, β) : Parameters for QLAI

Result : Classification of the new Query as malicious or not

```

1 begin
2   Initialization Dictionary Attr  $\leftarrow \emptyset$ 
3   AttrCount  $\leftarrow \emptyset$ 
4   AttrTotal  $\leftarrow \emptyset$ 
5   Match  $\leftarrow \emptyset$ 
6   ROC  $\leftarrow \emptyset$ 
7   ROE  $\leftarrow \emptyset$ 
8   L0S  $\leftarrow \{ \}$ 
9   L1S  $\leftarrow \{ \}$ 
10  lcs  $\leftarrow \emptyset$ 
11  During Transaction
12  foreach Query  $q$  in TRN do
13    parseTree  $\leftarrow$  SQLParser( $q$ )
14    Sq  $\leftarrow$  SequenceGenerator(parseTree)
15    foreach Rule Set rule in R do
16      if rule is matching Sq then then
17        Match  $\leftarrow$  Match+1
18      end
19    end
20    ROC  $\leftarrow$  Match / TotalRules(R)
21    foreach Rule Set rule in R do
22      foreach or each operation Op( $a_i$ ) in rule do
23        if Attr has not  $a_i$  then
24          AttrTotal  $\leftarrow$  AttrTotal + 1
25          Attr  $\leftarrow$  update( $a_i$ )
26        end
27      end
28      foreach or each operation Op( $e_i$ ) in Sql do
29        if Attr has  $e_i$  then
30          AttrCount  $\leftarrow$  AttrCount + 1
31          Attr  $\leftarrow$  remove( $e_i$ )
32        end
33      end
34      L0S  $\leftarrow$  append ( (AttrCount)2 / (AttrTotal)2)
35      lcs  $\leftarrow$  LCS(rule,Sq)
36      L1S  $\leftarrow$  append (2*lcs / (len(rule) + len(Sq)))
37    end
38    Sq_ROE  $\leftarrow$  (3*L0S + 7*L1S) / 10
39    ROE  $\leftarrow$  maximum(Sq_ROE)
40    QLAI  $\leftarrow \alpha * ROC + \beta * ROE$ 
41    if QLAI >  $\eta$  then
42      Rollback – Raise Alarm
43    end
44  end
45  C  $\leftarrow$  classify(TRN)
46  if (C == malicious) then
47    Rollback – Raise Alarm
48  end
49  else
50    commit
51  end
52 end

```

In step 12, we use each of the queries present in the given transaction. Step 13 to 14, gives us the parse tree obtained through function `SQLParser()`. Generated parse tree is transformed into formatted sequence tokens using the function `SequenceGenerator()`. Steps 15 to 20 are responsible for calculating the number of matching rules with the sequence for each rule in the rule set. ROC metric is determined using the matched value.

Steps 21 to 27 use the dictionary to store the number of attributes that are new in the operation of read or write. Count of `AttrTotal` determines the total unique attributes in the present rule. In steps 28 to 33, a dictionary is used to retrieve the count for the data items present in the sequence which are either present in the rule or not.

In step 34, the value for LOS (Level-0-Similarity) is determined using the metrics `AttrCount` and `AttrTotal`. Steps 35 to 37 use the longest common subsequence algorithm to determine the LIS metric. LIS is defined with the metrics of length of rule, length of the sequence and the lcs value. In steps 39 and 40, we determine the ROE (Rule Overlay Extent) as the maximum value obtained from the weighted combination of the two metrics, LOS (Level-0-Similarity) and LIS (Level-1-Similarity) with appropriate ratios. In step 41 we determine the extent of rule matching with the value of QLAI (Query Log Affinity Index) which takes the metrics of ROC (Rule Overlay Count) and ROE (Rule Overlay Extent).

From steps 42 to 45, QLAI is sent to the detection engine for determining whether the query is malicious or not. If the query is malicious then, we raise the alarm and rollback the transaction. Finally in steps 46 to 53, we classify the given transaction through EM Classifier to determine whether the query is malicious or not. If the query is malicious then, we raise the alarm and rollback the transaction else we commit the given transaction to the database.

3.3.2 EM classification algorithm

Incoming transactions which are determined to be malicious by IPPS rules are rolled back without committing to the database whereas those which satisfy data dependencies are further inspected to satisfy user-access patterns by using EM classification module. The algorithm iterates over all the clusters, and computes the membership probability of the incoming transaction associated with each of the transaction profile clusters. Then it computes the maximum of such membership probabilities and following the calculation of the Cluster Deviation Index classifies the transactions as malicious or non-malicious.

Definition 14 (Cluster Deviation Index CDI) Let k be the list of membership probabilities of given transaction X from all the K clusters. Then γ_{\max} can be defined as the maximum membership probability of X among all clusters and the corresponding cluster defines the role profile of user executing X . Finally Cluster Deviation Index (CDI) is formulated as:

$$CDI = 1 - \gamma_{\max} \quad (13)$$

$$where, \gamma_{\max} = \max(\gamma_k) \forall k \in [1, K] \quad (14)$$

In steps 3-4, using the means, covariances and mixing coefficients learnt by the EM clustering Algorithm, membership probabilities for a given transaction with each cluster is computed. In steps 6-7, the maximum of membership probabilities is used to compute Cluster Deviation Index (CDI) to measure the deviation of given transaction from all the clusters. In steps 8-12, Class (C) is assigned to the given transaction based on CDI values obtained in step 7. Here δ is used as a dissimilarity threshold to identify malicious and

non-malicious transactions. If $CDI > \delta$ is greater than, then the transaction is classified as malicious else it is classified as non-malicious.

Algorithm 4 EM classification algorithm.

Data : δ : dissimilarity threshold
 K : number of clusters
 $\mu^{(k)}$: means of K clusters
 $\Sigma^{(k)}$: covariances of K clusters,
 $\pi^{(k)}$: mixing coefficients of K clusters

Input : X : transaction record
Output: C : Class (Malicious or Non-malicious)

```

1 begin
2   Classification Phase
3   for  $k \leftarrow 1$  to  $K$  do
4      $\gamma_k \leftarrow \frac{\pi_k p(X/k)}{\sum_{j=1}^k \pi_j p(X/j)}$ 
5   end
6    $\gamma_{max} \leftarrow \max(\gamma_k)$ 
7    $CDI \leftarrow 1 - \gamma_{max}$ 
8   if  $CDI > \gamma_{max}$  then
9      $C \leftarrow$  malicious
10  end
11  else
12     $C \leftarrow$  non-malicious
13  end
14 end

```

4 Results and discussion

To demonstrate the efficiency of our proposed approach, several experiments were carried out on synthetically generated transactions targeted towards conventional banking databases adhering to TPC-C (Transaction Processing Performance Council 2002) benchmark [58]. The dataset consists of two types of logs — one consisting of malicious user transactions and the other with the normal user transactions carried out by authorized users. In total 25,000 transactions were generated from the two logs to carry out the performance evaluation. In this section, we first detail the transaction generation methodology, followed by the study of the performance of EMSPM on the synthetically generated datasets.

4.1 Synthetic transactions generation

EMSPM took into account the fact that any dataset which conformed to our approach did not exist which prompted us to make effective use of synthetically generated datasets. This dataset contains two different modules for the two different types of transactions that we operated on — malicious transactions generation module and normal transactions generation module, the description and working of which has been given below.

4.2 Malicious transactions generation

To assess our approach on a wide gamut of attacks, we generated two different sets of malicious transactions. First set of transactions were generated randomly with the assumption that the attacker may be unaware of normal database dependencies. This set is generated by replacing attributes randomly of legitimate queries which constitute the usual behavior and by replacing legitimate transactions of each.

The second set of transactions were generated such that they do not conform to the average user activity and are indicators of users trying to perform transactions outside of their scope and permissions.

4.3 Normal transactions generation

Normal transactions are those that satisfy data dependencies and user-access patterns at the time of execution. To generate normal transactions, we define a fixed set of SQL queries which adhere to the TPC-C benchmark and the schema used. Each user is assigned individual attributes that are within the scope of their access and permissions and the corresponding transactions are generated using those attributes thus satisfying the user-access patterns.

4.4 Selection of dissimilarity threshold

In the detection phase, a transaction is classified as malicious depending upon the value of the dissimilarity threshold. To select appropriate value of δ , we try to model our algorithm for different values of δ and compare the performance of our approach plotting graphs for various performance metrics like Precision, Recall, F1-score and Accuracy against the number of transactions..

From Fig. 3 and Table 4 we can observe that for $\delta = 0.12$, the line plot for Precision remains well above all the other plots for different threshold values suggesting it to be the accurate threshold value. For δ less than 0.12, the number of false positives increases, resulting in low value of Precision. For δ greater than 0.12, the model fails to identify many true

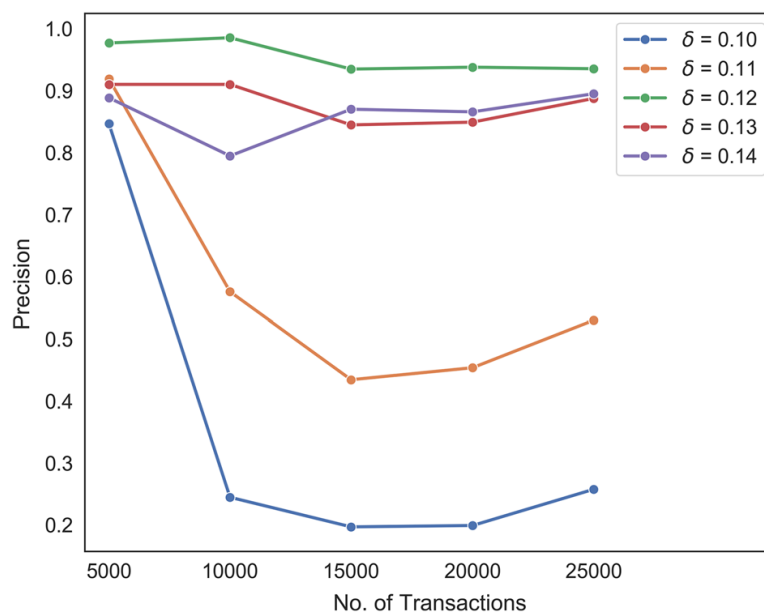


Fig. 3 Variation in Precision with Number of Transactions for different dissimilarity threshold

Table 4 Precision for different dissimilarity threshold δ

δ	Number of Transactions				
	5000	10000	15000	20000	25000
0.10	0.8467	0.2448	0.197	0.1992	0.2577
0.11	0.9191	0.5758	0.4344	0.4538	0.5302
0.12	0.977	0.9855	0.9349	0.938	0.9354
0.13	0.9102	0.9102	0.8448	0.8493	0.8876
0.14	0.8886	0.7948	0.8702	0.8659	0.8951

positives again resulting in low value of precision. Thus if we consider only the Precision, $\delta = 0.12$ performs best for our given model.

In comparison of recall in Fig. 4 and Table 5, though $\delta = 0.11$ comes very close to $\delta = 0.12$, but the best performance is attained by $\delta = 0.12$ when the model is evaluated over different numbers of transactions. This is because $\delta = 0.11$ fails to capture malicious transactions as effectively as $\delta = 0.12$.

Similar to the Precision and Recall, the optimum threshold for best F1-score observed from Fig. 5 and Table 6 is also $\delta = 0.12$ which could be realised from the fact that F1-score is calculated from both Precision and Recall and thus the value of having equally better values of both Precision and Recall will achieve the highest F1-score. Since all the other threshold values had poor results in either Precision or Recall or both, F1-score of $\delta = 0.12$ is well above all other thresholds.

Accuracy of our approach from Fig. 6 and Table 7 was highest in case of $\delta = 0.11$ for lesser number of transactions but as the number of transactions increased, the accuracy with $\delta = 0.11$ drops below 0.90. But for $\delta = 0.12$, the model consistently has an accuracy of over 0.93 with a maximum of 0.98373.

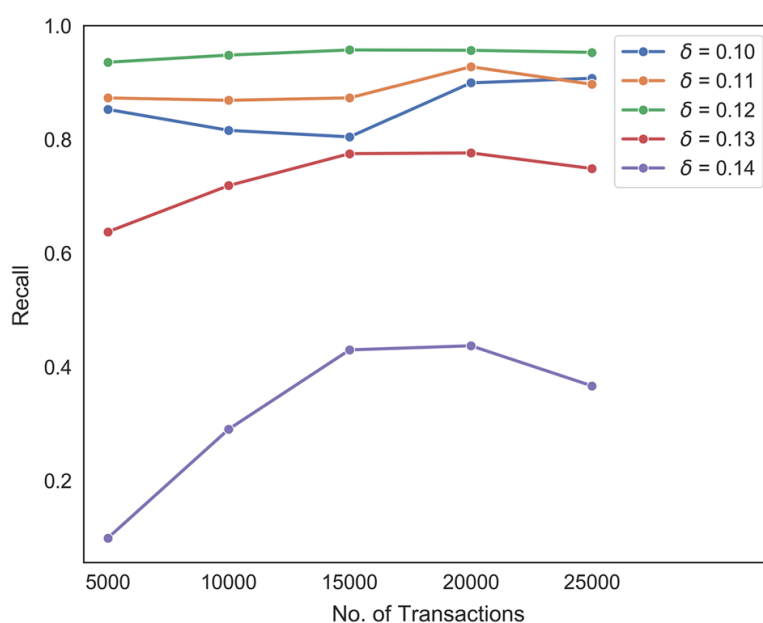
**Fig. 4** Variation in Recall with Number of Transactions for different dissimilarity threshold

Table 5 Recall for different dissimilarity threshold δ

δ	number of Transactions				
	5000	10000	15000	20000	25000
0.10	0.853	0.8162	0.8047	0.8998	0.9079
0.11	0.9361	0.9487	0.9578	0.9571	0.9534
0.12	0.9361	0.9487	0.9578	0.9571	0.9534
0.13	0.6374	0.719	0.7752	0.7765	0.7489
0.14	0.099	0.2907	0.4304	0.4376	0.3669

Therefore, by comparing various performance metrics we reach to a conclusion that $\delta = 0.12$ is the optimal dissimilarity threshold to classify transactions in detection phase for our approach.

4.5 Performance evaluation of our approach

The performance metrics used to carry out the assessment were precision, recall and F1-score. Precision is defined as the ratio of correctly identified malicious transactions known as True Positives (TP) to the total number of transactions classified as malicious from the database logs, a combination of False Positives (FP) and True Positives (TP). Recall is defined as the ratio of correctly identified malicious transactions known as True Positives (TP) to the total number of malicious transactions existing in the database logs, a combination of False Negatives (FN) and True Positives (TP). In case of imbalanced dataset, high precision and low recall as well as low precision and high recall describes a poor model. Thus F1-score is used which takes both precision and recall into account for performance evaluation. F1-score is defined as the harmonic mean of precision and recall.

Figures 7, 8, 9 depicts the variation among Precision, Recall and F1-score with the number of transactions. It can be observed from the graph that Precision first increases

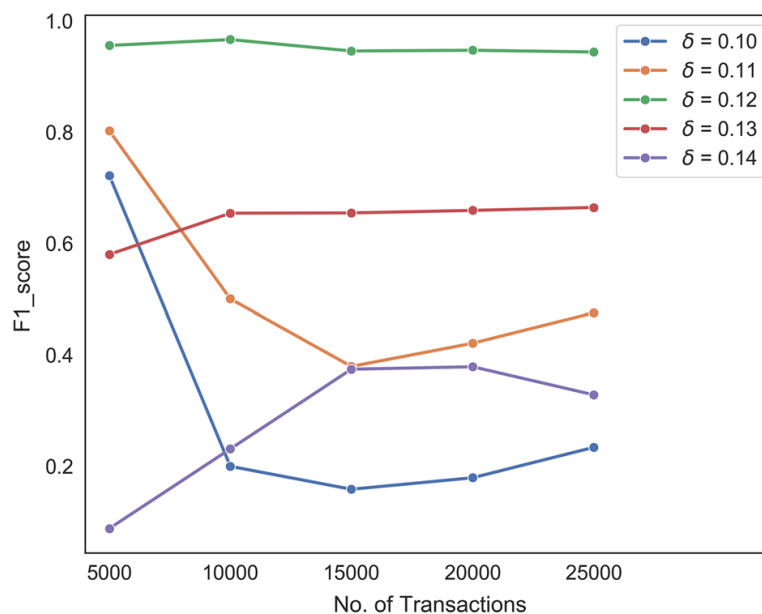
**Fig. 5** Variation in F1-Score with Number of Transactions for different dissimilarity threshold

Table 6 F1-Score for different dissimilarity threshold δ

δ	Number of Transactions				
	5000	10000	15000	20000	25000
0.10	0.7223	0.1998	0.1586	0.1793	0.2339
0.11	0.8026	0.5005	0.3794	0.4212	0.4757
0.12	0.9561	0.9667	0.9462	0.9475	0.9444
0.13	0.5802	0.6544	0.6549	0.6594	0.6647
0.14	0.088	0.2311	0.3745	0.3789	0.3284

from 0.97698 to 0.98546 and then decreases to 0.93544 with further increase in the number of transactions. It can be understood from the fact that as the number of transactions increases, the number of rules generated for each transaction increases which also results in an increased number of False Positives. Though the value of Recall is low for a lesser number of transactions starting from 0.93608, it increases upto 0.95776 as the number of transactions are increased and then attains a stable value. The initial low value of recall can be understood from the fact that with lesser number of transactions, the model fails to accurately identify all the malicious transactions leading to higher False Negatives, hence reducing the value of recall.

Since F1-score is a harmonic mean of Precision and Recall, it depicts a fair balance between precision and recall and stays always between the two values maintaining a value between 0.94435 and 0.96671.

Figure 10 illustrates the variation between accuracy number of transactions with a dissimilarity threshold = 0.12.

Accuracy is defined as the ratio of correctly identified transactions i.e. True Positives and True Negatives (TP+TN) to total number of transactions (TP+FP+TN+FN). Initially with a lesser number of transactions, accuracy observed was 0.93635. But as the number of transactions increased, the accuracy increased gradually upto 0.98373. This can be justified

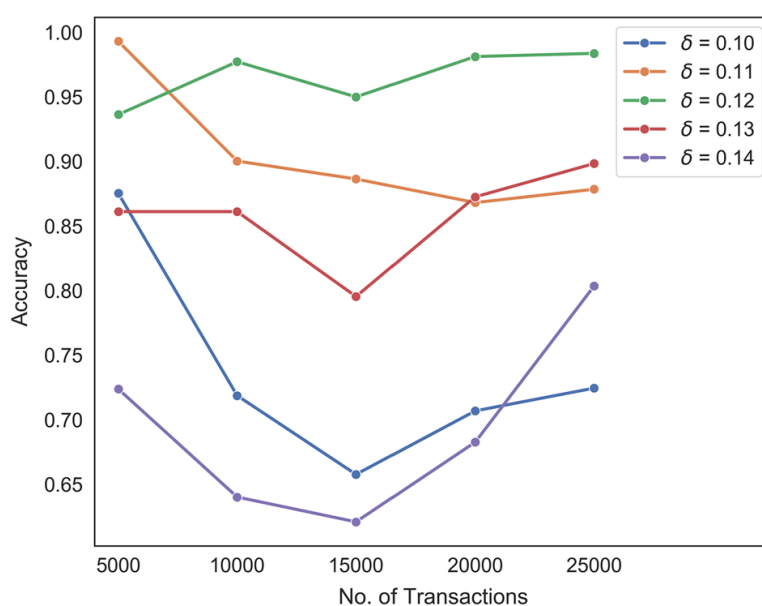
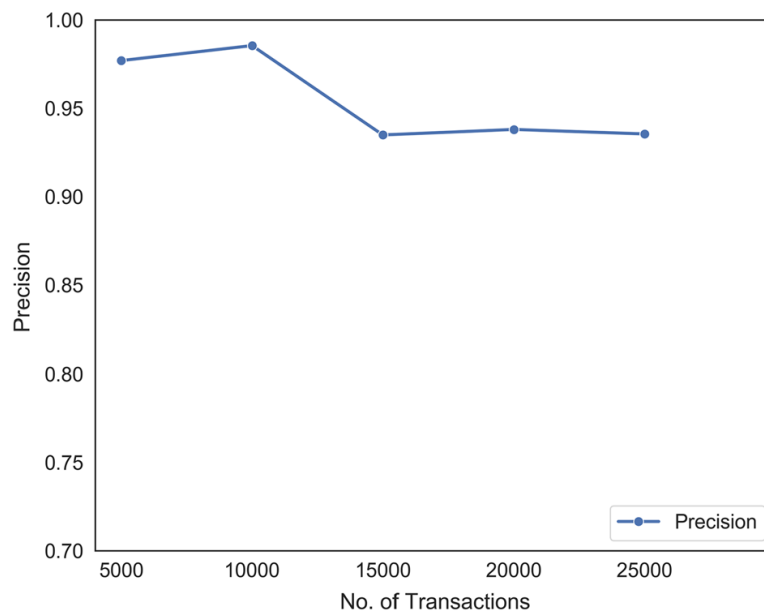
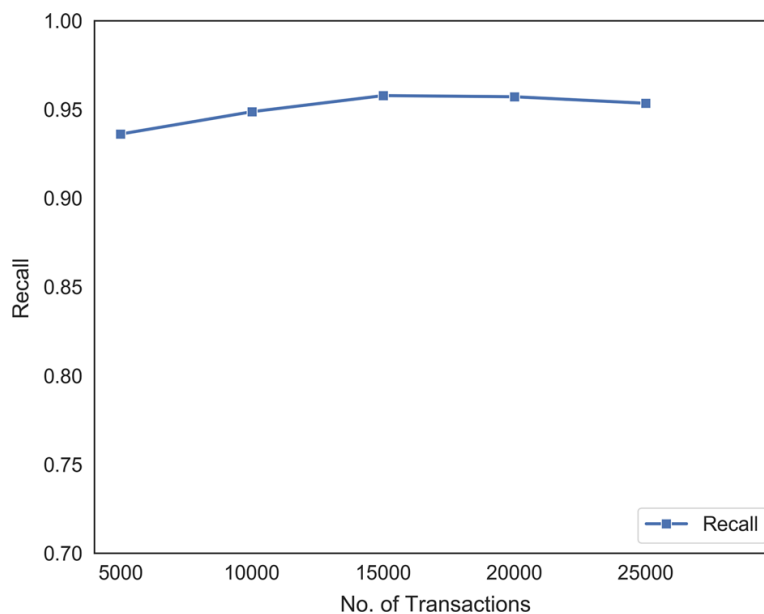
**Fig. 6** Variation in Accuracy with Number of Transactions for different dissimilarity threshold

Table 7 Accuracy for different dissimilarity threshold δ

δ	Number of Transactions				
	5000	10000	15000	20000	25000
0.10	0.8753	0.7187	0.6577	0.7069	0.7245
0.11	.9931	0.9003	0.8865	0.8682	0.8785
0.12	0.9364	0.9772	0.95	0.9813	0.9837
0.13	0.8612	0.8612	0.7954	0.8726	0.8984
0.14	0.7238	0.6401	0.6209	0.6827	0.8035

**Fig. 7** Variation in Precision, Recall and F1-score with Number of Transactions for dissimilarity threshold = 0.12**Fig. 8** Variation in Recall with Number of Transactions for dissimilarity threshold = 0.12

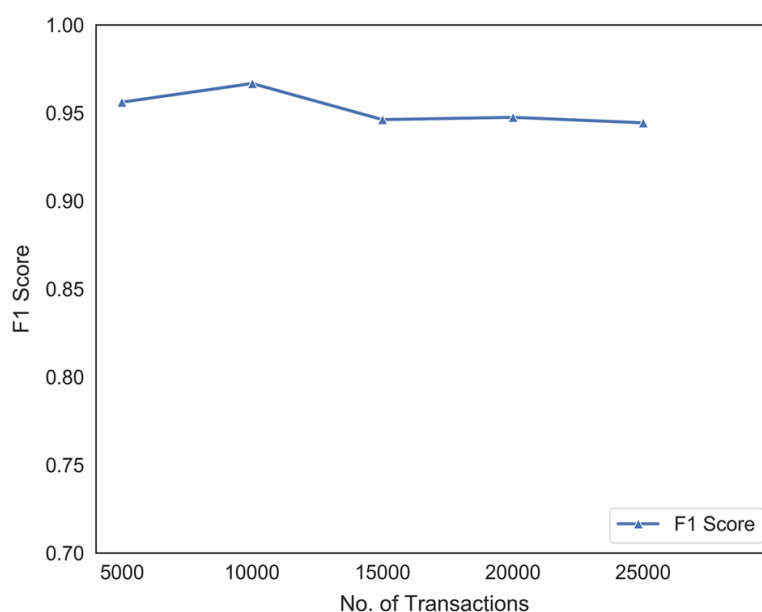


Fig. 9 Variation in F1-score with Number of Transactions for dissimilarity threshold = 0.12

from the fact that with increased number of transactions resulting in larger numbers of rules and well-defined clusters, the model classified both types of transactions pretty accurately. This trend is also justified by the fact that with an increase in the number of transactions available, the model can mine increasingly consistent rules, making it more robust.

Therefore, on a complete dataset our EMSPM approach performs with an accuracy of 0.98373 for 25000 transactions.

4.6 Comparison of IPPS vs EMSPM

Our EMSPM algorithm uses two techniques to classify an incoming transaction as malicious or non-malicious namely IPPS (iteratively Pruned Prefix Span) algorithm and EM

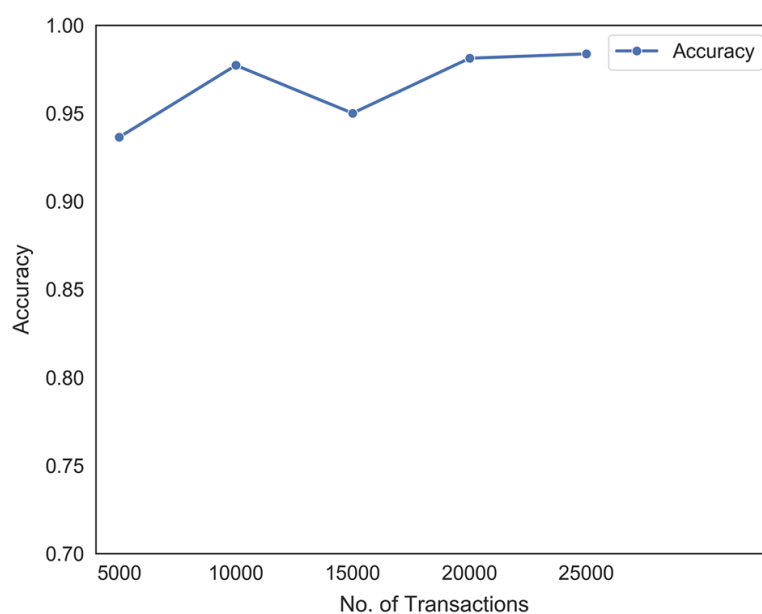


Fig. 10 Variation in Accuracy with Number of Transactions for dissimilarity threshold = 0.12

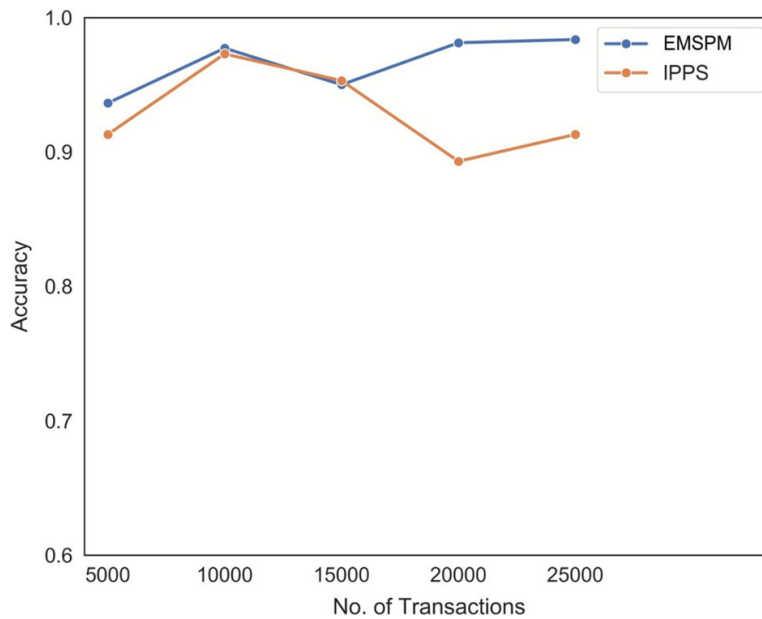


Fig. 11 Comparison of IPPS and EMSPM for variation in Accuracy with Number of Transactions

(Expectation maximization) clustering. In our EMSPM algorithm, the first step is to check whether a transaction adheres to rules generated by IPPS, if not an alarm is raised and it is classified as malicious, otherwise if IPPS rules fail to identify the transaction as malicious, then EMSPM uses EM classification algorithm to evaluate the transaction for user-access patterns and detect the transaction as malicious or not.

Figure 11 demonstrates the effectiveness of the EMSPM algorithm over the IPPS module. It can be seen that the EM algorithm improves the classification capabilities of the IPPS algorithm significantly. If the classification is done purely with IPPS algorithms the overall accuracy for 25000 transactions is 0.913. This is due to the fact that in the IPPS algorithm, if a transaction that complies with the data dependency rules can be classified as non-malicious without checking for user-access patterns that govern a particular role profile.

EMSPM combines both IPPS generated data dependency rules as well as role profiles clusters generated by the usage of an EM clustering algorithm that can determine whether the user executing the transaction adheres to the access privileges allocated or performs transactions outside the scope of their roles. This combined approach increases the accuracy of classification from 0.913 of the IPPS to 0.9837 of EMSPM for the complete dataset.

Table 8 contrasts various techniques based on the algorithm used, the ability to avoid intrusion, and the performance metrics - precision and recall. The performance was evaluated for different support values for each technique considered and maximum values were taken for precision and recall. The output values show that our algorithm performs better than alternative methods. These improvements can be attributed to low sensitivity of our algorithm to change in user patterns due to our consideration of relative adherence to data dependencies.

This is expressed in the transactions in the real world where the transactions never fully comply with the data dependencies. Our algorithm has the benefits of both statistical-based detection and anomaly-based detection methods and hence, we are able to reduce both the false positive and false negative errors.

Table 8 Performance comparison of EMSPM with other techniques

Author Names	Technique	Command Syntax	Scalability	RBAC Used	Anomaly Detection	Approach for query evaluation	Intrusion Detection Capabilities	Performance
Hu et al. [60]	Integrated dependency with sequence alignment analysis	✓	×	×	×	Syntax Centric	Partial	Recall = 0.90 Precision = 0.64
Srivastava et al. [55]	Integrated dependency with sequence alignment analysis	✓	×	×	×	Syntax Centric	Partial	Recall = 0.77 Precision = 0.80
Panigrahi et al [40]	User Behavior Mining	✓	✓	✓	×	Syntax Centric	Partial	Recall = 0.93 Precision = 0.91
Hashemi et al [20]	Temporal Mining	✓	×	×	×	Syntax Centric	Yes	Recall = 0.90 Precision = 0.75
Kamra et al [24]	Dependency And Relation Analysis	✓	✓	✓	✓	Syntax Centric	Yes	Recall = 0.63 Precision = 0.77
Sohrabi et al. [53]	Mining dependencies	✓	×	×	×	Data Centric	Yes	Recall = 0.65 Precision = 0.77
Doroudian et al. [17]	Mining dependencies	✓	×	×	×	Syntax Centric	Yes	Recall = 0.89 Precision = 0.91
Ronao and Cho [45]	Weighted Random Forest	✓	✓	✓	✓	Syntax Centric	Yes	Recall = 0.95 Precision = 0.89
Sallam et al [47]	Anomaly Detection using Bayesian Classifier	✓	×	✓	✓	Data and Syntax Centric	Yes	Recall = 0.97 Precision = 0.90
Seok-Jun et al. [7]	Convolutional Neural-based Classifier	✓	✓	✓	×	Data Centric	Yes	Recall = 0.93 Precision = 0.92
Sharmila Subudhi et al. [56]	OPTICS Clustering with Ensemble Learning	✓	×	×	✓	Data Centric	Yes	Recall = 0.92 Precision = 0.95
EMSPM	Mining Dependencies and transaction role profiling with EM Clustering	✓	✓	✓	✓	Data and Syntax Centric	Yes	Recall = 0.95 Precision = 0.97

5 Conclusion and future work

In this paper, we presented a DIDS which can safeguard a database from insider as well as external threats, and in general prevent it from attacks by users that are unacquainted with the normal data dependencies between the data items and the intricate syntactic features of legitimate transactions. Our intrusion detection system incorporates a Rule Mining module and an Expectation maximization (EM) Clustering module. The Rule Mining module mines user information access patterns using modified Prefixspan and the Expectation maximization (EM) Clustering module creates unique profiles of intrusion-free transactions by clustering the user activity parameters from information logs. The incoming transactions are assessed against the two levels and the extent of conformity to the mined rules together with membership of the present user profile within the role profile clusters work to classify the transaction as either malicious or non-malicious.

Considering only those frequent patterns whose lengths were at least three to remove redundant information resulted in a significant decrease in the number of false positives and increase in the overall performance. Furthermore, the use of EM clustering, which is widely popular due to its low computation time and high accuracy increased the overall efficiency of our approach. So the use of EM clustering makes the model not only time efficient but also cost efficient. The cost efficiency makes the model accessible to be open sourced and also available to all stratas requiring this intrusion detection technique. In future, we will investigate a lot of refined options for incorporating user behavior and improvise methods for analysing knowledge dependencies for mining patterns together with the employment of real-world knowledge. The future work be more efficient in overcoming present limitations as well like investigating more sophisticated features for login user behavior which were not present in this. Our future work will emphasize on an extra careful consideration of sensitivity of operations, which may enhance the performance of the system

References

1. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proc. 20th int. conf. very large data base, VLDB, vol 1215, pp 487–499
2. Agrawal R, Srikant R (1995) Mining sequential patterns. In: Proceedings of the eleventh international conference on data engineering, pp 3–14
3. Assaad HE, Samé A, Govaert G, Aknin P (2016) A variational expectation–maximization algorithm for temporal data clustering. *Comput Stat Data Anal* 103:206–228
4. Bertino E, Sandhu R (2005) Database security-concepts, approaches, and challenges. In: *IEEE Transactions on Dependable and secure computing* 2.1, pp 2–19
5. Bertino E, Terzi E, Kamra A, Vakali A (2005) Intrusion detection in RBAC-administered databases. In: 21st Annual computer security applications conference (AC-SAC'05), IEEE, 10–pp
6. Bilmes JA et al (1998) A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. In: *International computer science institute* 4.510, p 126
7. Bu S-J, Cho S-B (2020) A convolutional neural-based learning classifier system for detecting database intrusion via insider attack. *Inf Sci* 512:123–136
8. Cappelli DM, Moore AP, Trzeciak RF (2012) The CERT guide to insider threats: how to prevent, detect, and respond to information technology crimes (Theft Sabotage Fraud). Addison-Wesley
9. Cárdenas AA, Amin S, Lin Z-S, Huang Y-L, Huang C-Y, Sastry S (2011) Attacks against process control systems: risk assessment, detection, and response. In: *Proceedings of the 6th ACM symposium on information, computer and communications security*, pp 355–366
10. Chen M-S, Han J, Yu PS (1996) Data mining: an overview from a database perspective. *IEEE Trans Knowl Data Eng* 8.6:866–883

11. Chung CY, Gertz M, Levitt K (1999) Demids: A misuse detection system for database systems. In: Working conference on integrity and internal control in information systems, Springer, pp 159–178
12. Corney MW, Mohay GM, Clark AJ (2011) Detection of anomalies from user profiles generated from system logs. In: Conferences in research and practice in information technology (CRPIT). vol. 116, Australian Computer Society, Inc. pp 23–32
13. Debar H, Dacier M, Wespi A (1999) Towards a taxonomy of intrusion-detection systems. *Comput Netw* 31.8:805–822
14. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J Royal Stat Soc Ser B Methodol* 39.1:1–22
15. Denning DE (1987) An intrusion-detection model. *IEEE Trans Softw Eng* 2:222–232
16. Do CB, Batzoglu S (2008) What is the expectation maximization algorithm? *Nature biotechnol* 26.8:897–899
17. Doroudian M, Shahriari HR (2014) A hybrid approach for database intrusion detection at transaction and inter-transaction levels. In: 2014 6th Conference on information and knowledge technology (IKT), IEEE, pp 1–6
18. Han J, Pei J, Mortazavi-Asl B, Chen Q, Dayal U, Hsu M-C (2000) FreeSpan: frequent pattern-projected sequential pattern mining. In: Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining. pp 355–359
19. Han J, Pei J, Mortazavi-Asl B, Pinto H, Chen Q, Dayal U, Hsu M (2001) Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In: Proceedings of the 17th international conference on data engineering. Citeseer, pp 215–224
20. Hashemi S, Yang Y, Zabihzadeh D, Kangavari M (2008) Detecting intrusion transactions in databases using data item dependencies and anomaly analysis. *Expert Syst* 25.5:460–473
21. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media
22. Heady R, Luger G, Maccabe A, Servilla M (1990) The architecture of a network level intrusion detection system. Tech. rep. Los Alamos National Lab., NM (United States); New Mexico Univ. Albuquerque...
23. Hoglund AJ, Hatonen K, Sorvari AS (2000) A computer host-based user anomaly detection system using the self-organizing map. In: Proceedings of the IEEE-INNS-ENNS international joint conference on neural networks. IJCNN 2000. neural computing: new challenges and perspectives for the new millennium. vol. 5. IEEE, pp 411–416
24. Kamra A, Terzi E, Bertino E (2008) Detecting anomalous access patterns in relational databases. *VLDB J* 17.5:1063–1077
25. Kim T-Y, Cho S-B (2019) CNN-LSTM neural networks for anomalous database intrusion detection in RBAC-administered model. In: International conference on neural information processing, Springer, pp 131–139
26. Kuang F-J, Zhang S-Y (2017) A Novel Network Intrusion Detection Based on Support Vector Machine and Tent Chaos Artificial Bee Colony Algorithm. *J Netw Intell* 2.2:195–204
27. Lan G-C, Hong T-P, Lee H-Y (2014) An efficient approach for finding weighted sequential patterns from sequence databases. *Appl Intell* 41.2:439–452
28. Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions, and reversal. *Soviet Physics doklady* 10. 8.:707–710
29. Liao H-J, Lin C-HR, Lin Y-C, Tung K-Y (2013) Intrusion detection system: A comprehensive review. *J Netw Comput Appl* 36.1:16–24
30. Lin JC-W, Fournier-Viger P, Koh YS, Kiran RU, Thomas R (2017) A survey of sequential pattern mining. *Data Sci Pattern Recogn* 1.1:54–77
31. Liu P-Y, Gong W, Jia X (2011) An improved prefixspan algorithm research for sequential pattern mining. In: 2011 IEEE international symposium on IT in medicine and education. vol. 1, IEEE, pp 103–108
32. Lunt TF, Tamaru A, Gillham F (1992) A real-time intrusion-detection expert system (IDES). SRI International Computer Science Laboratory
33. Luo C, Chung SM (2005) Efficient mining of maximal sequential patterns using multiple samples. In: Proceedings of the 2005 SIAM international conference on data mining. SIAM, pp 415–426
34. Mazzawi H, Dalal G, Rozenblat D, Ein-Dor L, Ninio M, Lavi O (2017) Anomaly detection in large databases using behavioral patterning. In: 2017 IEEE 33rd international conference on data engineering (ICDE). IEEE, pp 1140–1149
35. McLachlan GJ, Krishnan T (2007) The EM algorithm and extensions, vol 382. Wiley, New York
36. Mitra P, Pal SK, Siddiqi MA (2003) Non-convex clustering using expectation maximization algorithm with rough set initialization. *Pattern Recogn Lett* 24.6:863–873
37. Neal RM, Hinton GE (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Learning in graphical models, Springer, pp 355–368

38. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48.3:443–453. issn: 0022-2836. [https://doi.org/10.1016/0022-2836\(70\)90057-4](https://doi.org/10.1016/0022-2836(70)90057-4). <http://www.sciencedirect.com/science/article/pii/0022283670900574>
39. Ordonez C, Omiecinski E (2002) FREM: fast and robust EM clustering for large data sets. In: Proceedings of the eleventh international conference on Information and knowledge management, pp 590–599
40. Panigrahi S, Sural S, Majumdar AK (2013) Two-stage database intrusion detection by combining multiple evidence and belief update. *Inform Syst Front* 15.1:35–53
41. Pei J, Han J, Mortazavi-Asl B, Wang J, Pinto H, Chen Q, Dayal U, Hsu M-C (2004) Mining sequential patterns by pattern-growth: The prefixspan approach. *IEEE Trans Knowl Data Eng* 16.11:1424–1440
42. Rahman MM, Ahmed CF, Leung CK-S (2019) Mining weighted frequent sequences in uncertain databases. *Inf Sci* 479:76–100
43. Rahman MM, Ahmed CF, Leung CK, Pazdor AGM (2018) Frequent sequence mining with weight constraints in uncertain databases. In: Proceedings of the 12th international conference on ubiquitous information management and communication, pp 1–8
44. Rashid T, Agrafiotis I, Nurse JRC (2016) A new take on detecting insider threats: exploring the use of hidden markov models. In: Proceedings of the 8th ACM CCS international workshop on managing insider security threats, pp 47–56
45. Ronao CA, Cho S-B (2016) Anomalous query access detection in RBAC-administered databases with random forest and PCA. *Inf Sci* 369:238–250
46. Sallam A, Bertino E (2019) Result-based detection of insider threats to relational databases. In: Proceedings of the ninth ACM conference on data and application security and privacy, pp 133–143
47. Sallam A, Fadolalkarim D, Bertino E, Xiao Q (2016) Data and syntax centric anomaly detection for relational databases. In: Wiley interdisciplinary reviews: data mining and knowledge discovery 6.6, pp 231–239
48. Sandhu RS, Coyne EJ, Feinstein HL, Youman CE (1996) Role-based access control models. *Computer* 29.2:38–47
49. Sandhu R, Ferraiolo D, Kuhn R et al (2000) The NIST model for role-based access control: towards a unified standard. In: ACM workshop on Role-based access control. Vol. 10. 344287.344301
50. Shirkhorshidi AS, Aghabozorgi S, Wah TY (2015) A comparison study on similarity and dissimilarity measures in clustering continuous data, *PloS one* 10.12
51. Shou Z, Di X (2018) Similarity analysis of frequent sequential activity pattern mining. *Trans Res Part C Emerg Technol* 96:122–143
52. Smith TF, Waterman MS et al (1981) Identification of common molecular subsequences. *J Mol Biol* 147.1:195–197
53. Sohrabi M, Javidi MM, Hashemi S (2014) Detecting intrusion transactions in database systems: a novel approach. *J Intell Inf Syst* 42.3:619–644
54. Srikant R, Agrawal R (1996) Mining sequential patterns: Generalizations and performance improvements. In: International conference on extending database technology, Springer, pp 1–17
55. Srivastava A, Sural S, Majumdar AK (2006) Database intrusion detection using weighted sequence mining. *J Comput* 1.4:8–17
56. Subudhi S, Panigrahi S (2019) Application of OPTICS and ensemble learning for database intrusion detection. In: Journal of king saud university-computer and information sciences
57. Talpade R, Kim G, Khurana S (1999) NOMAD: Traffic-based network monitoring framework for anomaly detection. In: Proceedings IEEE international symposium on computers and communications (Cat. No. PR00250). IEEE, pp 442–451
58. TPC-C Benchmark. <http://www.tpc.org/tpcc/default.asp>
59. Yi H, Brajendra P (2003) Identification of malicious transactions in database systems. In: Seventh international database engineering and applications symposium, 2003 Proceedings. IEEE, pp 329–335.
60. Yi H, Brajendra P (2004) A data mining approach for database intrusion detection. In: Proceedings of the 2004 ACM symposium on applied computing, pp 711–716
61. Yip RW, Levitt EN (1998) Data level inference detection in database systems. In: Proceedings. 11th IEEE computer security foundations workshop (Cat. No. 98TB100238). IEEE, pp 179–189
62. Zahedeh Z, Feizollah A, Anuar NB, Kiah LBM, Srikanth K, Kumar S (2019) User profiling in anomaly detection of authorization logs. In: Computational science and technology. Springer, pp 59–65



Indu Singh is an Assistant Professor in Computer Science Engineering Department at the Delhi Technological University, Delhi. Singh has B.TECH in Computer Science & Engineering and an M.TECH degree in Information Security from Ambedkar Institute of Advanced Communication Technologies & Research, Guru Gobind Singh Indraprastha University, Govt. of NCT Delhi. She is currently pursuing her Ph.D in Computer Science and Engineering, specializing in Data Mining and Information Security at CSE Department, Delhi Technological University. Her research interests include Database Systems, Data Mining, Information Security, Machine Learning, Fuzzy systems, Swarm Intelligence and Pattern Recognition. She has published papers in International conferences and Journals of IEEE, Elsevier, Springer and ACM. She has also received IEEE Best Paper Award in ICACCI-2016. Singh has also served as a reviewer for several conferences of IEEE and Springer in India and abroad. Her e-mail is indusingh@dtu.ac.in



Dr. Rajni Jindal is currently heading the Department of Computer Science & Engineering at Delhi Technological University (erstwhile Delhi College of Engineering). She is working here as faculty since 1992. She completed her PhD (Computer Engineering) from Faculty of Technology, Delhi University in the area of Data Mining. She received her M.E. (Computer Technology & Applications) degree from Delhi college of Engineering. Jindal joined Indira Gandhi Technical University for Women (IGDTUW) as Professor in 2012 on lien. She worked as Head (IT) and Dean (Research & Collaboration) at IGDTUW till Feb 2015 before returning back to DTU. Her major areas of interest are Database Systems, Data Mining and Operating systems. She has supervised more than 55 ME/MTech thesis. Jindal has authored/co-authored around 80 research papers and articles for various international journals/conferences. There are 12 Ph.D students working under her supervision and 4 Ph.D have already been awarded to her students. She has completed AICTE sponsored project in the area of education data mining as Co- Principal Investigator. She has authored books on “Data Structures using C” and “Compiler-Construction and Design”. She has successfully organized IEEE International Conferences ICDMIC-2014 and IICIP 2016. She is a life member of professional bodies like CSI, ISTE and Senior member of IEEE, USA. Her e-mail is rajnijindal@dce.ac.in

Fast Under Water Image Enhancement for Real Time Applications

Aruna Bhat

Department of Computer Science and
Engineering
Delhi Technological University
Delhi, India
aruna.bhat@dtu.ac.in

Aadhar Tyagi

Department of Software Engineering
Delhi Technological University
Delhi, India
tyagi.aadhar@gmail.com

Aarsh Verdhan

Department of Software Engineering
Delhi Technological University
Delhi, India
aavrdhen123@gmail.com

Vaibhav Verma

Department of Software Engineering
Delhi Technological University
Delhi India
vaibhavv1904@gmail.com

Abstract— Ocean exploration is a major challenge that we are facing today. With advancements to fields of marine engineering and aquatic robotics, we are capable of performing autonomous and complex decision making deep underwater. Significance of online Underwater Computer Vision Algorithms is ever increasing. Underwater images, however, suffer from inaccurate colors, hazing, colour cast and degradations because of unequal absorption of light by water. Algorithms designed for detection/enhancement in the air are of no use underwater. Although a lot of underwater image enhancement algorithms have come up in recent times, most of them are not suitable for real-time applications like AUV, due to their high computational times. These algorithms are more suitable for offline analysis. In this paper, we propose an algorithm which is fast enough for real-time systems such as AUVs/ROVs and is comparable to the offline state of the art image enhancement algorithms. We will be exploring histogram equalization techniques for dehazing and automatic white balancing algorithms for color correction. UIEB (Underwater Image Enhancement Benchmark) is used for evaluation of our algorithm. The codes and results are available at <https://github.com/opgp/underwater-image-processing>.

Keywords— *underwater image enhancement, real-time, color-cast, CLAHE, white balancing*

I. INTRODUCTION

In the last two decades an exponentially growing interest has been observed in ocean exploration and marine robotics. This has led to significant advancements in the field of aquatic robotics enabling them to perform increasingly challenging tasks autonomously underwater [1]. Underwater image processing thus becomes an essential area of research, such algorithms are deployed in AUVs (Autonomous underwater vehicles). At present, underwater image processing algorithms are used in underwater mine detection [2], submerged robots [3], underwater imaging[4], underwater archaeology [5], ocean basement mapping [6], some of the commercial devices such as cameras, video cameras are also integrated with such application. Even though a lot of exceptional methods for underwater imaging have been proposed. Most of the methods are suitable to be performed on the recordings for research purposes rather than online applications due to the computational power and



Fig. 1. Proposed method for underwater image enhancement

time required [7]. Transportation characteristics of light in underwater environment makes image enhancement challenging problem.

Water is hundreds of times denser compared to the air. As the light passes through water a lot of energy gets attenuated which results in low colour and contrast of underwater images thus resulting in distortion of information about the image [8]. To tackle this problem, the use of artificial lighting was proposed but that produced another problem. Artificial lighting produces a bright spot at the centre of the image and the intensity reduces as we move away from the centre which results in non-uniform illumination[6]. Apart from non-uniform lighting conditions, light's unique absorption and scattering characteristics cause the degradation of underwater images.

Absorption: Specific wavelengths are absorbed at different depths. The red colour is absorbed much more than green and blue colours at a much lesser depth. Resulting in a blue or green colour cast, in most underwater photos [9].

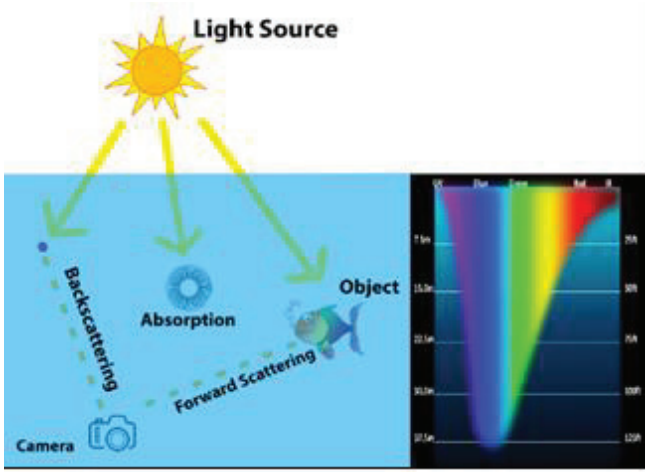


Fig. 2. Degradation of image underwater through scattering and absorption.

Scattering: Suspended particles underwater are much larger than particles found in the air. Leading to incident light reflected from objects to be scattered from particles and resulting in dull images. Contrast and edges are lost due to this phenomenon [10].

Backscattering: Artificial light might illuminate those suspended particles as well, resulting in a lot of noise in underwater images making tasks such as segmentation challenging.

Marine Snow: Macroscopic remains of organic matter from living organisms or inorganic matter present in underwater images from oceans. This results in added noise [11].

The main contributions of the paper are summarized as follows:

- One of the fastest underwater image enhancement method for real-time applications and computationally light and suitable to be run on CPU (Central Processing Unit) only.
- An online enhancement algorithm comparable to state-of-the-art methods that are used for offline analysis. Comparison based on UIEB (underwater image enhancement benchmark) [7].
- Method capable of removing colour casts, enhance colours, boost contrast and dehazing, with minimal parameters to be manually trained.

II. EXISTING METHODOLOGIES

Exploring the underwater world has become an active issue in recent years. Underwater image enhancement is gaining more and more attention in the research field [12]–[14]. We can classify the types of underwater image enhancement methods into four groups.

A. External Hardware-based Methods

To improve the visibility of underwater images, these models use the supplementary information from multiple images, special cameras/filters, like:

- Polarization filtering [15].
- Range-gated imaging [16]–[18].
- Fluorescence imaging [19].

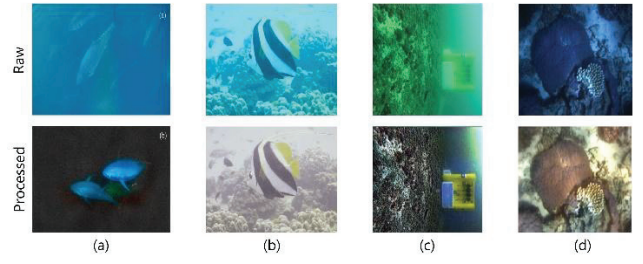


Fig. 3. Different types of Methods (a) Range Gated Underwater Imaging (b) White Balancing based enhancement (c) DCP based enhancing (d) WaterGan

These models are not suitable for challenging situations like dynamic scenes, real-time systems, etc. In these situations, more versatile underwater image enhancement is more suitable.

B. Pixel Value Manipulation Methods:

In these methods, the image pixel value is modified in order to improve the underwater image quality. These are some examples where these methods showed good results:

- K. Iqbal et al. in their paper enhanced the saturation and contrast of an underwater image by stretching the pixel range of HSV and RGB colour space [20].
- Ghani and Isa in their papers modified the work discussed in the last point and reduced the over/enhanced regions. They achieved this by reshaping the stretching process and followed the Raleigh distribution [21], [22].
- Ancuti, C. O. Ancuti, and P. Bekaert in their paper introduced a method for underwater image enhancement in which they blended a colour-corrected image and a contrast-enhanced image in a multi-scale fusion strategy [23], [24].
- X. Fu, Z. Fan, and M. Ling proposed a two-step method which included these algorithms: Contrast Enhancement and Colour Correction [25].
- X. Fu, P. Zhang, Y. Huang presented a retinex model-based approach for underwater image enhancement [26].
- Zhang et al. did research on an extended multiscale retinex-based enhancement model for underwater images [27].

C. Physical Modelling based Methods

In the context of the physical model-based methods for underwater image enhancement, the problem is not as simple as to remove the unwanted properties from the image itself, here we use the image as the source to generate latent parameters. These methods solve the problem of underwater degradation by modelling the underwater environment and applying operations to reverse those degradations.

This problem is generally solved by the same set of methods –

- 1) A physical model for the given degradation is built.
- 2) Unknown variables for the model are then estimated.

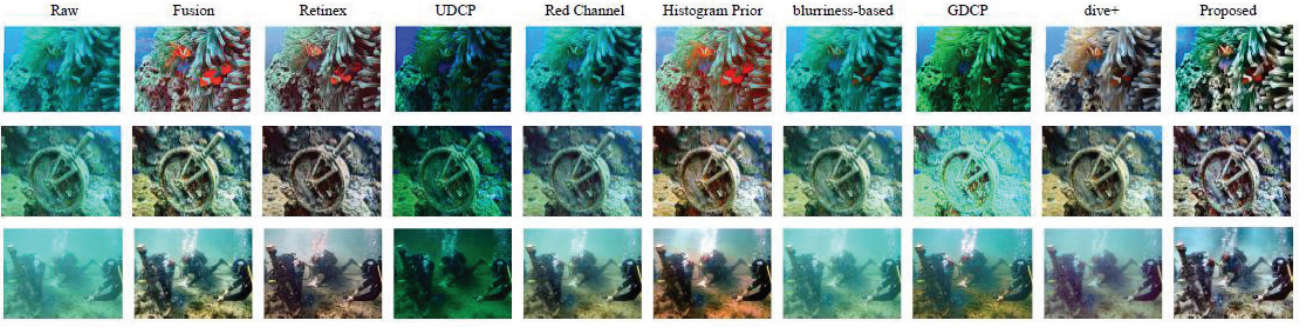


Fig. 4. Comparison of Different Methodologies, from left to right are raw images, fusion based, retinex based, UDCP, Red Channel, Histogram Prior, Bluriness based, GDCP, dive+ and finally the proposed method.

- 3) After the previous two methods the focus is played on the problem of generating latent parameters from the input image.

The research is going on to solve this problem and we have a few methods at hand that have been well researched in order to counter these inverse problems. One such method is to tweak DCP (Dark Channel Prior). Depth map of the underwater image was obtained using median filter in [28]. Then DCP was used for dehazing of the image. Due to the greatest absorption of red colour, loss of information is a problem in underwater images, a solution UDCP [29] (Underwater Dark Channel Prior) was formulated. It was observed that the dark channel for an image that was captured underwater tends towards a zero map Liu and Chau minimized a cost function with the aim of maximizing contrast in the image by formulating an optimal transmission map. GDCP (generalized dark channel prior) was introduced by Peng in order to restore the images by using an image formation model with the help of adaptive colour correction [30].

D. Artificial Intelligence and Machine Learning Methods

Recently there has been an increase in a shift towards Deep Learning algorithms for problems related to low-level computer vision. For training of a CNN (Convolutional Neural Network), original and ground truth images are needed. Since its almost impossible to obtain ground truth images of underwater objects, the only option left is to synthetically generate underwater images from ground truth images. Underwater images depend on temperature, depth and even turbidity of water, hence the deep learning algorithms based on underwater images cannot get the same success as in other low-vision problems.

WaterGan [31], a deep learning-based algorithm was recently proposed. WaterGAN algorithm works by taking the images captured underwater and stimulates it in air image along with the depth pairing using an unsupervised pipeline. The authors for the algorithm created a two-staged network for restoration of images especially for removing the colour casts. An UWCNN[32] (underwater CNN) that was trained using ten types of images captured underwater was proposed. The training images were synthesized using underwater scene variables using an image formation model. Water CycleGAN [33] model was recently proposed on the basis of Cycle Consistent Adversarial networks. This model eliminates the need for paired images in training dataset cause of its network architecture. Thus, allowing the training images to be taken in remote locations. However, the results produced in some cases aren't fully authentic due to multiple possible outputs. Hence the robustness of Deep learning

algorithms for underwater image enhancement is still lagging.

III. PROPOSED METHODOLOGY

This section contains the proposed algorithm for Fast Underwater Image Enhancement. The algorithm can be divided into three major parts, the first being enhancement of the colours in the image by splitting the image into RGB channels and applying adaptive contrast correction on individual channels (to enhance colours lost due to absorption of light underwater). The second part is enhancing the contrast and dehazing the image on the luminance channel. This is achieved by converting the colour channels to YCbCr from RGB, to preserve enhanced colours and applying Adaptive Contrast Correction on the Luminance channel (Y) (to equalize brightness in the image). A similar technique was used in [34], where they used contrast stretching instead of CLAHE. The third step is related to the smoothening of the images and removal of colour cast. Smoothening is performed using denoising algorithms (to remove the noise in the image due to the backscattering and marine snow) after converting the image back to RGB colour space. The colour cast if present if removed based on automatic white balancing by histogram stretching technique (to remove blue or green tint due to the absorption of light underwater). A similar technique was used in [35]. Results generated can be further used for other real-time applications like object detection, segmentation and so on. Now the following subsections will be discussing all the steps and algorithms used in detail.

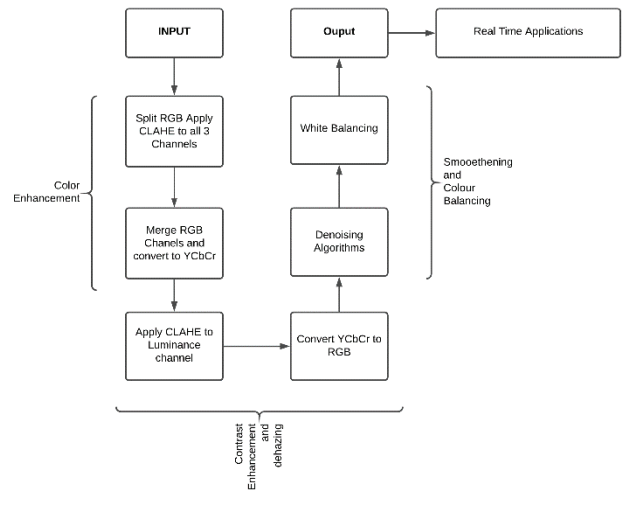


Fig. 5. Proposed Methodology

A. Colour Enhancement

The colours in the images captured underwater have unnatural variations in the colour intensities due to the absorption and scattering of light underwater, these variations result in problems such as dull edges, loss of colours and colour cast. To remove these unwanted effects from the image, it will be operated on in the RGB colour space in this part.

The first step will be to split the RGB components of the image into separate channels and to apply Adaptive Contrast Correction using an approach called CLAHE [36] (Contrast Limited Histogram Equalization) on the individual channels. Where window size = 8×8 and clip limit = 1.

Contrast Limited Histogram Equalization (CLAHE):

Histogram equalization is a technique of distributing intensities throughout a given range. However, instead of taking input from the complete image and generating an equalization function the AHE (adaptive histogram equalization) method is a better alternative, it generates different histograms for different parts of the image and then equalizes the contrast based on those values.

However, AHE is susceptible to noise Amplification in some cases where regions are relatively homogeneous. Hence variant of Adaptive Histogram Equalization called CLAHE will be used. AHE may result in overamplification, to overcome this CLAHE clips the histogram at some value before computing the CDF.

The procedure used for implementing CLAHE is adopted from [36]. Since the intensity values can lie in the range from 0 to 255, let F_k be defined as the frequency of pixel intensity k in the image, then

$$F_k = n_k ; 0 \leq k \leq 255 \quad (1)$$

Where n_k = number of pixels with an intensity value k

Now, cdf (cumulative distribution function) of intensity value at x,y will be calculated.

$$cdf_{I(x,y)} = \sum_k^{I(x,y)} F_k \quad (2)$$

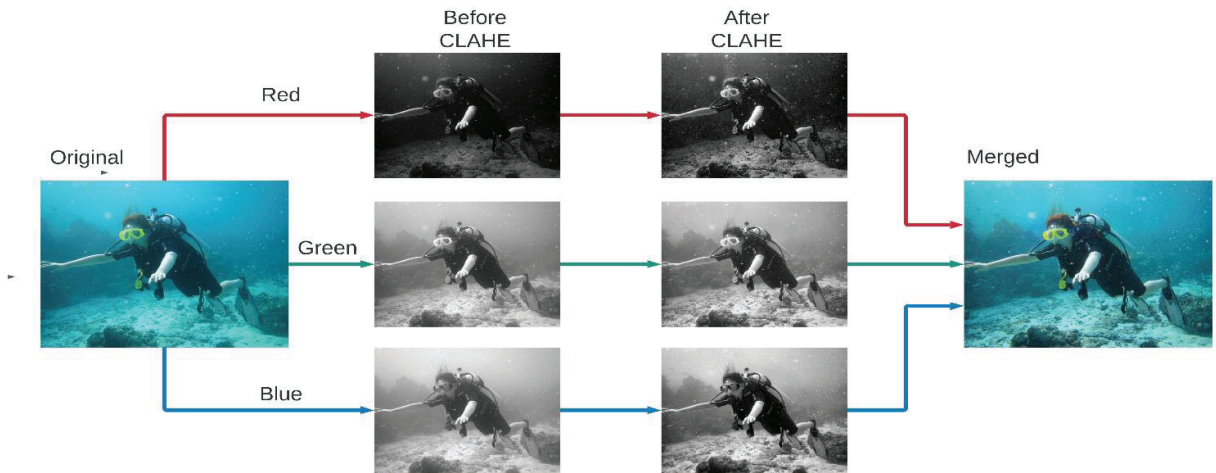


Fig. 7. CLAHE on RGB Channels

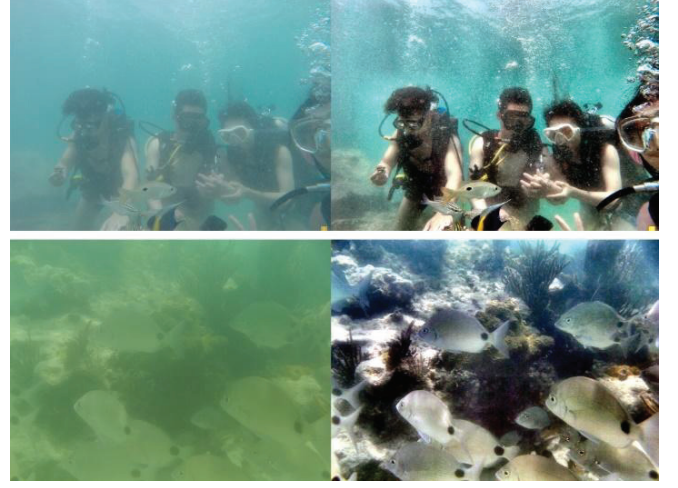


Fig. 6. Effects of dehazing; Left column: Raw; Right Column: processed

Where $x = 1$ to M (number of rows), $y = 1$ to N (number of columns) and $I(x,y)$ = intensity value at x,y .

Now, calculating Histogram equalized Intensity value for each x,y .

$$I'(x,y) = \left\{ \frac{(cdf_{I(x,y)} - cdf_{min})}{(M \times N - cdf_{min})} \times 255 \right\} \quad (3)$$

Where, cdf_{min} is the minimum cdf value for the segment. For implementing CLAHE a clip limit of 1 is put to avoid over-amplification of noise. After equalization of the intensities across all the channels they will be merged in the image. Now, as it can be seen in Fig. 7 since the red channel has low pixel intensities (darker image), CLAHE increased the intensities (made the single channel image brighter). Blue channel had a lot of high intensities, so the single-channel image got darker after CLAHE.

B. Dehazing and Contrast Enhancement

Another main problem for the images obtained underwater is the variation of brightness or luminance. Due to the behaviour of light underwater, some parts appear lighter while some parts appear darker, making it impossible to detect some fine edges in the image. To overcome this issue the image is converted from RGB colour space to YCbCr colour space. The conversion from RGB to YCbCr is given below:

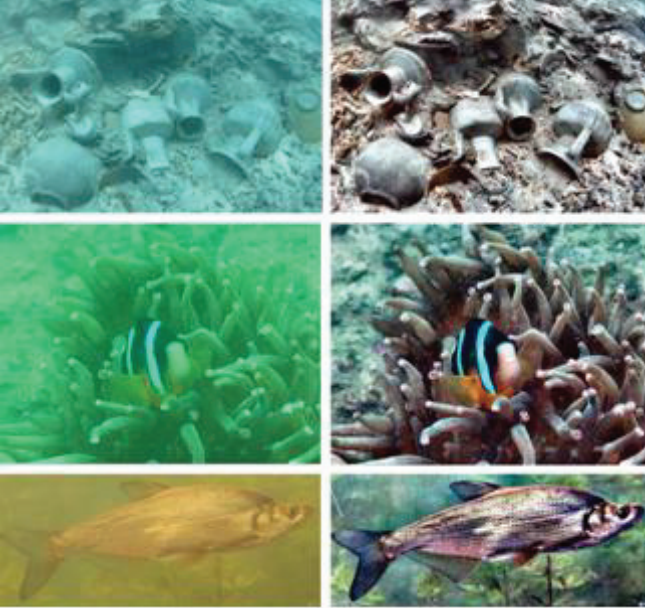


Fig. 8. Top Row: Blue colour Cast; Middle Row: Green colour Cast; Bottom Row: Yellow colour Cast

$$\begin{bmatrix} Y' \\ P_B \\ P_R \end{bmatrix} = \begin{bmatrix} K_R & K_G & K_B \\ -\frac{1}{2} \cdot \frac{K_R}{1-K_B} & -\frac{1}{2} \cdot \frac{K_G}{1-K_B} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \cdot \frac{K_G}{1-K_R} & -\frac{1}{2} \cdot \frac{K_B}{1-K_R} \end{bmatrix} \begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} \quad (4)$$

Where $K_R + K_G + K_B = 1$. Y here represents Luminance while C_b and C_r represent blue difference and red difference chroma component. CLAHE is applied on the Y (Luminance) channel to equalize brightness in the image, again since the variation in brightness is non-homogenous over the image, using a single Histogram Equalization function is not a good idea. Image is now converted back to RGB according to the following matrix:

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 - 2 \cdot K_R \\ 1 & -\frac{K_B}{K_G} \cdot (2 - 2 \cdot K_B) & -\frac{K_R}{K_G} \cdot (2 - 2 \cdot K_R) \\ 1 & 2 - 2 \cdot K_B & 0 \end{bmatrix} \begin{bmatrix} Y' \\ P_B \\ P_R \end{bmatrix} \quad (5)$$

C. Colour Balancing

After applying Histogram Equalization over various channels, there is a need for colour cast removal. White balancing is the process of ensuring white colour is actually white in a picture. Due to the scattering of light underwater the blue and green intensities are usually higher than red intensities resulting in blue/green colour casts. Tint in the images makes it impossible to apply thresholding,

segmentation using colour ranges. To overcome this issue, we will be White Balancing [35] our images.

White balancing in the images is performed by Histogram Stretching [35]. The very first step will be to compute the R G B channel colour histograms. The second step will be to compute two thresholds Higher and Lower proceeded by processing every individual pixel of the R, G, B channel.

Let H be the colour threshold higher than 98% of all the pixels and L be the colour threshold lower than 98% of all the pixels. 2% is left here to maintain robustness.

$$I_{out} = \left\{ \frac{(I_{in} - L)}{(H - L)} \times 255 \right\} + I_{min} \quad (6)$$

Where I_{out} = Output tonal value, I_{in} = Input tonal value and I_{min} = minimum tonal value possible; 0 for range [0,255]. Image obtained after white balancing as shown in Fig. 8 can be observed to be free of unnatural green and blue tint because of scattering of light underwater.

D. Smoothing (Optional)

Underwater images contain a large amount of noise due to effects such as backscattering and marine snow. It depends on the application, if image smoothing is required or not. Algorithms performing segmentation or thresholding tasks perform better in the absence of noise. On the other hand, for human perception or manual analysis, denoising is not necessary. The Primal-dual algorithm was used for denoising [42]. The images will be first converted to CIELAB. Then L and AB channels are denoised. All the results calculated below will be done without this step.

IV. RESULTS AND EVALUATION

A. Dataset and Benchmark

We used the UIEB [7] for benchmarking and evaluation of our method. Dataset consists of 950 underwater images. Out of which 890 images have a corresponding reference image. Remaining 60 images are classified as challenging and no reference image is present. Reference images are generated using 12 models in total out of which 9 image enhancement methods (i.e., fusion-based [24], two-step-based [25], retinex-based [26], UDCP [29], regression-based [38], GDCP [32], Red Channel, histogram prior [45], and blurriness-based [41]), 2 image dehazing methods (i.e., DCP and MSCNN), and 1 commercial application for enhancing underwater images (i.e. dive+). After all the images are generated, 50 volunteers voted for each algorithm, pairwise and the best image was chosen for each raw image. Thus, UIEB provides a method to evaluate our method against the best results in a wide range of underwater image enhancement methods.

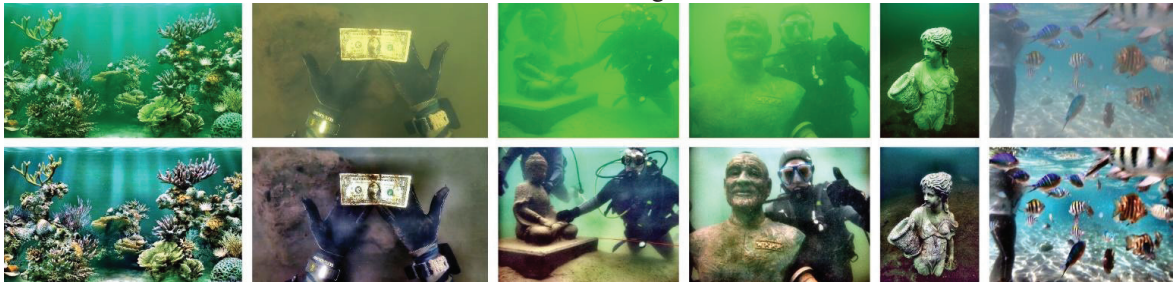


Fig. 9. Results; Top Row: Raw Image; Bottom Row: Final Result; It is evident that the method is capable of colour correction, colour cast removal, dehazing on a variety of images taken in different kinds of water at different depths.

TABLE I. NON-REFERENCE IMAGE QUALITY EVALUATION

Method	Metrics	
	UCIQE ↑	UIQM ↑
Fusion-based[24]	0.6414	1.5310
Two-step-based[25]	0.5776	1.4002
Retinex-based[26]	0.6062	1.4338
UDCP[29]	0.5852	1.6297
Regression Based[38]	0.5971	1.2996
GDCP[30]	0.5993	1.4301
Red Channel[39]	0.5421	1.2147
Histogram Prior[40]	0.6778	1.5440
Blurriness Based[41]	0.6001	1.3757
dive+	0.6227	1.3410
Proposed Method	0.6471	1.7702

*Best in red, Second best in blue

Along with reference images, in [7], non-reference metrics i.e., UCIQE [43] (Underwater Colour Image Quality Evaluation Metric) and UIQM [44] (Universal Image Quality Metric). Reference Metrics, i.e., PSNR (Peak signal-to-noise ratio), SSIM (Structural Similarity Index Metric) [45], MSE (Mean Square Error) is also available for comparison with reference images.

B. Non-reference metrics

In most of the cases, the ground truth images are not available for the test images, we can always encounter unidentified objects in the ocean. Especially, for underwater images, it can be extremely hard or even impossible to get ground truth images. For evaluating the quality of those kinds of images, we can use metrics like dynamic range independent image quality assessment [46], visible edges in an image [47] and image entropy, we can also make use of applications like edge detection, feature point matching, etc. for evaluation. Here we specifically use two metrics (i.e., UIQM[44] and UICQE[43]) which are commonly used for evaluating underwater image quality [38][30][41][40].

1) UICQE

UICQE measures the contrast, saturation and chroma component of an image and then combines them in a linear manner to give results. Images having a better balance between these attributes are likely to get a better UICQE score. As it can be seen in Table I, our algorithm performs second best after histogram prior.

2) UIQM

UIQM focuses on the attributes like contrast, colourfulness and sharpness of an image for evaluation, inspired by the human visual perception. A higher UIQM score implies that the image is more perceivable to the human eye. Again, as seen in Table I, our method outperforms all the other methods.

Even though a good score of UICQE and UIQM should correspond to a more visually perceivable image for humans but it is not always the case. Their results are not always consistent as we are not yet evolved to see properly underwater so it might be possible that after enhancing an

TABLE II. FULL REFERENCE IMAGE QUALITY EVALUATION

METHOD	METRICS		
	MSE ($\times 10^3$) ↓	PSNR (dB) ↑	SSIM ↑
Fusion-based[24]	0.8679	18.7461	0.8162
Two-step-based[25]	1.1146	17.6596	0.7199
Retinex-based[26]	1.3531	16.8757	0.6233
UDCP[29]	5.1300	11.0296	0.4999
Regression Based[38]	1.1365	17.5751	0.6543
GDCP[30]	3.6345	12.5264	0.5503
Red Channel[39]	2.1073	14.8935	0.5973
Histogram Prior[40]	1.6282	16.0137	0.5888
Blurriness Based[41]	1.5826	16.1371	0.6582
dive+	0.5358	20.8408	0.8705
Proposed Method	1.1014	18.5279	0.7865

image, we get good scores of UIQM and UICQE even though the image isn't visually pleasing for humans. It is because our way of perceiving underwater image is not accurate, we focus on the flashing details of images like colours, familiar objects, etc. To overcome this problem, we can evaluate our algorithm against reference images chosen by volunteers in [7]. Below we will be discussing the full reference metrics used.

C. Full Reference Metrics

Full-reference metrics are employed when the ground truth image is available for the test image. Pair of test images and corresponding ground truth images for full-reference evaluation is usually prepared artificially by taking objects underwater or by simulating the underwater environment. We compare the features of the result images with the reference images in the full-reference evaluation method. Although, sometimes the reference images might be different from ground truth images. Here we used three commonly used metrics, PSNR, MSE, and SSIM[45]. A higher SSIM score suggests that the texture and structure of the result image is more similar to that of the reference image. Similarly, a lower MSE score and a higher PSNR score signifies the content similarity of result and test image.

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (7)$$

Where, n = number of data points, Y_i = observed values, \hat{Y}_i = predicted value.

$$PSNR = 20 \cdot \log_{10}(MAX_I) - 10 \cdot \log_{10}(MSE) \quad (8)$$

Where MAX_I = Maximum value, MSE = Mean square error. Our algorithm achieves the **third best** score in SSIM, MSE, PSNR amongst all the algorithms discussed in [7].

D. Runtime Evaluation

We implemented our method in C++17. MATLAB codes of other methods discussed in [7] are converted to C++ code using MATLAB coder and some manually so that we can compare runtimes. All the experiments are conducted on a PC with an Intel(R) i7-7700HQ, CPU, 16GB RAM, on Ubuntu 18.04 LTS. OpenCV was used for the implementation of computer vision algorithms. Average runtimes over all the 950 images available in UIEB dataset is shown in Table III. Images were resized to fit the dimensions to calculate average runtime.

TABLE III. AVERAGE RUNTIME IN SECONDS

METHOD	DIMENSIONS OF IMAGE		
	500 × 500	640 × 480	1280 × 720
Fusion-based[24]	0.065	0.072	0.170
Two-step-based[25]	0.030	0.041	0.120
Retinex-based[26]	0.075	0.080	0.230
UDCP[29]	0.210	0.290	0.840
GDCP[30]	0.312	0.422	0.942
Red Channel[39]	0.243	0.310	0.922
Histogram Prior[40]	0.509	0.563	1.76
Blurriness Based[41]	4.103	4.682	14.294
Proposed Method	0.014	0.018	0.044

*Best in red, second best in blue

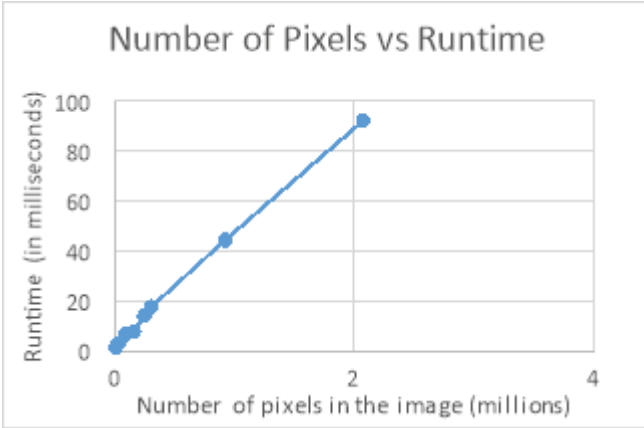


Fig. 10. Change in runtime with respect to changes in total number of pixels

The chrono library in C++ was used to calculate the runtime of the algorithms. As seen in Table III, proposed methodology is the fastest among all the evaluated algorithms. As expected from [7], two step based ranks second. Regression based method was skipped due to complexity of converting it into a C++ implementation.

The entire UIEB (950 images) was processed at an average of 21 milliseconds/image. As shown in Fig. 10, the time complexity of the method is linear, i.e. with respect to changes in the size of input, the runtime will increase linearly. Thus, even a 4K (3840×2160) image will be processed in 370 milliseconds.

V. CONCLUSION, FUTURE WORK AND LIMITATIONS

An underwater image enhancement algorithm is essential for computer vision tasks. In this paper, we proposed an algorithm suitable for real-time and online applications. The proposed method is fast enough to provide 20 FPS (Frames per second) on an HD (High Definition, 720×1280) video. Along with speed, the quality of images generated is comparable to reference images selected manually by volunteers. When it comes to non-reference metrics (UICQE and UIMQM), quality is evaluated as per contrast and colours present, our method performs exceptionally well. The proposed method deals with unwanted colour casts, lost colours, blurriness due to absorption. For the future, performance and effect of the method needs to be evaluated for algorithms such as object detection, automatic thresholding, edge detection etc. Underwater image

degradation is a challenging problem, an algorithm that gives robust results on images with different lighting conditions, different depths, different properties of water and objects is a challenging task. The Proposed algorithm might give unsatisfactory results if the image contains multiple colour casts of variable colours and brightness. Since the white balancing algorithm used is a global algorithm. Adaptive white balancing could be considered for solving the problem, more research is required on the topic.

REFERENCES

- [1] E. Zereik, M. Bibuli, N. Mišković, P. Ridao, and A. Pascoal, "Challenges and future trends in marine robotics," *Annual Reviews in Control*, vol. 46, 2018, doi: 10.1016/j.arcontrol.2018.10.002.
- [2] D. P. Williams, "On optimal AUV track-spacing for underwater mine detection," 2010, doi: 10.1109/ROBOT.2010.5509435.
- [3] J. Henderson, O. Pizarro, M. Johnson-Roberson, and I. Mahon, "Mapping submerged archaeological sites using stereo-vision photogrammetry," *Int. J. Naut. Archaeol.*, 2013, doi: 10.1111/1095-9270.12016.
- [4] F. M. Caimi, D. M. Kocak, F. Dalglish, and J. Watson, "Underwater imaging and optics: Recent advances," 2008, doi: 10.1109/OCEANS.2008.5152118.
- [5] P. Drap, "Underwater Photogrammetry for Archaeology," in *Special Applications of Photogrammetry*, 2012.
- [6] Y. Cho and A. Kim, "Visibility enhancement for underwater visual SLAM based on underwater light scattering model," 2017, doi: 10.1109/ICRA.2017.7989087.
- [7] C. Li *et al.*, "An Underwater Image Enhancement Benchmark Dataset and beyond," *IEEE Trans. Image Process.*, vol. 29, 2020, doi: 10.1109/TIP.2019.2955241.
- [8] D. Akkaynak and T. Treibitz, "A Revised Underwater Image Formation Model," 2018, doi: 10.1109/CVPR.2018.00703.
- [9] D. Akkaynak, T. Treibitz, T. Shlesinger, R. Tamir, Y. Loya, and D. Iluz, "What is the space of attenuation coefficients in underwater computer vision?," 2017, doi: 10.1109/CVPR.2017.68.
- [10] J. Y. Chiang and Y. C. Chen, "Underwater image enhancement by wavelength compensation and dehazing," *IEEE Trans. Image Process.*, 2012, doi: 10.1109/TIP.2011.2179666.
- [11] A. L. Alldredge and M. W. Silver, "Characteristics, dynamics and significance of marine snow," *Progress in Oceanography*, 1988, doi: 10.1016/0079-6611(88)90053-5.
- [12] S. Corchs and R. Schettini, "Underwater image processing: State of the art of restoration and image enhancement methods," *EURASIP J. Adv. Signal Process.*, 2010, doi: 10.1155/2010/746052.
- [13] M. Han, Z. Lyu, T. Qiu, and M. Xu, "A Review on Intelligence Dehazing and Color Restoration for Underwater Images," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020, doi: 10.1109/TSMC.2017.2788902.
- [14] R. Cui, L. Chen, C. Yang, and M. Chen, "Extended State Observer-Based Integral Sliding Mode Control for an Underwater Robot With Unknown Disturbances and Uncertain Nonlinearities," *IEEE Trans. Ind. Electron.*, 2017, doi: 10.1109/TIE.2017.2694410.
- [15] M. Calisti, G. Carbonara, and C. Laschi, "A rotating polarizing filter approach for image enhancement," 2017, doi: 10.1109/OCEANSE.2017.8084722.
- [16] G. R. Fournier, "Range-gated underwater laser imaging system," *Opt. Eng.*, 1993, doi: 10.1117/12.143954.
- [17] C. S. Tan, A. Sluzek, G. G. L. Seet, and T. Y. Jiang, "Range gated imaging system for underwater robotic vehicle," 2006, doi: 10.1109/OCEANSAP.2006.4393938.
- [18] P. Mariani *et al.*, "Range-gated imaging system for underwater monitoring in ocean environment," *Sustain.*, 2018, doi: 10.3390/su11010162.
- [19] T. Treibitz *et al.*, "Wide Field-of-View Fluorescence Imaging of Coral Reefs," *Sci. Rep.*, 2015, doi: 10.1038/srep07694.
- [20] K. Iqbal, M. Odetayo, A. James, R. A. Salam, and A. Z. H. Talib, "Enhancing the low quality images using unsupervised colour correction method," 2010, doi: 10.1109/ICSMC.2010.5642311.

- [21] A. S. Abdul Ghani and N. A. Mat Isa, "Underwater image quality enhancement through integrated color model with Rayleigh distribution," *Appl. Soft Comput. J.*, 2015, doi: 10.1016/j.asoc.2014.11.020.
- [22] A. S. Abdul Ghani and N. A. Mat Isa, "Enhancement of low quality underwater image through integrated global and local contrast correction," *Appl. Soft Comput. J.*, 2015, doi: 10.1016/j.asoc.2015.08.033.
- [23] C. O. Ancuti, C. Ancuti, C. De Vleeschouwer, and P. Bekaert, "Color Balance and Fusion for Underwater Image Enhancement," *IEEE Trans. Image Process.*, 2018, doi: 10.1109/TIP.2017.2759252.
- [24] C. Ancuti, C. O. Ancuti, T. Haber, and P. Bekaert, "Enhancing underwater images and videos by fusion," 2012, doi: 10.1109/CVPR.2012.6247661.
- [25] X. Fu, Z. Fan, M. Ling, Y. Huang, and X. Ding, "Two-step approach for single underwater image enhancement," 2017, doi: 10.1109/ISPACS.2017.8266583.
- [26] X. Fu, P. Zhuang, Y. Huang, Y. Liao, X. P. Zhang, and X. Ding, "A retinex-based enhancing approach for single underwater image," 2014, doi: 10.1109/ICIP.2014.7025927.
- [27] S. Zhang, T. Wang, J. Dong, and H. Yu, "Underwater image enhancement via extended multi-scale Retinex," *Neurocomputing*, 2017, doi: 10.1016/j.neucom.2017.03.029.
- [28] H. Y. Yang, P. Y. Chen, C. C. Huang, Y. Z. Zhuang, and Y. H. Shiau, "Low complexity underwater image enhancement based on dark channel prior," 2011, doi: 10.1109/IBICA.2011.9.
- [29] P. L. J. Drews, E. R. Nascimento, S. S. C. Botelho, and M. F. M. Campos, "Underwater depth estimation and image restoration based on single images," *IEEE Comput. Graph. Appl.*, 2016, doi: 10.1109/MCG.2016.26.
- [30] Y. T. Peng, K. Cao, and P. C. Cosman, "Generalization of the Dark Channel Prior for Single Image Restoration," *IEEE Trans. Image Process.*, 2018, doi: 10.1109/TIP.2018.2813092.
- [31] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson, "WaterGAN: Unsupervised generative network to enable real-time color correction of monocular underwater images," *IEEE Robot. Autom. Lett.*, 2018, doi: 10.1109/LRA.2017.2730363.
- [32] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern Recognit.*, 2020, doi: 10.1016/j.patcog.2019.107038.
- [33] P. Liu, G. Wang, H. Qi, C. Zhang, H. Zheng, and Z. Yu, "Underwater Image Enhancement with a Deep Residual Framework," *IEEE Access*, 2019, doi: 10.1109/ACCESS.2019.2928976.
- [34] J. Banerjee, R. Ray, S. R. K. Vadali, S. N. Shome, and S. Nandy, "Real-time underwater image enhancement: An improved approach for imaging with AUV-150," *Sadhana - Acad. Proc. Eng. Sci.*, vol. 41, no. 2, pp. 225–238, 2016, doi: 10.1007/s12046-015-0446-7.
- [35] S. Wang, Y. Zhang, P. Deng, and F. Zhou, "Fast automatic white balancing method by color histogram stretching," 2011, doi: 10.1109/CISP.2011.6100338.
- [36] S. M. Pizer *et al.*, "ADAPTIVE HISTOGRAM EQUALIZATION AND ITS VARIATIONS," *Comput. vision, Graph. image Process.*, 1987, doi: 10.1016/S0734-189X(87)80186-X.
- [37] A. Buades, B. Coll, and J.-M. Morel, "Non-Local Means Denoising," *Image Process. Line*, 2011, doi: 10.5201/ipol.2011.bcm_nlm.
- [38] C. Li, J. Guo, C. Guo, R. Cong, and J. Gong, "A hybrid method for underwater image correction," *Pattern Recognit. Lett.*, 2017, doi: 10.1016/j.patrec.2017.05.023.
- [39] A. Galdran, D. Pardo, A. Picón, and A. Alvarez-Gila, "Automatic Red-Channel underwater image restoration," *J. Vis. Commun. Image Represent.*, 2015, doi: 10.1016/j.jvcir.2014.11.006.
- [40] C. Y. Li, J. C. Guo, R. M. Cong, Y. W. Pang, and B. Wang, "Underwater image enhancement by Dehazing with minimum information loss and histogram distribution prior," *IEEE Trans. Image Process.*, 2016, doi: 10.1109/TIP.2016.2612882.
- [41] Y. T. Peng and P. C. Cosman, "Underwater Image Restoration Based on Image Blurriness and Light Absorption," *IEEE Trans. Image Process.*, 2017, doi: 10.1109/TIP.2017.2663846.
- [42] M. Zhu, "Fast numerical algorithms for total variation based image restoration," *Thesis*, 2008.
- [43] M. Yang and A. Sowmya, "An Underwater Color Image Quality Evaluation Metric," *IEEE Trans. Image Process.*, 2015, doi: 10.1109/TIP.2015.2491020.
- [44] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, 2002, doi: 10.1109/97.995823.
- [45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, 2004, doi: 10.1109/TIP.2003.819861.
- [46] T. O. Aydm, R. Mantiuk, K. Myszkowski, and H. P. Seidel, "Dynamic range independent image quality assessment," 2008, doi: 10.1145/1399504.1360668.
- [47] N. Hautière, J. P. Tarel, D. Aubert, and É. Dumont, "Blind contrast enhancement assessment by gradient ratioing at visible edges," *Image Anal. Stereol.*, 2008, doi: 10.5566/ias.v27.p87-95.

Generation of COVID-19 Chest CT Scan Images using Generative Adversarial Networks

Prerak Mann¹
Computer Engineering
Delhi Technological
University
New Delhi, India
prerakmann_2k17co241@dtu
.ac.in

Sahaj Jain¹
Computer Engineering
Delhi Technological
University
New Delhi, India
sahajjain_2k17co291@dtu.ac
.in

Saurabh Mittal¹
Computer Engineering
Delhi Technological
University
New Delhi, India
saurabhmittal_2k17co309@d
tu.ac.in

Aruna Bhat²
Computer Science and Engineering
Delhi Technological University
New Delhi, India
aruna.bhat@dtu.ac.in

Abstract—SARS-CoV-2, also known as COVID-19 or Coronavirus, is a viral contagious disease that is infected by a novel coronavirus, and has been rapidly spreading across the globe. It is very important to test and isolate people to reduce spread, and from here comes the need to do this quickly and efficiently. According to some studies, Chest-CT outperforms RT-PCR lab testing, which is the current standard, when diagnosing COVID-19 patients. Due to this, computer vision researchers have developed various deep learning systems that can predict COVID-19 using a Chest-CT scan correctly to a certain degree. The accuracy of these systems is limited since deep learning neural networks such as CNNs (Convolutional Neural Networks) need a significantly large quantity of data for training in order to produce good quality results. Since the disease is relatively recent and more focus has been on CXR (Chest XRay) images, the available chest CT Scan image dataset is much less. We propose a method, by utilizing GANs, to generate synthetic chest CT images of both positive and negative COVID-19 patients. Using a pre-built predictive model, we concluded that around 40% of the generated images are correctly predicted as COVID-19 positive. The dataset thus generated can be used to train a CNN-based classifier which can help determine COVID-19 in a patient with greater accuracy.

I. INTRODUCTION

COVID-19 is on the spread, and without any known vaccine or treatment, a significant number (~10%) of people having fatal reactions, and a mortality rate of ~2-3%, there is an urgent need to test and isolate people. According to Chinese authorities' publications, the diagnosis of COVID-19 has to be verified by gene sequencing of respiratory or blood specimens or RT-PCR (reverse-transcription polymerase chain reaction). However, due to the limitations of transportation and sample collection, as well as the testing kit's performance, throat swab samples have RT-PCR's total positive rate to be approximately around 30% - 60% only. Another factor is that one has to wait for the lab test results, which usually takes around 24-36 hours.

In a study of more than 1000 patients[2], chest-CT outperformed RT-PCR lab testing when diagnosing COVID-19 patients. From the results, the chest CT scans of 88% of the patients were positive, while only 59% had positive RT-PCR results. This goes to show that chest CT scans are more accurate for the screening of the novel coronavirus disease.

The number of chest CT Scan images available for COVID-19 patients is very less due to which the accuracy of a Convolutional Neural Network (CovNet/CNN) classifier is limited. Deep learning neural networks such as CNNs (convolutional neural networks) need a significantly large number of data for training in order to produce good quality results. Since the disease is relatively recent and more focus has been on CXR (Chest XRay) images, the available chest CT Scan image dataset is much less.

In order to tackle this situation, we came up with a solution of increasing the available Chest-CT scan dataset using synthetic images generated by a GAN model. This extended dataset can now be used to develop an improved CNN-based classifier model.

II. BACKGROUND AND RELATED WORK

Image generation is the task of generating new synthetic images from a given dataset. There are many machine learning techniques such as Variational Autoencoders (VAEs), Autoregressive model, Flow model, Hybrid Models (a combination of these techniques), Generative Adversarial Networks (GANs). The latest of these techniques is GANs [4]. They belong to a set of generative models and can be used to generate text, audio, and images. GANs have had their application increase manifold in the past few years in fields such as science, fashion, art, advertising, etc., and have an advantage over other techniques when the task is to generate images that are very realistic and virtually indistinguishable from real images.

GAN consists of two primary networks - Generator ($G(z)$) and Discriminator ($D(x)$). The generator module in GANs is used to create artificial samples of data by incorporating feedback from the discriminator. Its objective is to deceive the discriminator into classifying the generated data as belonging to the original dataset and ultimately minimize $V(D, G)$ which is the cost value function.

During training, the generator encapsulates the probability distribution of original data and is trained to generate data that maximizes the probability of the discriminator mistaking the fake data to be real. The end goal of the generator is such that the discriminator is no longer able to differentiate between real data or synthetically generated data. The generator module is a neural network that consists of one input layer, one or more hidden layers, and an output layer. The discriminator module in GAN can be thought of as a classifier. The aim of the discriminator is to categorize data under analysis to real or fake. The architecture of the discriminator depends on the type of data it is classifying.

The combined loss of the GAN can be represented by the following equation -

$$\min_G \max_D V(D, G) = \min_G \max_D (E_{x \sim P_{data}(x)} [\log D(x)] + E_{z \sim P_z(z)} [\log(1 - D(G(z)))])$$

COVID-CT-Dataset[1] identified the difficulty of publically available COVID-19 CT datasets due to various privacy issues and came up with its own publicly available dataset. The dataset is composed of 349 chest CT scans of COVID-19 positive patients collected from 216 patients and it contains 463 non-COVID-19 CT images. It has also verified this dataset from a senior radiologist. It also came up with useful experimentation results where it demonstrated the usefulness of the dataset by building AI models

for diagnosing COVID-19. The diagnosis model achieves impressive results with an accuracy of 0.89 by exploiting multi-task learning and self-supervised learning.

III. METHODOLOGY

During the analysis of the dataset, we realized that the images available in the dataset are of different dimensions and have different brightness. So, the first step in our experiment was to preprocess the images and transform them to a fixed size and normalize the brightness. We decided to resize the images to 224 x 224 for our models. Using the PyTorch Vision library, we performed the following on the images—downscaling, random resized crop, random horizontal flip, and normalization.

We chose Deep Convolutional GANs (DCGANs) for generating images as our generative model. We designed the Discriminator and the Generator model such that their architectures are symmetric since it is the simplest and the most effective way of ensuring that both models are equally powerful and have fair competition. The complete architecture is trained using the PyTorch framework.

The generator takes input as a noise vector of shape 100 x 1 and outputs a single 224 x 224 x 3 image. The first layer of the network is a fully connected layer with 100 input features and 150528 output features. The output of this layer is provided as input to five transpose convolutional layers (ConvTranspose2D) to upsample the input vector, first to 3072 x 7 x 7, then 1536 x 14 x 14, then 768 x 28 x 28, then 384 x 56 x 56, then 192 x 112 x 112, and finally to 3 x 224 x 224. Except for the last layer, each convolution transpose layer is followed by a batch normalization layer (BatchNorm2D) and a Rectified unit (ReLU) activation layer. The transpose convolutional layers are configured to use a kernel size of (4, 4) and stride of (2, 2). The activation layer uses the ReLU function whereas the output layer uses the hyperbolic tangent (Tanh) function. The generator has ~30 million parameters. The output of the generator is an image of shape 224 x 224 x 3. The layer-based architecture of the Generator and Discriminator model is shown in Fig. 1 and Fig 2.

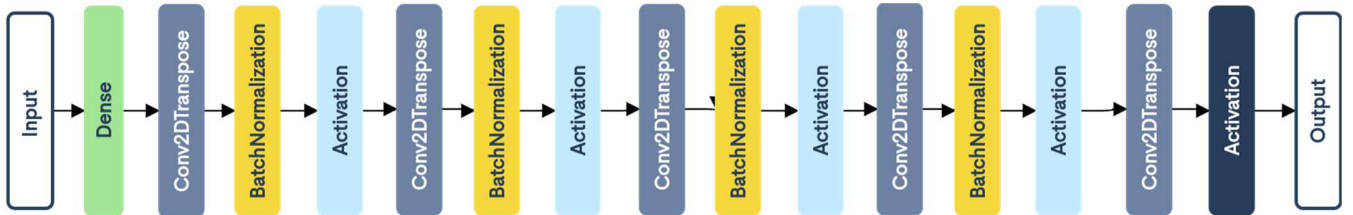


Fig 1: Layered architecture of the Generator model

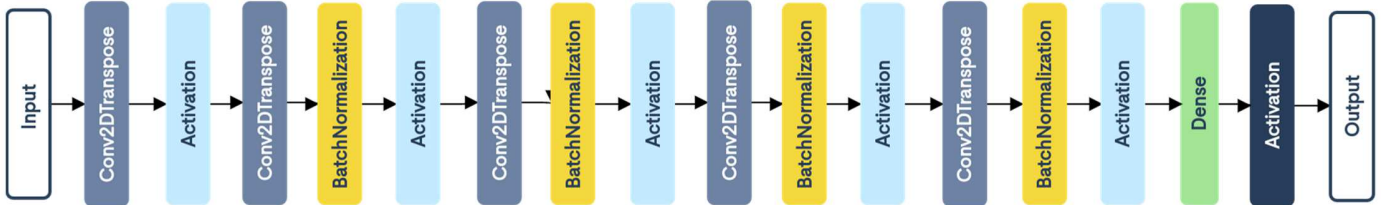


Fig 2: Layered architecture of the Discriminator model

A CNN based model is used as the Discriminator model which outputs whether the image is fake/real (class = 0/1). The input image is passed through five convolutional layers (Conv2D) each of which is followed by a batch normalization layer (BatchNorm2D) and a LeakyReLU activation layer. The input is downsampled from $224 \times 224 \times 3$ to $192 \times 112 \times 112$, then $384 \times 56 \times 56$, then $768 \times 28 \times 28$, then $1536 \times 14 \times 14$, then $3072 \times 7 \times 7$. Kernel size chosen for the model is (4, 4), the size of the stride chosen is (2, 2) and the activation function used is LeakyReLU with a slope of 0.2. There are approximately 5 million parameters in the Discriminator. The final output is reduced to a vector and passed to a dense layer. The output of the dense layer is used to estimate if the image is real or not. The output layer predicts the realness using the Sigmoid function. The layered architecture of the Generator and Discriminator models are shown in Fig. 3 & 4 respectively.

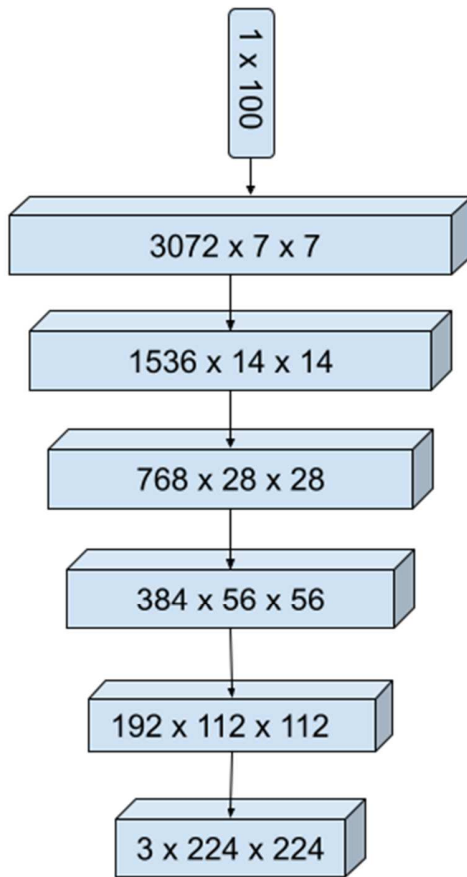


Fig 3: Generator architecture

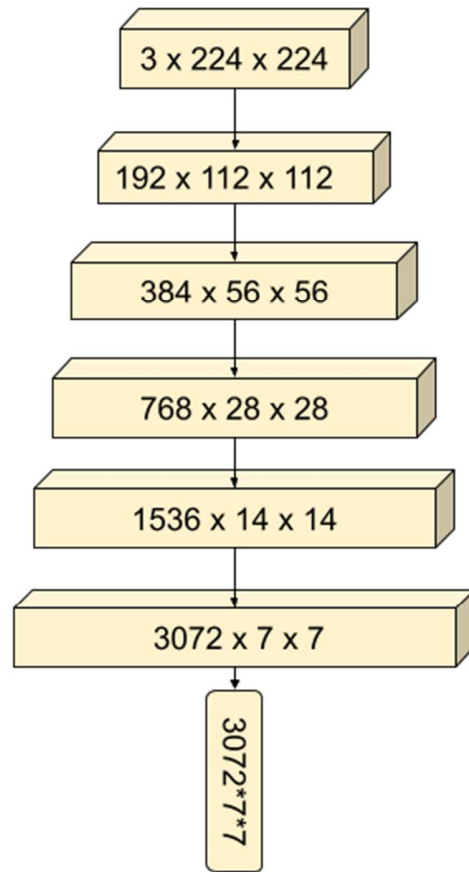


Fig 4: Discriminator architecture

The loss function used for the GAN is the “binary cross-entropy loss”. The loss can be represented by the simple equation -

$$L = \min_G \max_D [\log(D(x)) + \log(1 - D(G(z)))]$$

The major issue of GANs is the validation of the generated image i.e. the generated image is accurate or not. This task might be trivial in cases like Face Generation or Image to Image translation but not in this case. Generated images need to be validated on whether they are actually of COVID-19 positive Chest CT Scans. The ideal way would be to get a radiologist to review the generated images and handpick the images that can be added to our dataset. But since this is not possible, we decided to use a Baseline model[3] that can perform the same task. We used a CNN based model available with the dataset to measure the performance of the GAN model. The baseline model is provided by the COVID-CT repository[1]. The baseline model is an optimized version of the DenseNet-169 pre-trained model and it classifies the input images in two classes - COVID-19 positive and Normal.

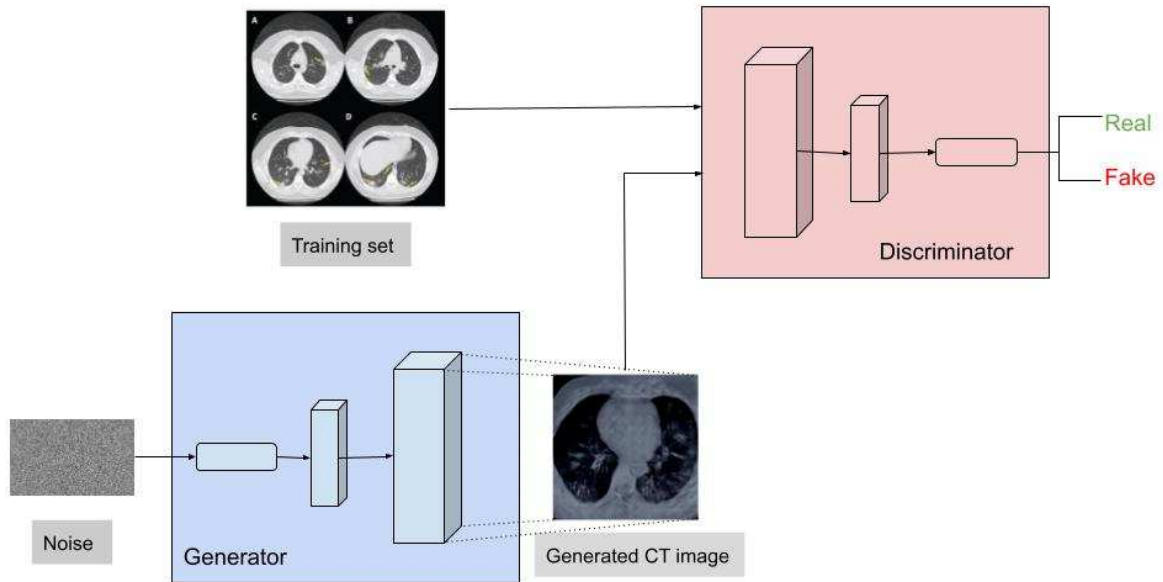


Fig 5: Model Training

IV. EXPERIMENTS AND RESULTS

The model is trained to generate COVID-19 positive Chest CT images. The image preprocessing involves resizing the image to 224 x 224 x 3 and then normalizing the image pixels to $[-1, 1]$ from $[0, 255]$. Adam is used as the optimizer function. The hyperparameters used for the training are as follows: “batch size: 16, learning rate: 0.0002, beta: 0.5 (Adam optimizer momentum) and the number of epochs: 3000.” The model has ~35 million parameters.

The images generated by our DCGAN were compared with the Baseline model provided by the COVID-CT repository[1]. The baseline model is an optimized version of the DenseNet-169 pre-trained model. We generated ten sets of 100 COVID-19 positive chest CT images using the Generator model and predicted their nature using the baseline model. Since all images should be COVID-19 positive the accuracy is calculated by counting the number of images marked COVID-19 positive by the baseline model. On

average, the model predicted about 40% of generated images to be COVID-19 positive. The generated images and calculated accuracies for the sets can be seen in Fig. 6 & 7 respectively.

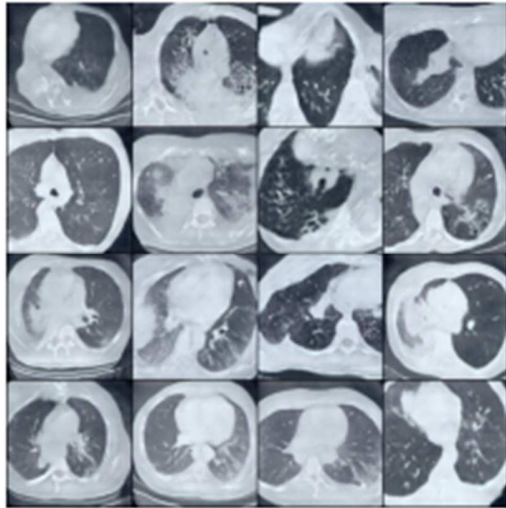


Fig 6: Generated Chest CT images

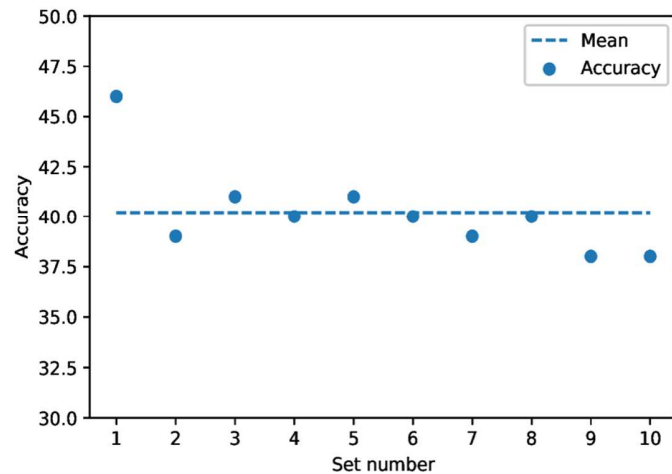


Fig 7: Scatter graph of accuracy vs Set.

V. CONCLUSION AND FUTURE WORK

In this paper, we have shown how GANs can be useful to generate synthetic images. Using the baseline models we can see that around 40% of images are being correctly predicted as being COVID-19 positive. The motive behind generating synthetic images was to extend the existing dataset of chest CT images which can be utilized to build a CNN-based predictive model. The models that are currently available for detection of SARS-CoV-2 using chest CT scans [5][6][7] were trained on small datasets of chest CTs and have accuracy in the range of 85-90%. CNN-based networks are prone to overfitting if trained using a small dataset. Therefore, if these available networks are trained on the extended dataset, they could perform better and give more accurate results. Also, the extended dataset could be published separately for the development of further predictive models.

REFERENCES

- [1] Xingyi Yang, Xuehai He, Jinyu Zhao, Yichen Zhang, Shanghang Zhang, Pengtao Xie, "COVID-CT-Dataset: A CT Scan Dataset about COVID-19", arXiv:2003.13865, 2020, [online] Available: <https://arxiv.org/abs/2003.13865>
- [2] Tao Ai, Zhenlu Yang, Hongyan Hou, Chenao Zhan, Chong Chen, Wenzhi Lv, Qian Tao, Ziyong Sun, Liming Xia, "Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases", [online] Available: <https://doi.org/10.1148/radiol.2020200642>
- [3] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Tudjman and P. R. Pinheiro, "CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection," in IEEE Access, vol. 8, pp. 91916-91923, 2020, DOI: 10.1109/ACCESS.2020.2994762.

- [4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., "Generative adversarial nets", Proc. Adv. Neural Inf. Process. Syst., pp. 2672-2680, 2014.
- [5] Pham, T.D. A comprehensive study on classification of COVID-19 on computed tomography with pre-trained convolutional neural networks. Sci Rep 10, 16942 (2020). <https://doi.org/10.1038/s41598-020-74164-z>
- [6] Yazdani, Shakib & Minaee, Shervin & Kafieh, Rahele & Saeedizadeh, Narges & Sonka, Milan. (2020). COVID CT-Net: Predicting Covid-19 From Chest CT Images Using Attentional Convolutional Network.
- [7] Singh, D., Kumar, V., Vaishali et al. Classification of COVID-19 patients from chest CT images using multi-objective differential evolution-based convolutional neural networks. Eur J Clin Microbiol Infect Dis 39, 1379–1389 (2020). <https://doi.org/10.1007/s10096-020-03901-z>

Handwriting Recognition for Medical Prescriptions using a CNN-Bi-LSTM Model

Tavish Jain

Department of Computer Science &
Engineering
Delhi Technological University
New Delhi, India
jaintavish@gmail.com

Rohan Sharma

Department of Computer Science &
Engineering
Delhi Technological University
New Delhi, India
rohan_dce@outlook.com

Ruchika Malhotra

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

Abstract - It is commonly seen that it is tough to read the handwritten text from medical prescriptions. It is mostly due to the different style of handwriting and the use of Latin abbreviations for medical terms which is usually unknown to the general public. This can make it difficult for both patients and even pharmacists to read the prescription, which can have negative or even fatal consequences if read incorrectly. This paper demonstrates the use of a CNN-Bi-LSTM model along with Connectionist Temporal Classification. The prescribed model consists of three components, the convolutional layers for feature extraction, the Bi-LSTM network for making predictions for each frame of the context vector and the final decoding to translate each character in the recognized sequence by LSTM layers into an alphabetic character using the CTC loss function. A linear layer is added after the bi-LSTM layer to compute the final probabilities, which will be decoded. We also built a corpus manually containing the terms widely used in the medical domain, commonly used in prescriptions. We then use string matching algorithms, and string distance functions to find the nearest word in the corpus, so that bias is given to medical terms for increasing accuracy of the predicted output.

Keywords - Long-short term memory networks, convolutional networks, neural networks, connectionist temporal classification, recurrent neural networks, character error rate, batch normalization, Seq2Seq networks, Adam Optimizer, PyTorch

I. INTRODUCTION

It is becoming increasingly common that people incorrectly read the medical prescriptions, and hence go on to consume wrong medicines or wrong dosage of medicines which is very harmful to their health, and may prove to be fatal in some cases. This mostly happens due to the fact that most doctors have illegible handwriting, and also due to the lack of medical knowledge of patients and the chemists. This is becoming an increasingly common problem, but can be solved using technology.

Deep Learning has been a major force in driving research advancements around text recognition [3]. Deep learning models have been a success due to the recent architectures and availability of large scale annotated data. There have been many attempts to leverage the power of deep learning to solve this issue in the past. But with the recent advancements in the field of Deep Learning, text recognition has become highly accurate and reliable which is a great solution to this problem. There are various different existing methods for simple text recognition, but the need of the hour is a custom technique specially suited for reading medical prescriptions. Such a technique could help remove errors in

reading the medicine or treatment names and dosage, and thus help save people's lives.

Through this paper we intend to develop a technique that is specially trained to recognize medical prescriptions correctly. The technique will take images of medical prescriptions as input, and return the text written in the prescription so that less mistakes are made in reading the prescription. The general steps that were used in the paper for handwriting recognition were sequenced as:

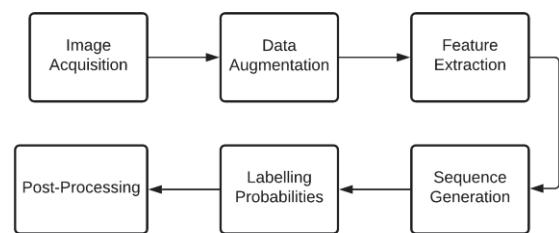


Fig. 1. Flow of logic

We have also included the methods used for building the dataset, writing the python code, the training process for the model and the results that were finally achieved.

II. LITERATURE REVIEW

A lot of researchers have researched into the topic of text recognition, and various approaches follow, which include Keyword Spotting [1], text orientation detection, information energy based on each pixel, stroke identification, MSER techniques, and also Deep Learning techniques that involve convolutional neural Networks [6].

Abhishek Bal et al. [8] presented a handwritten document analysis which uses segmentation and detecting the amount of pressure applied for the documents. The method is based on horizontal and vertical projections that divide the line and words. The technique also performs well in the presence of skewed and overlapped text. The method was tested on the IAM database.

Kanchan Keisham et al. [9] proposed a line segmentation approach on information energy that is calculated for all pixels, and the classification is done with the help of Artificial Neural Networks.

Nibaran Das et al. [4] demonstrated the use of the convex hull algorithm. A total of one hundred twenty-five features are extracted by the use of various attributes of the hull.

These experiments were carried out on the Bangla basic characters' dataset.

Nafiz Arica et al. [5] put forward recognition algorithms, which aimed to recognize cursive handwriting. The segmentation procedure included converting the image to grayscale, and then applying Hidden Markov Models for prediction of the characters.

Subhadip Basu et al. [7] presented the use of multi-layer perceptrons for recognition. Feature sets were designed for character recognition and used three kinds of topological features. These experiments were carried out on the Bangla basic characters' dataset.

Namrata Dave et al. [10] proposed techniques that could help segment the text. Three different levels of segmentation were proposed to be used. First a text level segmentation is done, followed by a word-level segmentation, and finally a character level segmentation has been explained.

III. PROPOSED ARCHITECTURE

A. Data Preparation

We used the publicly available IAM dataset for this paper. We registered on the website and downloaded images of lines, and its annotations, which were available in XML format, as well as the TXT format. In the images, there were "bounding boxes" around the words which theoretically gave additional context for a neural network to learn. We augmented the input images by distorting it. We pass in the complete images to the neural network, and its annotations in an encoded format, by creating a dictionary of all the characters that were used in the recognized text. We pass in the image of a line of textual data, along with the image, which is later on decoded when the model returns its output.

B. Model Architecture

We try to convert the input image into the text using a deep convolutional neural network, which converts the input image into a context vector, which is then sent as the input to the Bi-LSTM Decoder network, which outputs the predicted and converted sequence from the image. The network uses a complex architecture, using seven convolutional layers, along with optional batch Normalization layers, Max Pooling layers, ReLU and LeakyReLU activation functions for the Encoder, and a Bi-directional LSTM layer and a Linear

Layer as the Decoder, which finally returns us the predicted probabilities.

The built network takes in a variable width image as an input, where the length of the image is sixty pixels. The variable dimension of the images is normalized after the first convolutional layer. The first convolution operation changes the number of channels in the image from three to sixty-four which is increased up to five hundred twelve in further convolutional layers. Every layer is followed by ReLU, MaxPool layers and optional Batch normalization layers. This marks the end of the Encoder, which takes in the input as a processed image, and returns the context vector. Context Vectors can be said to be fixed-length vector representations which store model weights from the Encoder layers, and are fed as an input to the Decoder layer. The Decoder consists of two Bi-directional LSTM layers, along with a dropout value of 0.5. The Decoder is finally ended by adding a Linear Layer at the end, which has its count of input nodes as two thousand forty-eight, and the count of output nodes as the number of characters in the dictionary. The final Linear layer acts as an embedding layer. The final Linear layer gives us the output probabilities, which are of the shape [BATCH_SIZE, DICTONARY_SIZE, SEQUENCE_LENGTH], which is then passed to decode to characters.

To decode the output probabilities, we use the argmax function to find the index of the maximum probabilistic index, and the index is returned as a vector, which is then converted to the corresponding English character. Sequences of the returned English characters are accumulated together to eventually form the predicted text from the medical prescription image. from the mapping initially created. This however contains a lot of extra characters, which is then normalized by the use of Connectionist Temporal Classification. The neural network model can be trained end to end using widely available IAM Handwriting dataset. We also use data augmentation on input images by distorting the image, adding meshes to the image, applying linear and cubic interpolation methods and finally warping the image. Since the Encoder is fully convolutional, it is not restricted to fixed-size input.

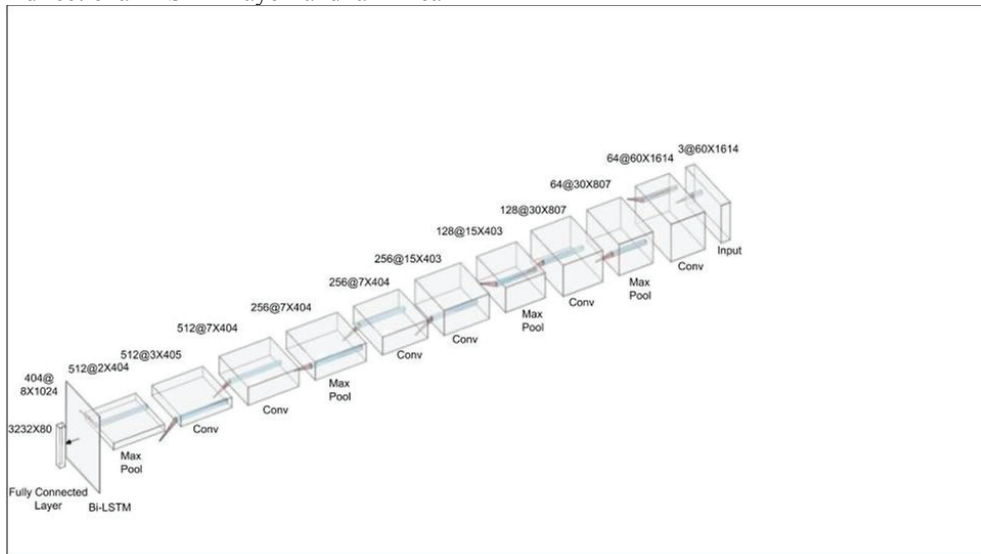


Fig. 2. Model Architecture

C. Predictions with Bi-LSTM

Long Short Term Memory Networks are used to recognize/classify the next character [2] from the input text.

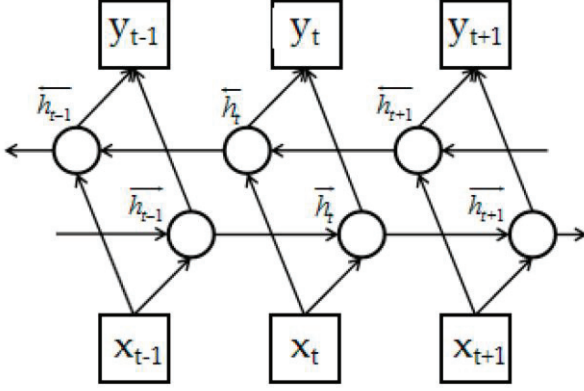


Fig. 3. Bi-LSTM Network Structure

The Encoder Layer returns us a tensor containing probabilities of occurrence of each character for a fixed sequence length. The argmax function is then used to find the character having the maximum probability, which is then finally used in the output character sequence. This output character sequence is then decoded into alphabets by reverse mapping the indices selected by the argmax function, to the corresponding alphabets. The Connectionist Temporal Classification technique is then used to remove all the

D. Corpus for Medical Terms

We also manually built a corpus containing the medical terms, which are used in prescriptions. We use string matching algorithms and string distance functions to find the nearest word in the corpus, so that bias is given to medical terms for increasing accuracy of the predicted output.

IV. EXPERIMENTAL RESULTS

Training the deep learning model using more and more images will help in increasing the accuracy/ reducing the loss of the prescribed deep learning model. We used the *Large*

Writer Independent Text Line Recognition Task which defines an experiment with well-defined training, test, and validation sets. This returns us with nine thousand plus training set images, which covers text-lines from over three hundred seventy writers, and thousand plus images for testing the built model, which covers text-lines from over one hundred twenty writers, all being mutually exclusive to each other.

We trained the model for thirty-two epochs. Training the model with more data and for longer epochs will help in increasing the accuracy of the model. We save the model configuration as and when we reach a minimum CER Loss value, and save the weights of the model. If the training loss comes out to be more than validation loss, it indicates that the model is under fitting. However, if the training loss is less than the validation loss, it indicates the model is overfitting. However, the pursued result is to have the training loss equivalent to validation loss. Training the model up to thirty-two epochs took six hours to finish.

The loss values as the model trained have been illustrated in the figure below. The orange curve denotes the test loss Vs Epochs. The blue curve denotes the training loss Vs Epochs of the model. Increasing the dataset collection will increase the accuracy rate.

Character Error Rate (CER) calculates the count of characters in the handwriting that the Deep Learning model did not read correctly. We prefer using CER as a metric, rather than WER (Word Error Rate), as it is not meaningful, as it would highly decrease the quality of the model. Predicting Words matching exactly without an error will be difficult as words could be of variable length with repeating characters, which might or might not be covered perfectly.

Logs are generated as epochs are completed showing training CER and test CER. The best model is saved/updated after every epoch. Some logs are attached below.

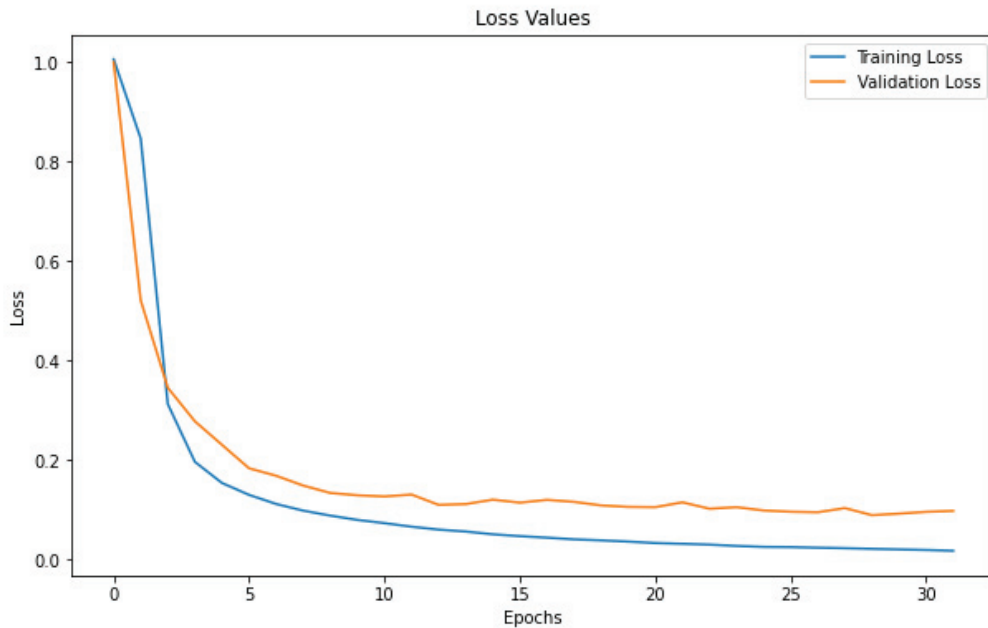


Fig. 4. Loss(CER) Vs Epochs


```
Starting for epoch: 18
Training CER 0.040290962858451795
Test CER 0.11566633122235549
```

```
Starting for epoch: 19
Training CER 0.038230602058216295
Test CER 0.1085081087030012
Saving Best
```

```
Starting for epoch: 28
Training CER 0.022497624425115867
Test CER 0.10314193567544347
```

```
Starting for epoch: 29
Training CER 0.021179178003325614
Test CER 0.08891598131148906
Saving Best
```

V. CONCLUSION

The paper deals with studying different techniques for handwritten text recognition. We've used data augmentation techniques to make the model more robust to noise, and also avoid overfitting. Multiple layers of Convolutional Neural Networks perform the feature extraction, and bi-LSTM's help in decoding the extracted features to English characters. Since the actual alignment between the input and the output is not known, we use Connectionist Temporal Classification to get around not knowing that alignment. More bias is given to words that are present in a manually created corpus to accurately recognize text specific to prescriptions offered by the doctors.

VI. ACKNOWLEDGMENT

The authors are very much grateful to the Department of Computer Science & Engineering of Delhi Technological University for giving us the opportunity to work on Handwritten Recognition for Medical Prescriptions. Both the authors sincerely express their gratitude to Dr. Ruchika

Malhotra for giving constant encouragement in doing research in the field of image processing

REFERENCES

- [1] Partha Pratim Roy, Ayan Kumar Bhunia, Ayan Das, Prithviraj Dhar, Umapada Pal, "Keyword spotting in doctor's handwriting on medical prescriptions", *Expert Systems with Applications*, Volume 76, 2017, Pages 113-128, ISSN 0957-4174
- [2] P. S. Dhande and R. Kharat, "Character Recognition for Cursive English Handwriting to Recognize Medicine Name from Doctor's Prescription," 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, 2017, pp. 1-5, doi:10.1109/ICCUBEA.2017.8463842.
- [3] N. Chumuang and M. Ketcham, "Model for Handwritten Recognition Based on Artificial Intelligence," 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Pattaya, Thailand, 2018, pp. 1-5, doi: 10.1109/iSAI-NLP.2018.8692958.
- [4] Nibaran Das, Sandip Pramanik, Subhadip Basu, Punam Kumar Saha, "Recognition of handwritten Bangla basic characters and digits using convex hull based feature set", 2009 International conference on Artificial intelligence and pattern recognition(AIPR-09).
- [5] Nafiz Arica, Student Member, IEEE, and Fatos T. Yarman-Vural, Senior Member, IEEE, "Optical Character Recognition for Cursive Handwriting", *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 6, june 2002.
- [6] Kamalanaban, E. & Gopinath, M. & Premkumar, S. (2018). Medicine Box: Doctor's Prescription Recognition Using Deep Machine Learning. *International Journal of Engineering and Technology(UAE)*. 7. 114-117. 10.14419/ijet. v7i3.34.18785.
- [7] Subhadip Basu, Nibaran Das, Ram Sarkar, Mahantapas Kundu, Mita Nasipuri, Dipak Kumar Basu, "A hierarchical approach to recognition of handwritten Bangla characters", Elsevier -2009
- [8] Abhishek Bala and Rajib Saha, "An Improved Method for Handwritten Document Analysis using Segmentation, Baseline Recognition and Writing Pressure Detection", 6th International Conference On Advances in Computing Communications, ICACC 2016, 6-8 September 2016, Cochin, India, Elsevier-2016.
- [9] Kanchan Keisham and Sunanda Dixit, "Recognition of Handwritten English Text Using Energy Minimisation", *Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing*, Bangalore, India, Springer-2016.
- [10] Namrata Dave, "Segmentation Methods for Hand Written Character Recognition", *International Journal of Signal Processing, Image Processing and Pattern Recognition* Vol. 8, No. 4 (2015), pp. 155-164.
- [11] U. Marti and H. Bunke. The IAM-database: An English Sentence Database for Off-line Handwriting Recognition. *Int. Journal on Document Analysis and Recognition*, Volume 5, pages 39 - 46, 2002.

Hindi-English Code Mixed Hate Speech Detection using Character Level Embeddings

Rahul

*Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
rahul@dtu.ac.in*

Vibhu Sehra

*Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
vibhusehra_2k17co368@dtu.ac.in*

Vasu Gupta

*Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
vasugupta_2k17co366@dtu.ac.in*

Yashaswi Raj Vardhan

*Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
yashaswirajvardhan_2k17co386@dtu.ac.in*

Abstract—Hinglish is a portmanteau word for 'Hindi' and 'English', and refers to the informal "language" predominantly used in the South-Asian (Indian) Sub-Continent, a blend of the two languages it derives its name from. It considerably differs from the English language in grammar, syntax, punctuations, phonetics and accent, as well as in sentiments.

As it is more convenient to use English for certain technical words, sports events, scientific phenomena, and other things, mixed usage of English and regional languages has gained considerable prominence in day-to-day conversations and Social Media. This research aims to create an independent and self-sufficient model that classifies Hinglish texts as Hate Speech, Abusive or Non-Offensive.

The prevalent use of code-mixed language in the subcontinent, the sensitive nature of hate speeches, and the need of a self-sufficient model for Hinglish, together serve as the motivation for this research.

We have used character level embeddings for Hinglish Language which has the potential to most efficiently extract the context from Hinglish sentences given the level of variation in syntax and semantics of the code-mixed (a language that is a combination of two or more languages) language. Later we trained various deep learning classifier models. Hybridisation of GRU with Attention Model performed best among more than 12 models experimented with. The use of Character Level Embeddings, GRU, and attention layer are novel to Hate Speech Detection in Hinglish Code-Mixed Language.

Index Terms—Hate Speech, Character-level Embeddings, Hinglish Sentiment Analysis, Deep Learning, Sequential Models, Code-Switched

I. INTRODUCTION

Social media is gaining pace, and with ever increasing users and the way in which it is being used is also changing. Users find it very convenient to blend a ubiquitous language (like English) with a native language (Hindi in this context), as it is much easier and relatable within the regional context. It is challenging to create a standard model that can classify tweets as hate-inducing or abusive, given a large number of regional languages.

Hate tweets may target a single person or express prejudice against a particular group¹, especially on the basis of race, colour, ethnicity, religion, nationality or sexual orientation, other identity factors that may extend to include a person's disability (mental or physical).

The mere presence of hate speech in public space has the potential to disrupt public order, harm communal harmony, or instigate mobs. Thus, removing these tweets is necessary and with the use of code-switched languages such as HINGLISH, it becomes difficult to remove these tweets using models designed for English tweets.

Abusive text on the other hand may be offensive in a vague sense with some degree of profanity [1]. While these texts are hurtful as well, they do not call for direct intervention. Most platforms deal with them when reported.

In 2019, India ranked first on the Social Hostilities Index, published by Pew Research Center², among the 25 most populous countries with an index value of 9.5 out of 10, rising from 8.7 (4th position) in 2014. This expresses the pressing need for a more vigilant social media moderation of hate inducing microblogs. It is also necessary that the mechanism efficiently differentiates between merely offensive from hate tweets so as to not curtail the freedom of expression, as it is guaranteed under most of the constitutions around the world [2], or create a social media policing system.

Hinglish language differs from English as well as Hindi due to the following reasons:

- Use of Roman script instead of Devanagari script (Script for Hindi)
- Absence of fixed grammatical use and high dependence on the region.
- Liberal use of more or fewer punctuations.

¹<https://www.un.org/en/genocideprevention/>

²<https://www.pewforum.org/essay/a-closer-look-at-changing-restrictions-on-religion/>

TABLE I
SAMPLE TWEETS

Tweet	Label
#teamIndia congrats on your win. #JaiHo	Non-Offensive
OMG. Jaldi ye offer use krlo. 80% off #Bachat www.abc.com	Non-Offensive
Hindu Muslim Bhai bhai #unity	Non-Offensive
Neem ka patta kadva hai, Salman s**la bha**a hai.	Abusive
@maj-tic_b-ra @ka-nj-h- R*ND* K PILLE TERI M** B*H*N KO K*THE PE BECH K AAYA HU AB	Abusive
@ai-t-in-a Bho**de k b**nch*d,4g ki offer dikhata hai!khud k ga**d k 2g v nhi hai,ch*t**a ulimited t**i	Abusive
Vi-t and A-sh-'s future kid An-h-a: Mamma bolo beta, mammaaaa Kid: Mm.. Ma.. Maa.. M*d*rc**d!	Hate-Inducing
ye mlmaano ko is desh se nikaalo	Hate-Inducing
@dasr-hu-r @na-n-amo-i @A-tSh- @BJP- M*d*rc**d brahmno se mafi mango	Hate-inducing

- Multiple variations in the transliteration of the same word.

The high magnitude of variations in the phonetics makes it challenging to interpret with reference to English speech. E.g., The slang "b*tch" can be translated to 'k*tiy*', 'k*ttiy*', 'k*tiyaa' etc. This causes a challenging situation during the preprocessing step as we want to avoid redundancies.

Merely translating the Hindi text or Hindi portions of Hinglish texts to English and then using an English Offensive Text Classification model might not always be successful as there is vast variation in the grammatical structures, and thus translations may fail to be adequate.

E.g.: 'Muje iske baare mei nhi sunna' original English translation should be 'I do not want to hear about it' but mere conversion into English leads to the sentence 'I it about in no hear' which is grammatically incorrect.

Our model classifies tweets as:

- Hate Inducing
- Abusive
- Non-offensive

Sample tweets of these categories can be seen in Table I. The degree of profanity varying in these categories can be clearly inferred from these examples.

The following models have provided us with the best robust model, capable of gaining all the relationships in the sentences:

- 1) Only GRU
- 2) Only GRU with Attention
- 3) Bidirectional LSTM with Attention + GRU: Attention after LSTM Layer

Our work can be condensed into 5 major steps:

- Preprocessing
- Creation of Character Embeddings
- Model Building

- Training & Hypertuning
- Testing trained models

Here, we have used character-level embedding models. Character level models process neither the word's semantic information nor the ecosystem of pre-trained word vectors. Instead, deep learning models working at the character level come with two key advantages: the vocabulary related issues in the input text of the model are circumvented, and on the output side, a computational bottleneck is averted.

Spelling mistakes, distorted vocabulary, and use of rare words are more common issues in Hinglish texts, as compared to any monolingual text, especially English. The use of character-level deep learning has made our model resilient to such problems and dramatically enhanced the vocabulary it can deal with. Along with this, the use of smaller tokens has made the output less computationally extensive.

The primary contributions of our work may be condensed as follows:

- Character level embeddings were used to deal with the variations in transliteration and grammatical liberties taken in Hinglish, as well as a computational bottleneck caused by word-level embedding
- An exhaustive survey of various Deep Learning architectures for training the model, including sequential stacked as well as unstacked architectures using GRU, Bi-directional LSTM, and Attention Layers, which has not been used for Hinglish Text.

The further sections of this paper contain related work, detailed methodologies, evaluation and lastly conclusion and possible future work in section 2 through 5.

II. RELATED WORK

In [3] analysis was carried out of posts on Facebook by Hinglish Bilingual users, which showed the prevalence of Code Mixed Language on Social Media, and thereby a need to monitor it. One common approach to the classification of Hinglish text has been simple translation of the text into the English language and then use a classifier suitable for English offensive texts. A breakthrough in this method came with [1] with the introduction of a Multi-Channel Transfer Learning-based model that uses Word Embedding of single words and combinations. It also uses sentimental scoring, 67 dimensions of LIWC features, and 210 dimensions of profanity vector. It utilised a hybrid model of CNN and BLSTM to carry out transfer learning and gives f1 score of 0.895.

The study in [4] compare CNN - 1D, LSTM, and Bi-directional LSTM, using domain-specific embeddings creating a 300 Dimension Vector for each word, of which CNN-1D gives the best results at F1-score of 0.8085. [5] created a dataset of Hinglish Text classified into two categories and used a supervised learning approach attaining the highest accuracy of 71.7% by use of SVM Classifier. A similar supervised classification and lexical baselining approach was used in [6], which uses character n-grams, word n-grams, and word skip grams. They were able to attain an accuracy of 78% by using three labels, separating hate speech from offensive language.

The study [7] put forward that hate speech has a presence in the 'long tail' of the dataset, and the lack of peculiar and distinct features make its detection a challenging task. They have used TF-IDF weighted word 1-gram, 2-gram, and 3-gram, number of mentions, hashtags, characters and words, alongside Part-of-Speech (PoS) tag and carried out a removal of candidates having document frequency under five. CNN+sCNN has performed better than CNN+GRU in all the tests, but the difference in F1 scores is only 1-5%.

Taking an example of work on non-English languages other than Hindi/Hinglish, [8] worked on Greek, presenting the first Offensive Greek Tweet Dataset (OGTD) containing 4,779 posts, with tweets annotated as Offensive, Not Offensive, and Spam. Similar to [7], the TF-IDF approach was taken, as well as part-of-speech (POS), and dependency relation tags and a 300-dimensional vector were used. They also experimented with some stacked and unstacked deep learning architectures obtaining the best results from LSTM and GRU with attention model. Work in [9] was carried out on a Dutch corpus from a popular question-answer-based social media platform Ask.fm, while [10] worked on a corpus of Facebook posts from anti-Islamic groups. [11] has worked on a corpus of hate tweets in German, targeting refugees.

The approach used in [12] for hate speech detection explore using Bag of words, N-grams, Character level n-grams, which might help with spelling problems as well as the frequency of @ mentions, punctuations, token length, words not found in English dictionary, variety of characters that are non-alphanumeric in tokens. As hate speech is applied to a small chunk of text, we might face the problem of sparsity with Bag of Words, and in that case, Word Generalisation comes into play. It may be done by word clustering or using the Latent Dirichlet Allocation(LDA) topic distribution technique. Thus, word embedding emerges as a feature. Sentiment analysis, lexical analysis, as well as linguistic features like POS and Dependency Relationships were examined.

Multi-tiered pipeline was created in [13]. The first being profanity modeling, second being deep graph embeddings, and the last one, author profiling. Their work uses targeted hate embeddings combined with social network-based features on various baselines and two real-world datasets. They have included an expert-in-the-loop algorithm within the pipeline framework for debiasing.

III. METHODOLOGY

A. Preprocessing

The data obtained was streamed through a series of preprocessing steps. Preprocessing is a crucial step as raw text often contains data that may contain errors and redundant information that can deter the model's training.

The first step in the preprocessing was as follows:

- 1) Converting tweets to lowercase
- 2) Removing all the @ mentions.
- 3) Removing hashtags (#) from the tweet. E.g., this is a # tweet → this is a tweet

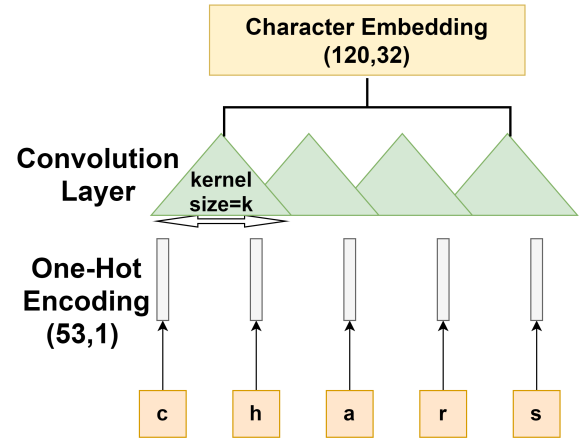


Fig. 1. Character-Level Embeddings Generation

- 4) Removing web links.
- 5) Removing digits

B. Character-Level Embeddings Generation

After preprocessing, the tweets are converted into character embeddings through the following process:

- 1) The average length of a tweet is determined. In this dataset, it is found that most of the preprocessed tweets are no longer than 120 characters. Therefore, the maximum length is taken as 120.
- 2) The total number of unique characters is determined in the tweets. We have 53 distinct characters after preprocessing.
- 3) The tweets are converted to one-hot vectors. Each tweet is of shape (MAXIMUM_LENGTH, TOTAL_CHARS) = (120,53)
- 4) These are then passed through a 1D-Convolutional Neural Network (CNN) layer with 32 kernels and kernel size being three, to obtain the character embeddings as shown in Figure 1.

C. Model Building

The output from the embedding layer is passed through our deep learning models, which predicts the probability of belonging to each class of this imbalance class classification scenario.

We have used Convolutional 1D Layers (same as in embedding layer), Batch Normalization Layer for faster convergence as discussed in [14], and Dense layers. We used ReLU activation with Convolutional Layers and the Dense Layers, the exception being the last dense layer. This layer's activation function was softmax.

The basic architecture for CNN, Bi-LSTM, and GRU are described as below:

CNN Model (Figure 2)

EMB_OUPUT → CONV1D(16,3) → BATCH-NORM() → CONV1D(12,3) → BATCHNORM()

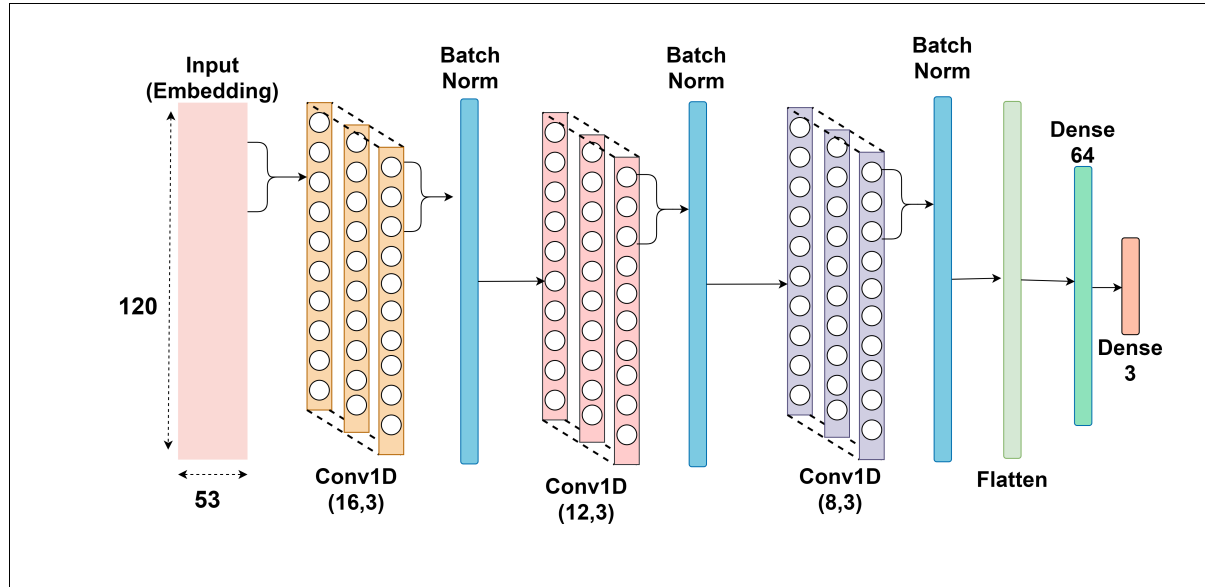


Fig. 2. Proposed CNN Model

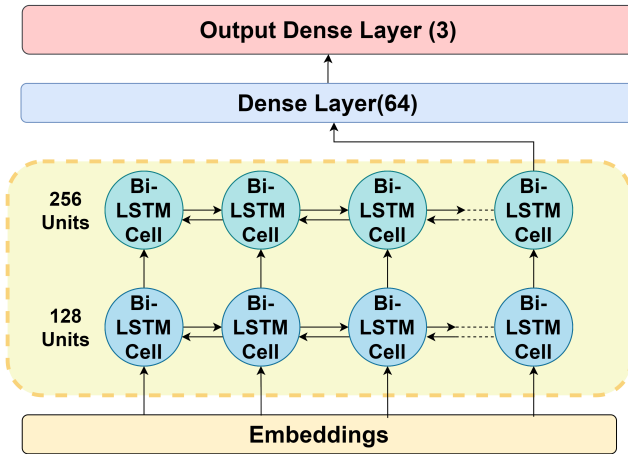


Fig. 3. Proposed Bi-LSTM Model

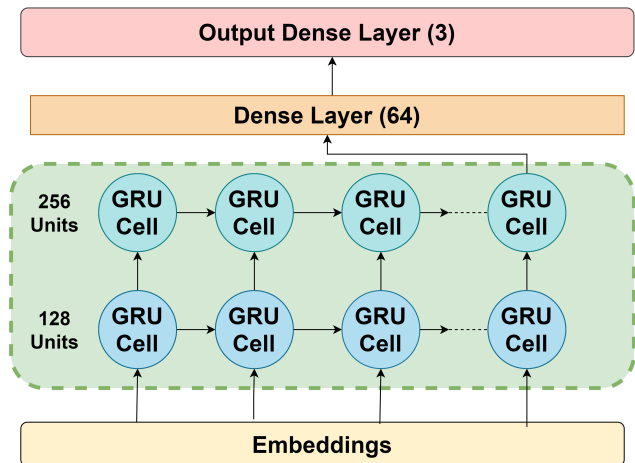


Fig. 4. Proposed GRU Model

→ CONV1D(8,3) → BATCHNORM() → DROPOUT(0.2)
→ FLATTEN() → DENSE(64) → DENSE(3)

Bidirectional LSTM Model (Figure 3)

EMB_OUPUT → BLSTM(128) → DROPOUT(0.25)
→ BLSTM(256) → DROPOUT(0.25) → DENSE(64) →
DENSE(3)

GRU Model (Figure 4)

EMB_OUPUT → GRU(128) → DROPOUT(0.25) →
GRU(256) → DROPOUT(0.25) → DENSE(64) →
DENSE(3)

Bidirectional GRU Model

EMB_OUPUT → BGRU(128) → DROPOUT(0.25) →
BGRU(256) → DROPOUT(0.25) → DENSE(64) →
DENSE(3)

We then used the above three basic architectures to make a combination of different models. We concatenated the output of the dense(64) layer of individual models and finally added a final dense output layer. The intent behind this was to see if combining the two models could give us an added advantage over the use of these models separately. CNN being a non-sequential model, whereas, Bi-LSTM and GRU being sequential models have their own set separate of advantages. All the different models created are listed below:

- CNN + GRU
- Bi-LSTM + GRU

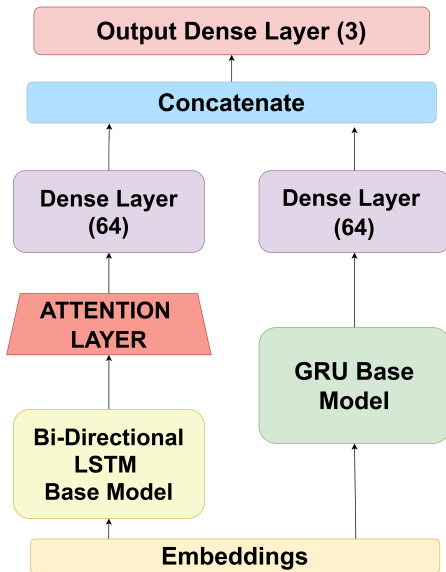


Fig. 5. Proposed Bi-LSTM(with Attention) + GRU Model

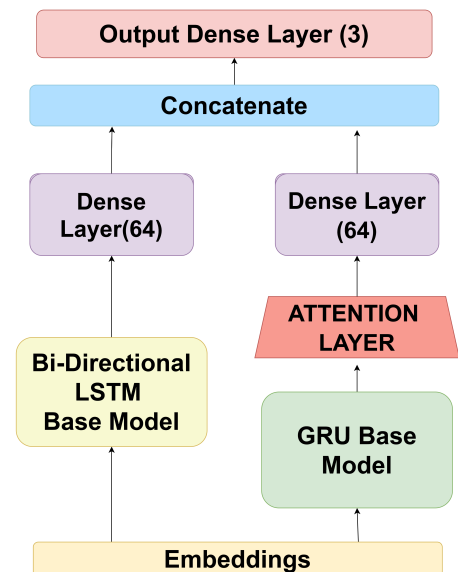


Fig. 6. Proposed Bi-LSTM + GRU Model(with Attention)

• CNN + Bi-LSTM

The idea behind stacking the dense layers from the two models (Bi-LSTM and GRU) is that when we concatenate the two layers, we are able to use the features extracted from the two models and add them to improve the overall understanding of the model.

Further, an attempt was made to enhance the results with the help of the attention layer [15], as it helps to provide weightage to output of the previous layer and pass on that weighted output to the next layer for better learning because some parts of the data are of greater importance for the prediction as compared to others. attention layer. We introduced attention in the following models:

- Bi-LSTM
- GRU
- Bi-GRU
- Bi-LSTM + GRU (with attention layer after Bi-LSTM layer) (Figure 5)
- Bi-LSTM + GRU (with attention layer after GRU layer) (Figure 6)

IV. EVALUATION

A. Dataset

The dataset is obtained from [1], and it contains 3189 tweets. This dataset contains tweets written in the Hinglish text, classified into three categories based on the profanity in the text. The three categories were namely: Hate Inducing, Abusive, and Non-Offensive. After preprocessing, some of the tweets were reduced to length zero. The results of baseline models like SVM and Random Forest using Character n-grams, Bag of Words and TF-IDF give an F1-score in the range of 0.574-0.723. The best performing baseline model is SVM classifier with TF-IDF feature with F1-score of 0.723.

TABLE II
DATASET DESCRIPTION

Label	Tweets Count
Non-Offensive	1018
Abusive	1764
Hate-Inducing	303
Total	3085

So, Table II is the final dataset size after preprocessing.

Further, we split the formed dataset with the train-test size ratio being 80:20. We used sklearn's train test split with random state 6. The final division of tweets belonging to various classes can be seen in Table VI.

B. Training Details

For the creation of model architectures, Keras Library with Tensorflow backend was used. To train each model, we used a loss function named Categorical Cross Entropy. We used Adam optimizer [16] with learning rate = 10^{-3} . All the model architectures were trained using ten-fold cross-validation. The batch size was kept as 8, and the model was trained for 50 epochs.

We used L2 regularization [17] with $\lambda = 10^{-5}$ in each layer as a kernel regularizer to prevent the model from overfitting as our dataset is relatively small, and there is a high probability of the model being overfitted.

C. Results and Discussion

As stated in the previous section, a train-test split of 80:20 had reserved 20% of the dataset for testing, which was used for the preparation of our results. A common problem in a classification scenario is when the ratio of observations in each class is disproportionate; that is, the distribution is biased or skewed, as is the case with our dataset, making it an imbalanced class. Therefore, accuracy is not an apt measure

TABLE III
RESULTS WITH BASIC PROPOSED MODELS

Model Name	Accuracy	F1-Score	Precision	Recall
CNN Model	0.72	0.71	0.70	0.72
Bidirectional LSTM Model	0.85	0.85	0.86	0.85
GRU Model	0.86	0.86	0.86	0.86
Bidirectional GRU Model	0.85	0.85	0.86	0.85

TABLE IV
RESULTS WITH STACKED MODELS

Model Name	Accuracy	F1-Score	Precision	Recall
CNN + GRU Model	0.72	0.71	0.71	0.72
Bidirectional LSTM + GRU Model	0.84	0.84	0.84	0.84
CNN + Bidirectional LSTM Model	0.73	0.71	0.70	0.73

TABLE V
RESULTS WITH MODELS CONTAINING ATTENTION LAYER

Model Name	Accuracy	F1-Score	Precision	Recall
Bidirectional LSTM with Attention Model	0.85	0.86	0.86	0.85
GRU with Attention Model	0.87	0.87	0.87	0.87
Bidirectional GRU with Attention Model	0.85	0.85	0.85	0.85
Bidirectional LSTM with Attention + GRU Model	0.86	0.86	0.87	0.86
GRU with Attention + Bidirectional LSTM Model	0.84	0.84	0.84	0.84

TABLE VI
DISTRIBUTION AFTER TRAIN-TEST SPLIT

Label	Train	Test
Non-Offensive	815	203
Abusive	1406	358
Hate-Inducing	247	56
Total	2538	651

for the performance of the models; instead, f1-score was taken for this purpose. F1-score is the harmonic mean of Precision, that is the positive predictive value, and Recall, the measure of sensitivity of the model.

The predictions were generated by passing the 20% data, previously reserved as test-data, through the models. The results were then tabulated using classification reports of the SciKitLearn library of python, which takes predicted and actual values of classes as function parameters.

The predictions were generated by passing the 20% data, previously reserved as test-data, through the models. The results were then tabulated using classification reports of the SciKitLearn library of python, which takes predicted and actual values of classes as function parameters.

The focus of this work was on implementing Character Level Embedding. This was done to avert the demerits of using word-level embeddings with Hinglish text, such as flawed grammar and vocabulary, as well as a computational bottleneck.

An exhaustive trial of various deep learning models and their layered combinations was carried out. It started with basic model architectures like CNN, Bi-LSTM and GRU. Table III shows results for the same. It can be seen, CNN did not perform well which was expected as CNN is too non sequential to suit in this case. Whereas, sequential models like GRU and LSTM performed significantly better. GRU models perform the best, with F1-Score of 0.86. This depicts that sequential architecture is able to extract suitable features from the sentiments and categorize them accordingly.

Even though, the individual performance of CNN model was below par with an f1-score of 0.71, , we stacked it with other sequential models with an idea to combine their advantages, however, it can be seen from Table IV, that stacked CNN Models did not perform well either, giving similar accuracy and f1-score.

For further experimentation, we drop the usage of CNN and introduce Attention Layer to our previously best performing architectures. Table V depicts the results for the same. The introduction of Attention layer enhances the performance in nearly all cases. This was expected as contribution of all parts is not uniform in building the meaning and sentiments of the sentence, and attention layer mechanism facilitates by giving more focus to those tokens which contribute more. The GRU model with Attention layer is the best performing model, among the various combinations we have attempted. It has the highest f1-score, accuracy, precision and recall at 0.87 each.

This was followed by Bidirectional LSTM with Attention layer and GRU, with an f1-score, accuracy and recall of 0.86 and precision at the same level as the previous model at 0.87.

In comparison, these results surpass the related work done in Hinglish Language like Bohra et al, which uses Random Forest Classifier and Support Vector Machine to achieve a maximum accuracy of 0.699 and 0.717 respectively. Kamble et al also remain limited to a highest f1-score of 0.808 and 0.804 with the use of CNN and Bi-LSTM respectively.

Our approach also significantly outperforms other works on Code-Mixed languages. Tulkens et al's detection of racism in Dutch social media using a dictionary based approach has a maximum f1-score of 0.50 and a highest AUC of 0.63 only.

V. CONCLUSION AND FUTURE WORK

We experimented with 12 model architectures to train our model on Character Level Embedding for multiclass labeling of Hinglish tweets into three categories(non-offensive, abusive, hate-inducing tweets). Among them, three models have performed considerably better than the rest and are recommended for use in this work: Only GRU, Only GRU with Attention, Bidirectional LSTM + GRU: with Attention after LSTM Layer. Our approach is robust and capable of learning complex dependencies known today or which may arrive in the near future.

The use of Character-Level Embedding instead of word embeddings has made our model resilient to the common defect in most Hinglish Code Mixed projects, which is the use of distorted vocabulary and a multitude of spelling mistakes. A possible enhancement to our work can be done if a larger corpus of tweets in Hinglish is available for the training of models. Another possibility is using an architecture that combines word-level embeddings and character-level embeddings while still averting the computational bottleneck caused by word-level embeddings and resolving the shortcomings of the character level approach. The use of transformer models has been proposed by various authors [18], [19], [20] and it may also enhance the working of our model.

REFERENCES

- [1] P. Mathur, R. Sawhney, M. Ayyar, and R. Shah, "Did you offend me? Classification of Offensive Tweets in Hinglish Language," 2019, pp. 138–148.
- [2] C. O'Regan, "Hate speech Online: An (intractable) contemporary challenge?" *Current Legal Problems*, 2018.
- [3] K. Bali, J. Sharma, M. Choudhury, and Y. Vyas, "'I am borrowing ya mixing ?' An Analysis of English-Hindi Code Mixing in Facebook," 2015.
- [4] S. Kamble and A. Joshi, "Hate speech detection from code-mixed Hindi-english tweets using deep learning models," 2018.
- [5] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, "A Dataset of Hindi-English Code-Mixed Social Media Text for Hate Speech Detection," 2018.
- [6] S. Malmasi and M. Zampieri, "Detecting hate speech in social media," in *International Conference Recent Advances in Natural Language Processing, RANLP*, 2017.
- [7] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? The challenging case of long tail on Twitter," *Semantic Web*, 2019.
- [8] Z. Pitenis, M. Zampieri, and T. Ranasinghe, "Offensive language identification in Greek," in *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 2020.
- [9] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste, "Automatic detection and prevention of cyberbullying," *International Conference on Human and Social Analytics (HUSO 2015)*, 2015.
- [10] S. Tulkens, L. Hilde, E. Lodewyckx, B. Verhoeven, and W. Daelemans, "The Automated Detection of Racist Discourse in Dutch Social Media," in *Computational Linguistics in the Netherlands Journal*, 2016.
- [11] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, "Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis," 2017.
- [12] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," 2017.
- [13] S. Chopra, R. Sawhney, P. Mathur, and R. Ratn Shah, "Hindi-English Hate Speech Detection: Author Profiling, Debiasing, and Practical Perspectives," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
- [14] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *32nd International Conference on Machine Learning, ICML 2015*, 2015.
- [15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *ArXiv*, vol. 1409, 09 2014.
- [16] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.
- [17] J. Schmidhuber, "Deep Learning in neural networks: An overview," 2015.
- [18] K. Pant and T. Dadu, "Towards Code-switched Classification Exploiting Constituent Language Resources," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*. Suzhou, China: Association for Computational Linguistics, dec 2020, pp. 37–43. [Online]. Available: <https://www.aclweb.org/anthology/2020.aacl-srw.6>
- [19] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A bert-based transfer learning approach for hate speech detection in online social media," 12 2019.
- [20] R. Mutanga, N. Naicker, and O. O., "Hate speech detection in twitter using transformer methods," *International Journal of Advanced Computer Science and Applications*, vol. 11, 01 2020.



Impact of multi threshold transistor in positive feedback source coupled logic (PFSCCL) fundamental cell

Ranjana Sivaram¹ · Kirti Gupta² · Neeta Pandey¹

Received: 18 April 2020 / Revised: 18 April 2020 / Accepted: 21 April 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

In this paper, a new fundamental cell in positive feedback source coupled logic is presented, which is an improvement over the existing fundamental cell employed in digital circuit design in various high resolution mixed-signal integrated circuits. The operation of the existing fundamental cell relies on using large sized transistor in its centre branch, resulting in significantly larger implementation area. The proposed fundamental cell incorporates multi-threshold transistor in the center branch thereby allowing designer to use reduce its dimension and hence the area. The impact of the proposed modification is examined by configuring the cell as two input exclusive OR (XOR2) gate. The behaviour is analysed in terms of static and propagation delay parameters which are modelled and a design procedure is also elaborated. The theoretical prepositions are verified by designing and simulating for various operating conditions using model parameters of 180 nm CMOS technology node. A maximum error of 27% is observed between the simulated and predicted parameters. The process variation study through Monte Carlo analysis and PVT variations identifies the proposed fundamental cell based circuit as less prone to variations in comparison to existing fundamental cell based counterparts. A full adder, as an application of the proposed fundamental cell, shows a significant (66%) area reduction while delay, power and PDP are within 4% of their corresponding values for the existing one.

Keywords Mixed-signal · Digital circuit · SCL · PFSCCL · Fundamental cell

1 Introduction

The advancement in integrated circuit technology has influenced the level of integration and facilitated mixed signal designs wherein analog and digital functions are encapsulated on the same substrate. The CMOS logic style is long established in designing digital functions due to negligible static power consumption and design ease [1]. It,

however, generates large switching noise which may lead to malfunctioning of analog circuit housed on the same substrate. Source coupled logic (SCL) is suggested as an alternative for mixed signal environment due to its inherent features such as frequency independent power consumption, high speed and low noise [2–4]. Two variants of SCL exist in literature—differential SCL popularly called MOS Current Mode Logic (MCML) and single ended SCL. The logic functions are implemented through series gating approach in differential SCL, while with single ended SCL, NOR/OR based implementation is used [5].

Positive feedback source coupled logic (PFSCCL) [6] is an improved form of single ended SCL that provides lower delay and smaller area compared to traditional counterpart. The PFSCCL based complex logic implementation translates into cascading of multiple gates because of NOR/OR based implementation scheme. A Fundamental Cell (FC), based on triple tail concept, is introduced in [7] to mitigate this issue. It reduces the number of gates required for logic implementation and leads to improved performance in

✉ Neeta Pandey
n66pandey@rediffmail.com

Ranjana Sivaram
ranjanasridhar@gmail.com

Kirti Gupta
kirtigupta22@gmail.com

¹ Department of Electronics and Communication Engineering,
Delhi Technological University, Delhi, India

² Department of Electronics and Communication Engineering,
Bharati Vidyapeeth's College of Engineering, New Delhi,
India

terms of speed and power consumption in comparison to PFSCCL NOR/OR based implementation scheme. The concept of fundamental cell is generalized in [8] which defines a configurable logic block (CLB) that results as an efficient circuit realisation technique and based on this, complex circuits such as comparators, adders, multipliers, test pattern generators etc. [9–13] have been implemented. Thus, the scheme is attractive, it however, requires bigger size transistors for proper operation of fundamental cell [7]. With the increasing focus on the design with smaller area, a modified fundamental cell is proposed in this paper, where multiple threshold voltage transistors are introduced so as to reduce transistor dimensions and hence, a reduction in the area.

In Sect. 2, a brief discussion on the existing fundamental cell is presented. The proposed fundamental cell is presented in Sect. 3. Its behaviour is elaborated and is modelled in terms of the static and delay parameters. The impact of the threshold voltage reduction factor on the delay and area is studied and a comparative discussion on area reduction is included. The proposed fundamental cell and the existing fundamental cell based XOR gates are designed and simulated for performance comparison. To study the behaviour of the proposed fundamental cell under process variations, Monte Carlo analysis and simulations under process-supply voltage-temperature are performed. In Sect. 4, a full adder based on the proposed fundamental cell has been designed and its performance has been evaluated through simulations. The paper is concluded in Sect. 5.

2 Fundamental cell

A fundamental cell is a circuit element which is being used in PFSCCL style for efficient function realizations. The cell is similar to a conventional PFSCCL gate i.e. it comprises of a pull down network (PDN), load and current source. The PDN of the cell realizes the functionality by employing two triple-tail cells biased by separate current sources of $I_{SS}/2$ value so that the total current drawn in the cell and conventional PFSCCL gate remains the same. Depending on the inputs, the bias current gets steered in the triple-tail cell and an appropriate output is obtained through the current to voltage conversion across the load transistors. The fundamental cell configured as a two input AND (AND2) gate is shown in Fig. 1a. The PDN consists of two triple-tail cells-TT-1: (Md1, Mc1, Md2) and TT-2: (Md3, Mc2, Md4). The centre transistor (Mc1/Mc2) in each triple-tail cell is connected between the power supply and the respective source-coupled node. The transistors Ms1 and Ms2 operate in saturation in order to maintain a constant bias current of $I_{SS}/2$ value. The four PMOS transistors (Mr1, Mr2, Mr3

and Mr4) work as load. At any given time, either of the two cells (TT-1/TT-2) gets activated and determines the output of the gate. An activated cell has its centre transistor OFF such that the bias current gets steered through its outer transistors. In the AND2 gate, for the case when B is asserted high, TT-1 is activated while TT-2 gets deactivated. The bias current $I_{SS}/2$ then flows through the transistor pair (Md1-Md2) and the output is generated accordingly. The other TT-2 does not contribute to the output since whole of the bias current $I_{SS}/2$ flows through Mc2.

The concept of fundamental cell is generalized by defining a configurable logic block (CLB) [8] and is being configured for realizing various other two input logic functions as well as 2:1 multiplexer. Its usage has also been extended to efficiently realise complex circuits such as comparators, adders, multipliers, test pattern generators etc. [9–13].

The use of fundamental cell offers high performance circuits but there exists a limitation in terms of area requirement. The proper operation of fundamental cell requires that the complete bias current $I_{SS}/2$ should flow through the centre transistor in a deactivated triple tail cell. But in practice it is difficult to achieve since the bias current $I_{SS}/2$ divides between the centre and one of the two outer transistors as both of them are driven by high inputs. To address this limitation and facilitate proper activation/deactivation, the aspect ratio of centre transistors is made N times of the outer transistors [7]. However, it is obvious that while realizing complex function this approach leads to significant area overhead due to larger aspect ratio of centre transistors. Thus, in the next section, an alternate mechanism for the activation/deactivation of the fundamental cell is proposed and its behaviour is analysed in terms of static parameters and delay.

3 Proposed fundamental cell

The proposed fundamental cell achieves activation/deactivation in triple-tail cells by lowering the threshold voltage of the centre transistor by a factor α in comparison to the outer transistors. The schematic of a generic proposed fundamental cell is shown in Fig. 2a. The PDN has two modified triple-tail cells MTT-1 and MTT-2 each biased by $I_{SS}/2$ with the inputs A, B and M. In the schematic, the low threshold voltage center transistors are made bold to differentiate from others having typical threshold voltage. The transistors Mr1-Mr4 act as loads and the output voltage is generated by combining any one of the two output nodes of modified triple-tail cells i.e. (MTT-1: either Q1 or Q2; MTT-2 either Q3 or Q4).

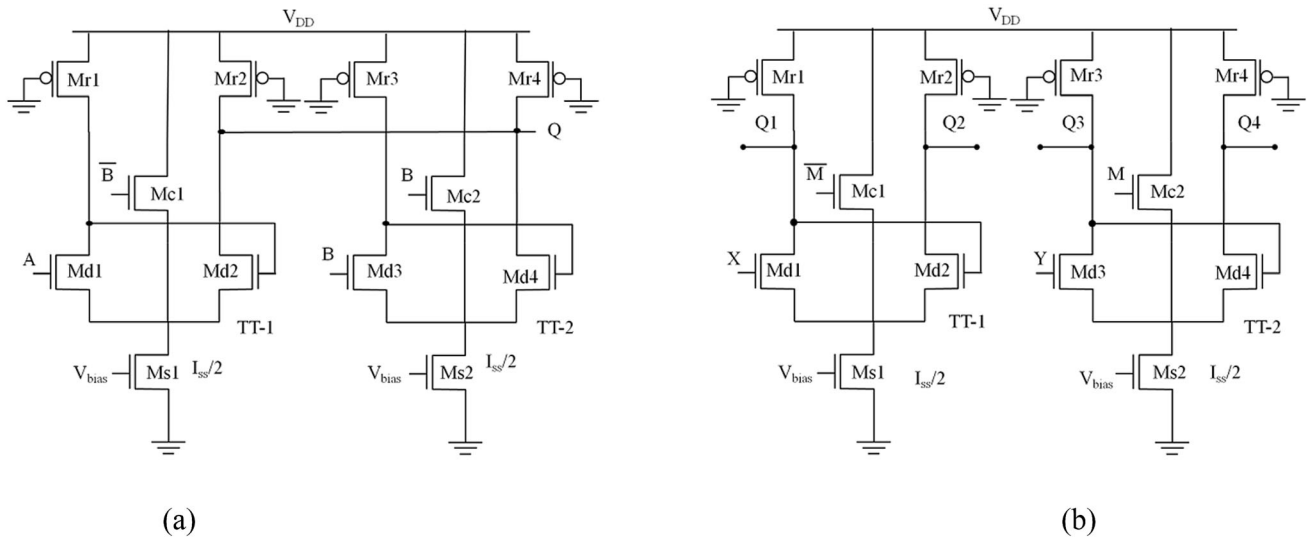


Fig. 1 Conventional fundamental cell as (a) AND2 gate [7] (b) configurable logic block (CLB) [8]

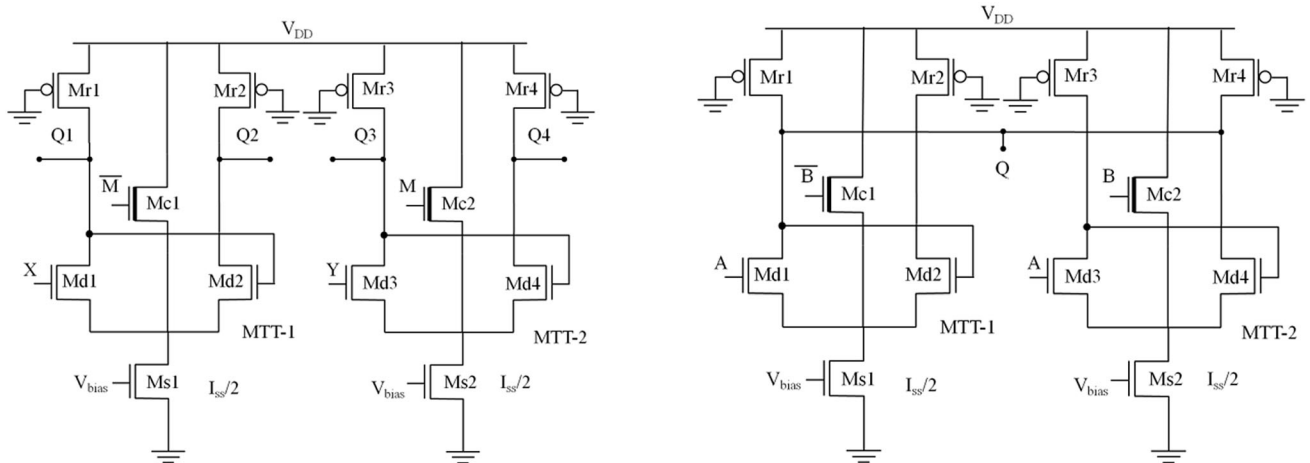


Fig. 2 a Proposed fundamental cell b Proposed fundamental cell based XOR2 gate

The study on proposed fundamental cell is performed by configuring it as a two input XOR (XOR2) gate. The schematic of the proposed fundamental cell based XOR2 gate is shown in Fig. 2b. The input B and its complement drives the low threshold voltage centre transistor Mc1 in MTT-1 and Mc2 in MTT-2 respectively whereas input A is connected to the outer transistors of the MTTs. The output node Q is obtained by connecting Q1 and Q4 from MTT-1 and MTT-2 respectively. For low value of input B, MTT-2 is activated while MTT-1 is deactivated; therefore the input A is available at the output. Analogously, for high values of input B, complement of A is available at the output as MTT-1 is activated and MTT-2 is deactivated. Thus, the functionality of the gate can be modelled as:

$$Q = \begin{cases} A & \text{if } B = 0 \\ \bar{A} & \text{if } B = 1 \end{cases} \quad (1)$$

Simulations have been carried out to verify the behaviour of the proposed fundamental cell based XOR2 gate and the results are shown in Fig. 3 by considering power supply and voltage swing of 1.1 V and 0.4 V respectively. It can be observed that for the cases when input B is at low logic level, the output is same as input A while it is complement of input A otherwise. Thus, the waveforms confirm the correct behaviour. This behaviour of the proposed fundamental cell is modelled in terms of static and delay parameters.

After this, it is necessary to provide an insight to the currents in MTTs of the proposed fundamental cell based XOR2 gate. In a deactivated MTT, it is clear that one of the outer transistors and the centre transistor are ON. Therefore, in such a situation, based on our design assumption that the threshold voltage of the center transistor is lower

Fig. 3 Proposed fundamental cell based XOR2 input output

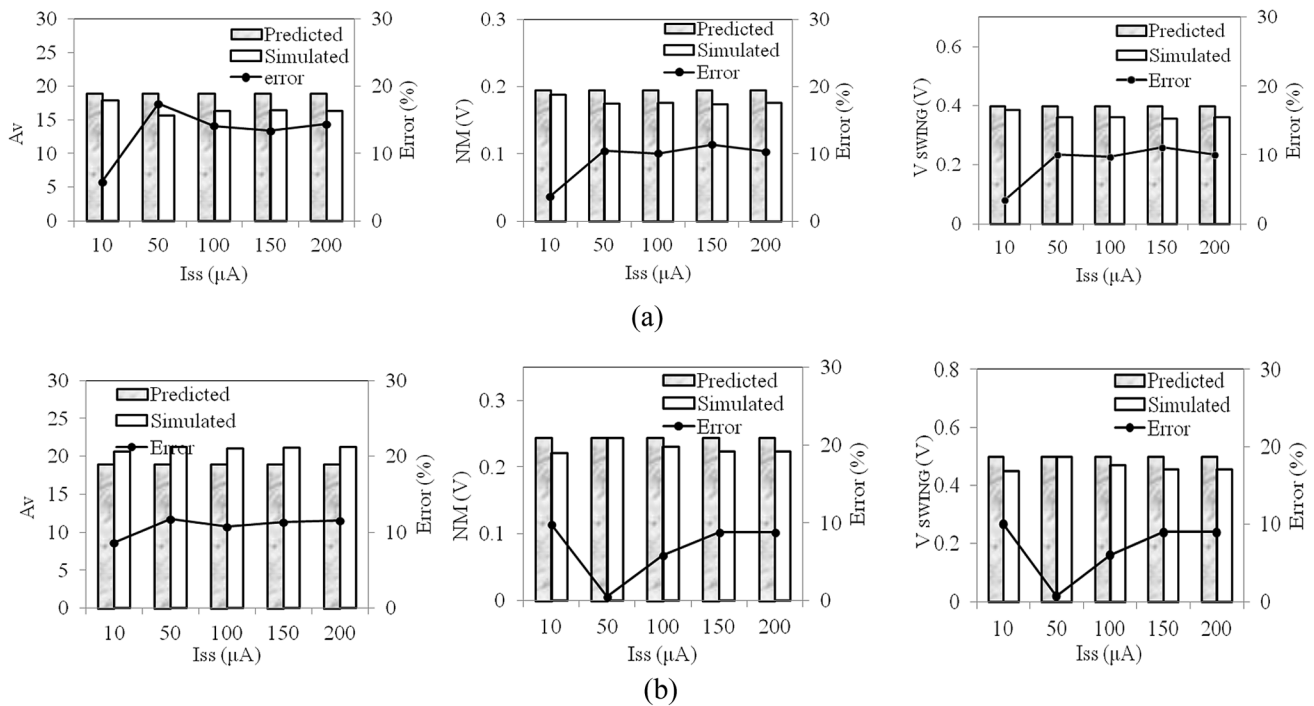
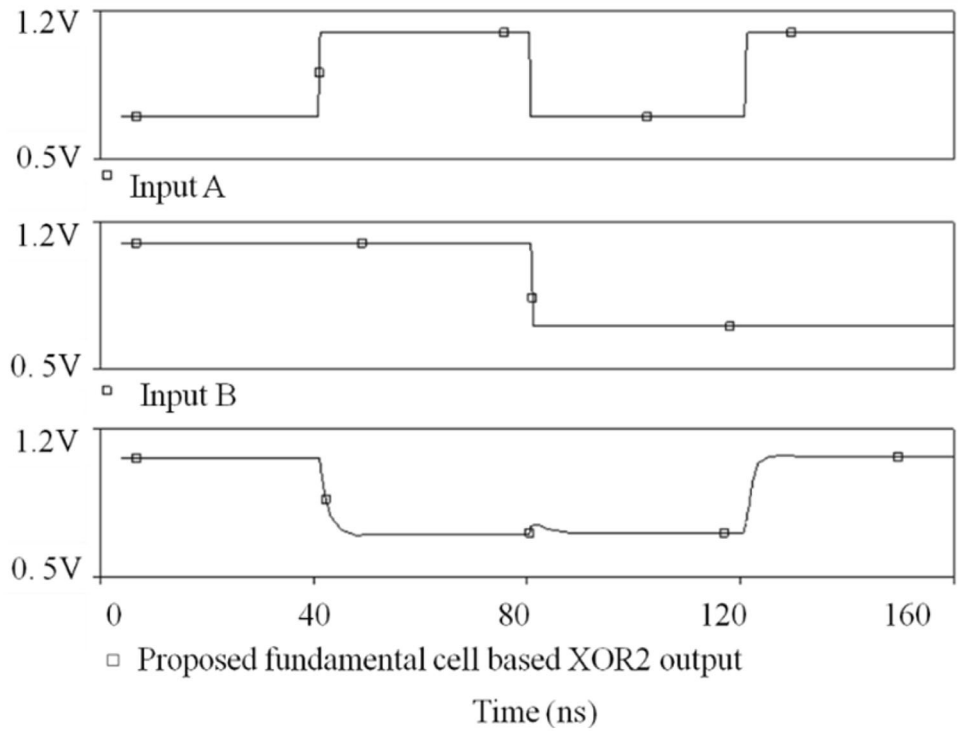


Fig. 4 Predicted and Simulated results with error versus Bias Current for static parameters with V_{SWING} of **a** 0.4 V and **b** 0.5 V

than the outer by α factor, the currents through the i th centre transistor (I_{Ci}) and the j th ON outer transistor (I_{Dj}) where $i \in (1,2)$ and $j \in (1,4)$, can be expressed as:

$$I_{Ci} = \frac{\mu_{eff} C_{ox} W_N}{2 N} (V_{GS} - \frac{V_{TN}}{\alpha})^2 \quad (2)$$

$$I_{Dj} = \frac{\mu_{eff} C_{ox} W_N}{2 L_N} (V_{GS} - V_{TN})^2 \quad (3)$$

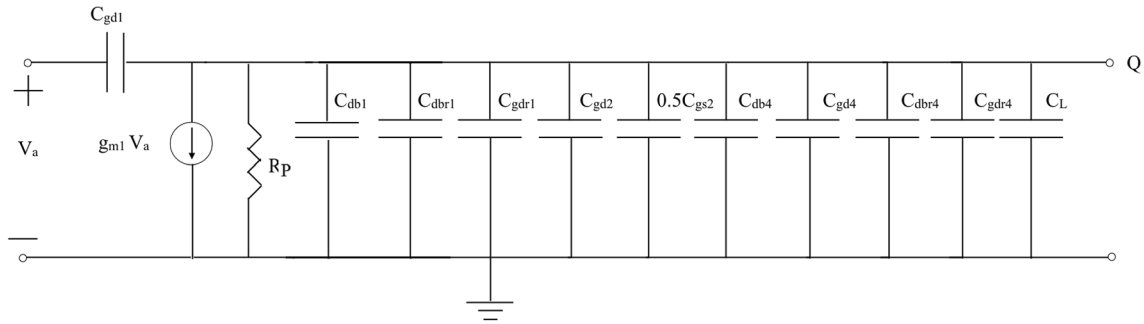
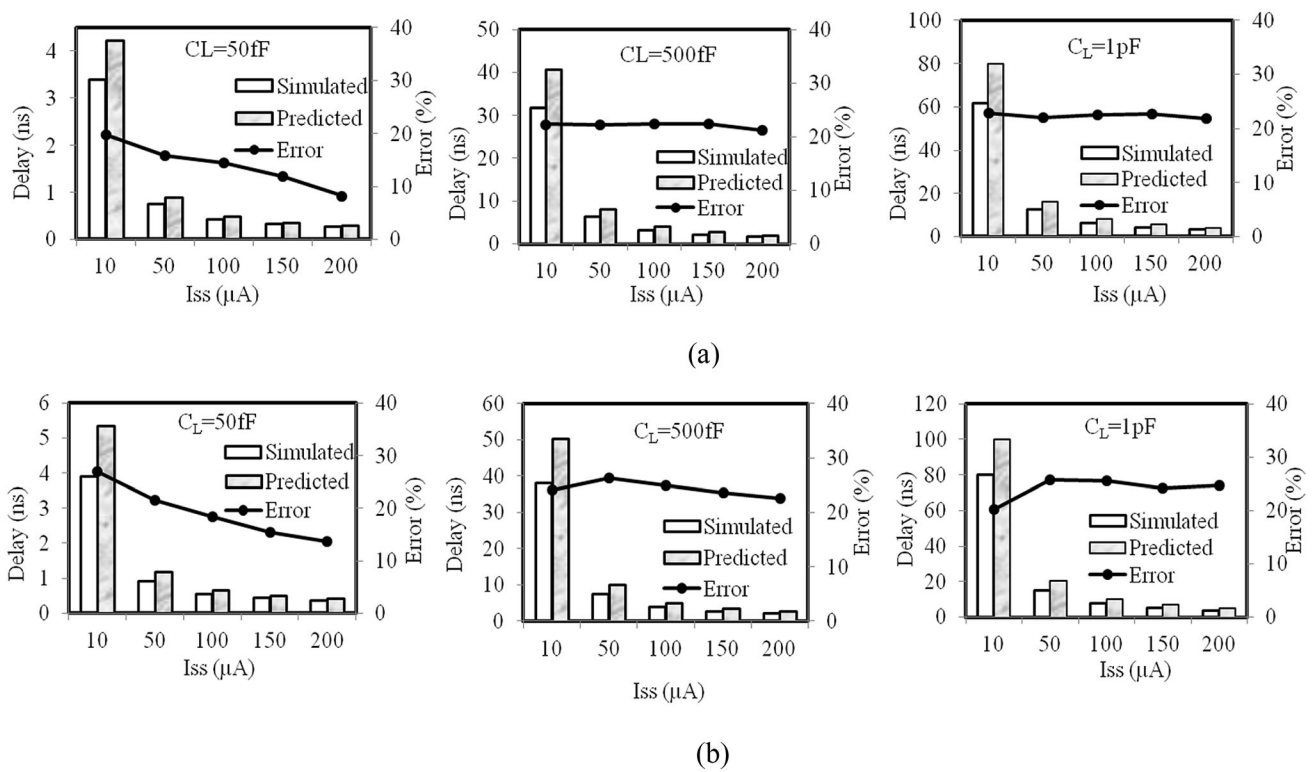
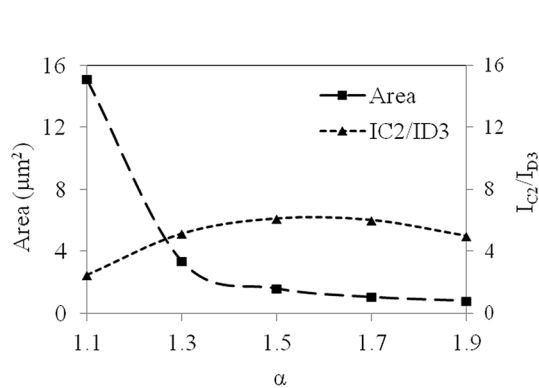


Fig. 5 Linear half circuit


 Fig. 6 Predicted and Simulated results with error versus Bias Current for **a** $V_{\text{SWING}} = 0.4 \text{ V}$ **b** $V_{\text{SWING}} = 0.5 \text{ V}$

 Fig. 7 Area vs I_{C2}/I_{D3} vs α for proposed fundamental cell based XOR2 gate

μ_{effn} , V_{GS} and V_{TN} are the effective electron mobility, the gate source voltage and the threshold voltage of NMOS transistor respectively. As each MTT-1 and MTT-2 are biased by the current source with bias current $I_{\text{SS}}/2$, the two currents can be related as:

$$I_{C_i} + I_{D_j} = \frac{I_{\text{SS}}}{2} \quad (4)$$

Using (2)–(4), I_{C_i} and I_{D_j} are derived as:

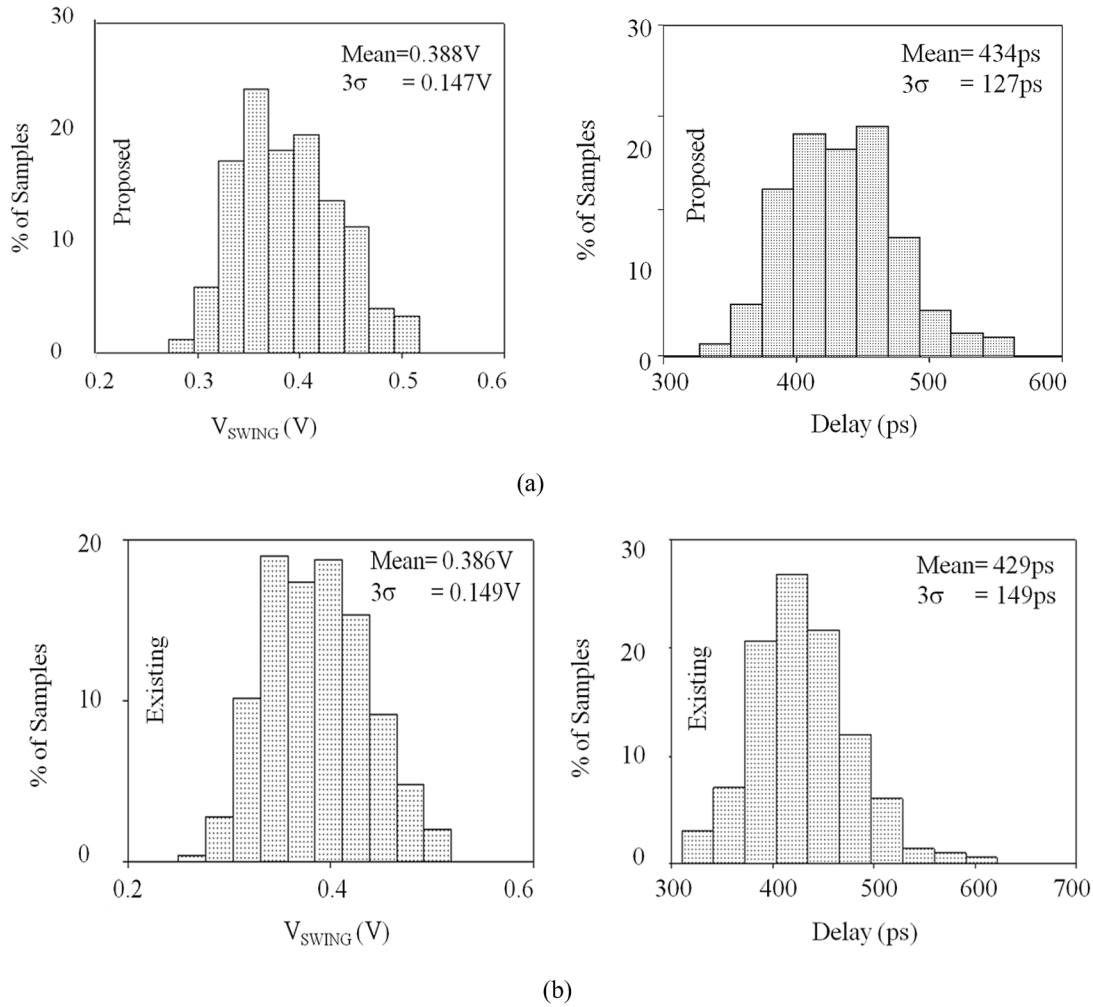


Fig. 8 Monte Carlo results for V_{SWING} and Delay for **a** proposed and **b** existing fundamental cell based XOR2 gate

$$I_{Ci} = \frac{I_{ss}}{2} \cdot \frac{1}{2} + \frac{\sqrt{\frac{\mu_{effn} C_{ox} W_N}{2 L_N} V_{TN}^2 \left(\frac{\alpha-1}{\alpha}\right)^2} \sqrt{\left(I_{ss} - \frac{\mu_{effn} C_{ox} W_N}{2 L_N} V_{TN}^2 \left(\frac{\alpha-1}{\alpha}\right)^2}\right)}{I_{ss}} \quad (5)$$

$$I_{Dj} = \frac{I_{ss}}{2} \left(1 - \frac{1}{2} - \frac{\sqrt{\frac{\mu_{effn} C_{ox} W_N}{2 L_N} V_{TN}^2 \left(\frac{\alpha-1}{\alpha}\right)^2} \sqrt{\left(I_{ss} - \frac{\mu_{effn} C_{ox} W_N}{2 L_N} V_{TN}^2 \left(\frac{\alpha-1}{\alpha}\right)^2}\right)}{I_{ss}}\right) \quad (6)$$

Substituting

$$p = \frac{1}{2} + \frac{\sqrt{\frac{\mu_{effn} C_{ox} W_N}{2 L_N} V_{TN}^2 \left(\frac{\alpha-1}{\alpha}\right)^2} \sqrt{\left(I_{ss} - \frac{\mu_{effn} C_{ox} W_N}{2 L_N} V_{TN}^2 \left(\frac{\alpha-1}{\alpha}\right)^2}\right)}{I_{ss}},$$

the current Eqs. (5) and (6) are simplified as:

$$I_{Ci} = \frac{I_{ss}}{2} \cdot p \quad (7)$$

$$I_{Dj} = \frac{I_{ss}}{2} (1-p) \quad (8)$$

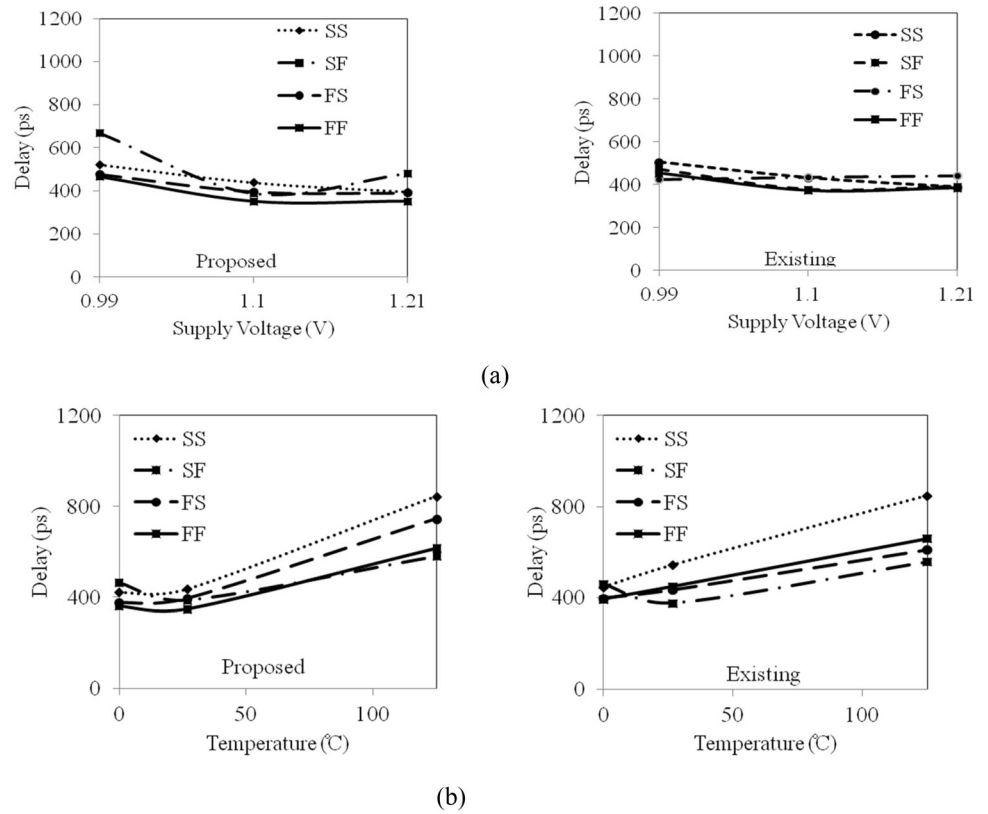
Based on the derived current expression, the output voltage for different input combinations fed to the proposed fundamental cell based XOR2 gate (Fig. 2b) is obtained and used to derive the V_{SWING} defined as the difference between high output voltage (V_{OH}) and low output voltage (V_{OL}) subsequently.

Case 1. Both inputs (A and B) are at high logic level: In this condition, the transistors Md1, Md3 and Mc2 are ON and transistors Mc1, Md2, Md4 are OFF. The current through the transistors Md1, Md3 and Mc2 is written as:

$$I_{D1} = \frac{I_{ss}}{2}; \quad I_{C2} = \frac{I_{ss}}{2} \cdot p \quad I_{D3} = \frac{I_{ss}}{2} (1-p) \quad (9)$$

This input condition produces a low output voltage V_{OL} computed as:

Fig. 9 PVT analysis results for Delay versus **a** V_{DD} **b** Temperature for proposed and existing fundamental cell based XOR2 gate



$$V_{OL} = V_{DD} - \frac{R_p I_{SS}}{2} \quad (10)$$

Case 2. Both inputs (A and B) are at low logic level: The transistors Md2, Mc1 and Md4 are ON and transistors Md1, Mc2, Md3 are OFF in this case. Therefore, the current through the transistors Md2, Mc1 and Md4 is found as:

$$I_{D2} = \frac{I_{SS}}{2}(1-p); I_{C1} = \frac{I_{SS}}{2} \cdot p; I_{D4} = \frac{I_{SS}}{2} \quad (11)$$

This input condition corresponds to low output voltage V_{OL} given as:

$$V_{OL} = V_{DD} - \frac{R_p I_{SS}}{2} \quad (12)$$

Case 3. Input A is high and input B is low logic levels: Under this condition, the transistors Md1, Mc1, Md3 are ON and transistors Md2, Mc2, Md4 are OFF. The current through the ON transistors Md1, Mc1, Md3 is computed as:

$$I_{D1} = \frac{I_{SS}}{2}(1-p); I_{C1} = \frac{I_{SS}}{2} \cdot p; I_{D3} = \frac{I_{SS}}{2} \quad (13)$$

Consequently, the expression for the high output voltage V_{OH} is evaluated as:

$$V_{OH} = V_{DD} - \frac{R_p I_{SS}(1-p)}{2} \quad (14)$$

Case 4. Input A is low and Input B is high: Here, the transistors Md2, Mc2 and Md4 are ON and the transistors

Md1, Mc1, Md3 are OFF. The current through the ON transistors Md2, Mc2 and Md4 are expressed as:

$$I_{D2} = \frac{I_{SS}}{2}; I_{C2} = \frac{I_{SS}}{2} \cdot p; I_{D4} = \frac{I_{SS}}{2}(1-p) \quad (15)$$

Consequently, the expression for the high output voltage V_{OH} is evaluated as:

$$V_{OH} = V_{DD} - \frac{R_p I_{SS}(1-p)}{2} \quad (16)$$

Using the above Eqs. (9)–(16), the V_{SWING} is expressed as:

$$V_{SWING} = V_{OH} - V_{OL} = \frac{p R_p I_{SS}}{2} \quad (17)$$

The small signal voltage gain A_v and the noise margin NM, for the proposed fundamental cell based XOR2 gate, is calculated as per [6] and is given in (18)–(19).

$$A_v = \frac{g_{mn} R_p / 2}{1 - g_{mn} R_p / 2} \quad (18)$$

where g_{mn} is the transconductance of the modified triple tail cell. The expression for $g_{mn} R_p / 2$ is derived as in Appendix where its dependence on the dimensions of Md1–Md4 is shown.

$$NM = \frac{V_{SWING}}{2} \left(1 - \frac{1}{A_v} \right) \quad (19)$$

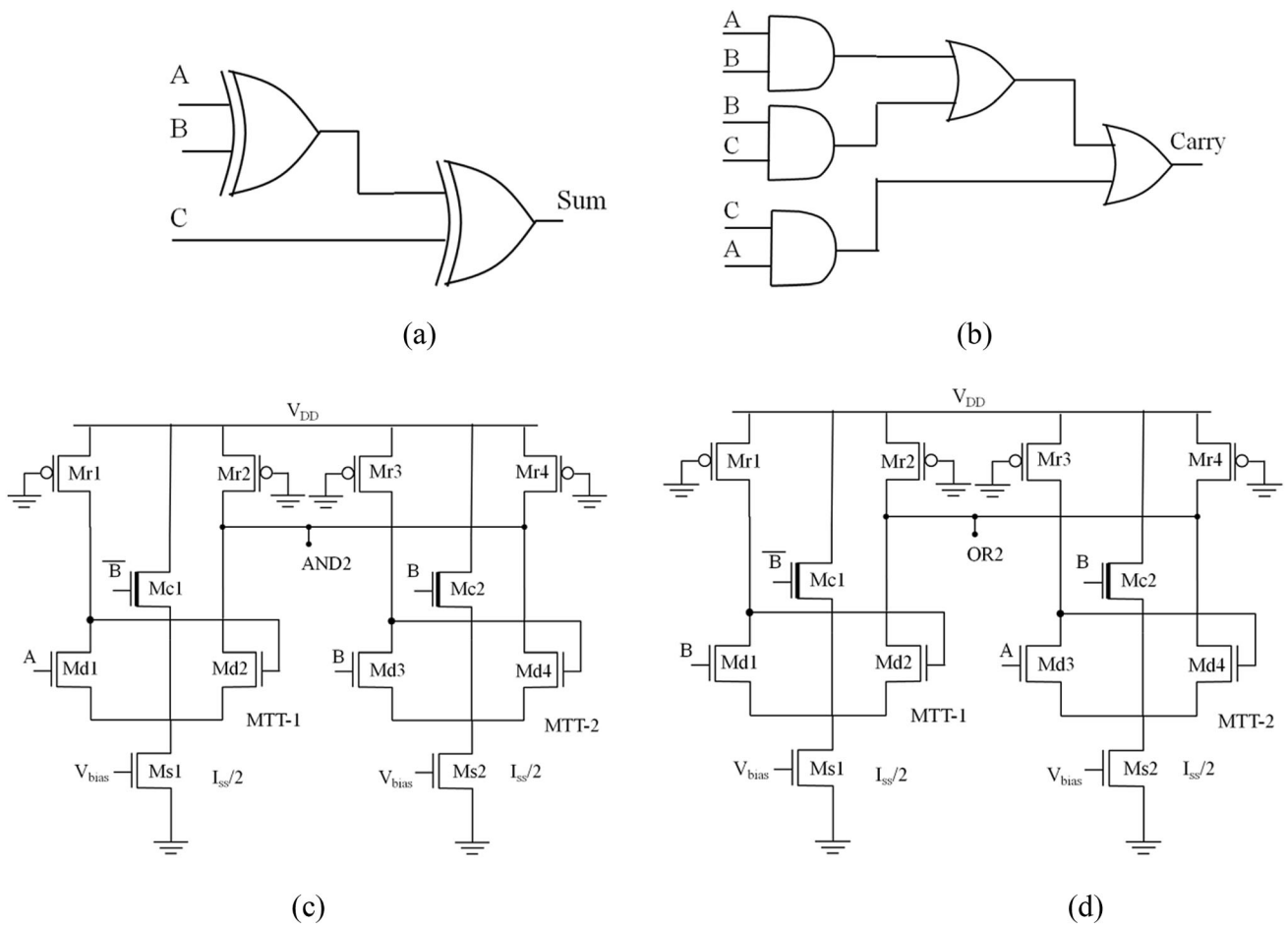


Fig. 10 Gate level schematic of **a** Sum and **b** Carry; Proposed fundamental cell based **c** AND2 gate and **d** OR2 gate

For a given V_{SWING} , A_v and NM, the Eqs. (17)–(19) are used in the design of the proposed fundamental cell and the design procedure for the cell is described in Appendix.

3.1 Validation of static model

The static model is verified by designing the proposed fundamental cell based XOR2 gate by considering design procedure outlined in Appendix for V_{DD} , A_v , α of 1.1 V, 19 and 1.3 respectively, wide range of I_{SS} (10 μ A to 200 μ A) and V_{SWING} (0.4 V and 0.5 V). The simulated results for variation of A_v , NM and V_{SWING} with respect to bias currents are recorded and are placed in Fig. 4a, b for V_{SWING} of 0.4 V and 0.5 V respectively. The predicted V_{SWING} , A_v and NM values using (17)–(19) are plotted along with corresponding simulated values in Fig. 4. In all plots of Fig. 4, percentage error in static parameters is also plotted and maximum error of 17.4% is observed.

3.2 Propagation delay modelling

The propagation delay depends on the contribution of parasitic MOS capacitances at the output node and the load capacitance. The parasitic capacitance for the proposed fundamental cell based XOR2 gate is calculated by considering input B as low such that MTT-2 is activated and MTT-1 is deactivated. Now, for a low-to-high transition on input A, total capacitance at the output node is depicted in Fig. 5 and can be expressed as:

$$C_{out} = C_{db1} + C_{gd1} + C_{dbr1} + C_{gdr1} + C_{gd2} + 0.5 \cdot C_{gs2} + C_{db4} + C_{gd4} + C_{dbr4} + C_{gdr4} + C_L \quad (20)$$

where C_{dbj} , C_{gdj} and C_{gsj} correspond to capacitance contributions of drain-bulk junction, gate-drain overlap and gate-source overlap capacitance of transistor Mdj. The capacitance $0.5 \cdot C_{gs2}$ is the Miller contribution of the gate-source capacitance Md2. The capacitances C_{dbrj} and C_{gdrj} represent contribution from drain-bulk junction and gate-

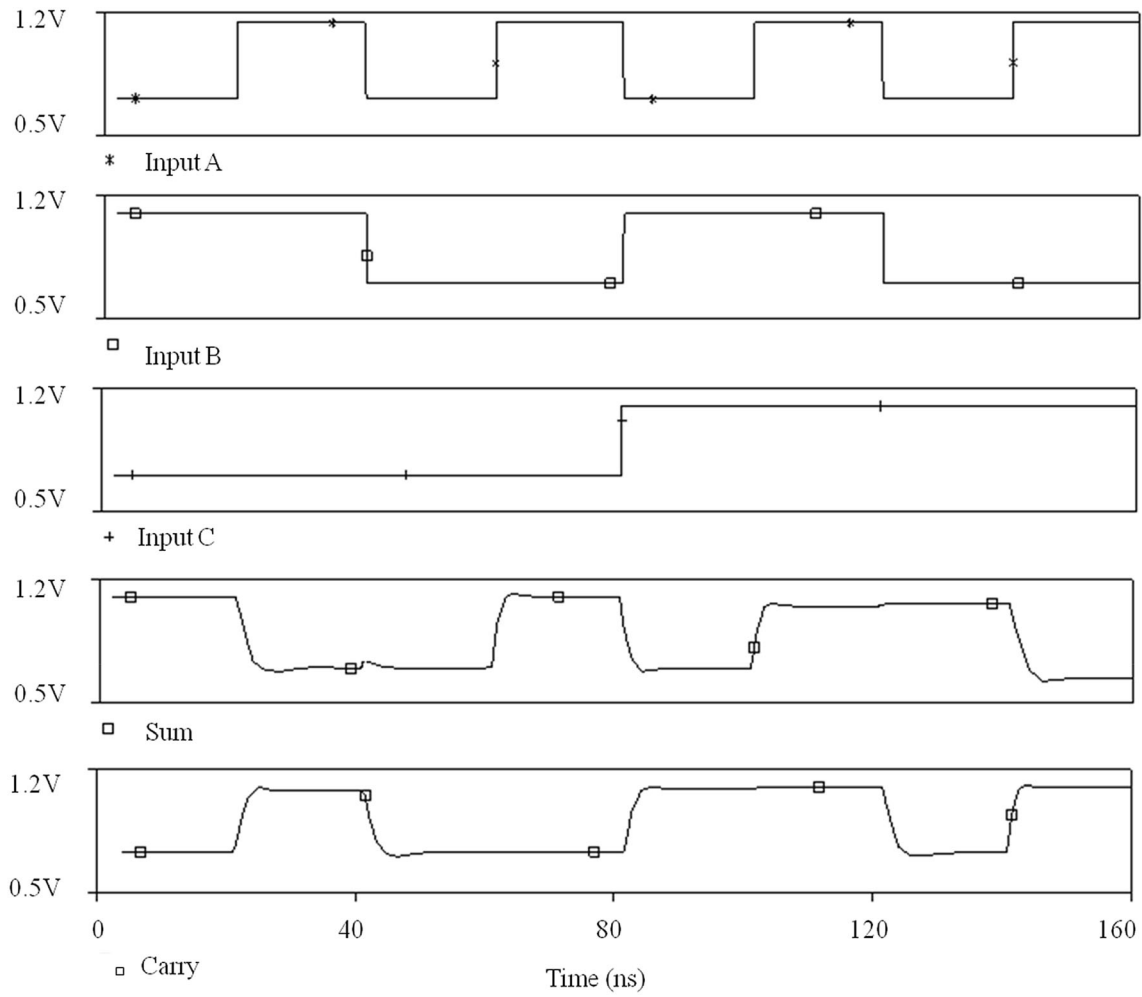


Fig. 11 Simulated input and output waveforms for the proposed fundamental cell based the full adder

Table 1 Performance summary of the full adder

Circuit	Scheme	Delay (ns)	Power (μ W)	PDP (fJ)	Area (μm^2)
Sum	Proposed	0.871	220	191.6	2.088
	Existing	0.870	220	191.4	6.148
Carry	Proposed	2.07	550	1138.5	5.22
	Existing	2.147	550	1180.85	15.37

drain overlap of PMOS load transistor M_{rj} , and C_L is the load capacitance.

The propagation delay τ_{PD} can now be expressed as:

$$\tau_{PD} = R_P C_{out} \quad (21)$$

$$= R_P (C_{db1} + C_{gd1} + C_{dbr1} + C_{gdr1} + C_{gd2} + 0.5 * C_{gs2} + C_{db4} + C_{gd4} + C_{dbr4} + C_{gdr4} + C_L) \quad (22)$$

Since the outer transistors (M_{d1} - M_{d4}) in the PDN are identical and the load transistors (M_{r1} - M_{r4}) are also identical, (22) can further be reduced as:

$$\tau_{PD} = R_P (2C_{db1} + 3C_{gd1} + 2C_{dbr1} + 2C_{gdr1} + 0.5 * C_{gs2} + C_L) \quad (23)$$

where the capacitances as per [14] can be expressed in terms of transistor dimensions as follows:

$$C_{db1} = W_N (L_{dn} C_j K_{eqn} + 2C_{jsw} K_{eqsw}) + 2L_{dn} C_{jsw} K_{eqsw} \quad (24)$$

$$C_{gd1} = C_{gd0} W_N \quad (25)$$

$$C_{dbr1} = W_P (L_{dp} C_j K_{eqp} + 2C_{jsw} K_{eqsw}) + 2L_{dp} C_{jsw} K_{eqsw} \quad (26)$$

$$C_{gdr1} = W_P \left(C_{gd0} + \frac{3}{4} A_{bulkmax} L_P C_{ox} \right) \quad (27)$$

$$C_{gs2} = \frac{2}{3} W_N L_N C_{ox} \quad (28)$$

where C_j is the junction capacitance, C_{jsw} is the sidewall capacitance, K_{eq} is the voltage equivalence factor for the substrate junction, K_{eqsw} is the voltage equivalence factor for the sidewall junction, L_d is the length of the source/drain junction, C_{gd0} is the gate drain capacitance per unit area, $A_{bulkmax}$ is the maximum value of the parameter A_{bulk} used to take into account the bulk charge effect and C_{ox} is the oxide capacitance. For the given design constraints, the appropriate values of the transistor dimensions can be evaluated by following the design procedure explained in Appendix.

Assuming I_{LOW} and I_{HIGH} as the minimum and the maximum bias current, which are set by the minimum dimensions of outer transistors (Md1-Md4) and the load transistors (Mr1-Mr4) respectively, the propagation delay for bias current I_{SS} in range of $I_{LOW} < I_{SS} < I_{HIGH}$ can be expressed as:

$$\tau_{PD} = R_P \left(A V_{SWING}^2 I_{SS} + B \frac{V_{SWING}}{I_{SS}} + (C + C_L) \right) \quad (29)$$

$$\text{where } A = \left(\frac{L_{Nmin}}{2\mu_n C_{ox} \left(\frac{g_{mn} R_P}{2} \right)^2 \left(V_{GS} - V_{TN} / \alpha \right)^4} \right) \\ (2L_{dn} C_j K_{eqn} + 4C_{jsw} K_{eqsw} + 3C_{gd0} + L_N C_{ox})$$

$$B = \frac{1}{p} (3A_{bulkmax} C_{ox} \mu_{effp} W_{Pmin} (V_{DD} - |V_{TP}|))$$

$$C = 4L_{dn} C_{jsw} K_{eqsw} + 4L_{dp} C_{jsw} K_{eqsw} \\ + 2W_{Pmin} (L_{dp} C_j K_{eqp} + 2C_{jswp} K_{eqswp}) \\ + 4L_{dp} C_{jswp} K_{eqswp}$$

$$+ 2W_{Pmin} C_{gd0p} \\ - \frac{3}{2} A_{bulkmax} C_{ox}^2 \mu_{effp} (V_{DD} - |V_{TP}|) R_{DSW} 10^{-6}$$

By using $R_P = \frac{2}{p} \frac{V_{SWING}}{I_{SS}}$, we get

$$\tau_{PD} = V_{SWING} \left(\frac{2}{p} (A) V_{SWING}^2 + (B) \frac{2V_{SWING}}{p I_{SS}^2} + \left(\frac{2}{p} \right) (C + C_L) \frac{1}{I_{SS}} \right) \quad (30)$$

This can be further rewritten as:

$$\tau_{PD} = V_{SWING} \left((a) V_{SWING}^2 + (b) \frac{V_{SWING}}{I_{SS}^2} + (c) \frac{1}{I_{SS}} \right) \quad (31)$$

where $a = \frac{2}{p} (A)$; $b = \frac{2}{p} (B)$ and $c = \frac{2}{p} (C + C_L)$.

3.3 Validation of delay expression

The derived delay expression in (31) is validated by designing and performing simulations for V_{DD} , A_v , α of 1.1 V, 19 and 1.3 respectively. The delay is measured for bias current I_{SS} ranging from 10 μ A to 200 μ A with V_{SWING} of 0.4 V and 0.5 V. The simulated delay and predicted delay values obtained by using the expression (31) for load capacitance value of 50fF, 500fF and 1 pF is plotted in Fig. 6 for V_{SWING} of 0.4 V and 0.5 V. It is observed that the propagation delay increases with increasing load capacitance. For a given load capacitance, the delay decreases with increasing I_{SS} , due to the availability of higher current for charging/discharging of load capacitance. Further, a maximum error of 27% can be observed from the error plot between the predicted and simulated values in Fig. 6.

3.4 Comparison between existing and proposed fundamental cells

In the previous subsections, the behaviour of the proposed fundamental cell as XOR2 gate is analysed. In this section, the performance of the proposed fundamental cell based XOR2 gate is compared with the existing fundamental cell based XOR2 gate. For this, an optimum value of α , the threshold voltage reduction factor, is determined. The proposed fundamental cell based XOR2 gate is simulated for α ranging from 1.1 to 1.9, for a bias current I_{SS} of 100 μ A. For the particular case when input A and input B are high and MTT-2 is deactivated, the ratio of currents through Mc2 and Md3 i.e. I_{C2}/I_{D3} was measured. The area and current ratio against α is plotted in Fig. 7. It is observed that the area reduces with the lowering threshold voltage of centre transistor. An optimum value of $\alpha = 1.7$ ($I_{C2}/I_{D3} = 6$) is chosen as it provides good activation/deactivation.

To compare the performance of the proposed fundamental cell with its existing counterpart, simulations are performed for measuring the propagation delay of XOR2 gate by keeping same current ratio ($I_{C2}/I_{D3} = 6$) in both the implementations. The delay of XOR2 is observed to be 405 ps in proposed cell based gate while it is 407 ps in existing counterpart. The corresponding area for the proposed fundamental cell based XOR2 gate is 1.048 μ m² while for the existing fundamental cell based XOR2 counterpart is 1.656 μ m². Thus, the proposed fundamental cell based XOR2 gate shows an area advantage of 36.7% with no negative impact on the propagation delay.

Further, the effect of parameter variations on voltage swing and delay of proposed and existing fundamental cell based XOR2 gate was studied by performing Monte Carlo analysis for 500 simulation runs. The corresponding variation in voltage swing and delay for both the gates are plotted in Fig. 8a, b, respectively. It is seen that voltage swing variations for the proposed and existing fundamental cell based XOR2 gate are 37.8% and 38.6% respectively and are of the same order. However, the delay shows lesser variation for the proposed fundamental cell based XOR2 (29.2%) as compared to the existing fundamental cell based XOR2 (34.7%).

The PVT analysis on delay for both proposed and existing fundamental cell based XOR2 gates is plotted in Fig. 9. In Fig. 9a, the delay is plotted for power supply voltage varied in the range of $\pm 10\%$ of nominal V_{DD} . $V_{DD} = 1.1$ V and different process corners at 27°C temperature. It can be observed that for a given process corner, the delay decreases as the V_{DD} increases. Also, for a given V_{DD} , the SS process corner gives the highest delay while the FF process corner leads to the lowest delay values. In Fig. 9b, the delay is plotted for $V_{DD} = 1.1$ V and different process corners at different temperatures. In terms of the temperature variation, the delay values increase with temperature across the all the corners.

4 Application

To showcase the advantage of the employing proposed fundamental cell in PFSCS circuit design, a full adder was designed and simulated using both the existing and proposed fundamental cell for $I_{SS} = 100 \mu\text{A}$ with $A_v = 19$ and $V_{SWING} = 0.4$ V, with the same $I_{Ci}/I_{Dj} = 6$. The gate level schematic for the sum and carry circuit is shown in Fig. 10a,b. The proposed fundamental cell based realization of the two input AND (AND2) and OR (OR2) are drawn in Fig. 10c, d, respectively. The simulation waveforms for the full adder using proposed fundamental cell is shown in Fig. 11 and the performance summary is tabulated in Table 1. It is seen that the proposed fundamental cell based design provides an area advantage of 66% while maintaining the same power and delay performance.

5 Conclusion

This paper presents a new fundamental cell that employs multiple threshold voltage transistors with the aim to reduce overall implementation area. The behaviour of the proposed fundamental cell is examined by configuring it as a two input exclusive XOR gate. Detailed analysis for static and delay models is put forward and a procedure to

design the cell for given constraint is derived. The proposed fundamental cell is designed for various design conditions namely bias currents and voltage swing and simulations are performed for validation using 180 nm CMOS technology parameters. The impact of process variations is also studied for the proposed cell. A full adder is designed as an application to compare the performance of the proposed and existing fundamental cells designs. It is found that full adder based on proposed scheme shows significant area saving (66%) while delay, power and PDP are within 4% of their corresponding values in existing counterpart.

Appendix: Design of proposed fundamental cell XOR2 gate

The design of the proposed fundamental cell XOR2 gate involves the method to determine the dimensions of various transistors for the given value of NM , A_v and I_{SS} . To begin, two bias currents namely, I_{HIGH} and I_{LOW} corresponding to bias current for minimum PMOS dimensions for a given V_{SWING} and the bias current corresponding to minimum NMOS dimensions are defined respectively.

For a given NM and A_v values and using the static model expressions, the required value of V_{SWING} , R_P is calculated as:

$$V_{SWING} = \frac{2NM}{1 - \frac{1}{A_v}} \quad (32)$$

$$R_P = \frac{2V_{SWING}}{p.I_{SS}} \quad (33)$$

Thus, the expression for I_{HIGH} can be written as:

$$I_{HIGH} = \frac{2V_{SWING}}{pR_{Pmin}} \quad (34)$$

where R_{Pmin} represents the resistance of minimum sized PMOS load transistor (Mr1-Mr4).

The calculated I_{HIGH} is compared with the required bias current I_{SS} value. For values of $I_{SS} > I_{HIGH}$, R_P will be less than R_{Pmin} and to calculate its value, L_P is set to minimum L_{Pmin} and W_P is calculated using (33) and [14].

$$W_P = \frac{pI_{SS}}{2V_{SWING}} \cdot \frac{L_{Pmin}}{\mu_{eff} C_{ox} (V_{DD} - |V_{TP}|) (1 - R_{DSW} 10^{-6} \mu_{eff} C_{ox} (V_{DD} - |V_{TP}|))} \quad (35)$$

Similarly, for values of $I_{SS} < I_{HIGH}$, R_P will be greater than R_{Pmin} and to calculate its value, W_P is set to W_{Pmin} and L_P is calculated as per following expression, derived using (33) and [14] and mathematical simplification.

$$L_P = \mu_{\text{effp}} C_{\text{ox}} W_{\text{Pmin}} (V_{\text{DD}} - |V_{\text{TP}}|) \left[\frac{2V_{\text{SWING}}}{PI_{\text{SS}}} - \frac{R_{\text{DSW}} 10^{-6}}{W_{\text{Pmin}}} \right] \quad (36)$$

After this, the dimensions of transistors in the PDN is derived by substituting

$$\frac{g_{\text{mn}} R_P}{2} = \sqrt{2\mu_n C_{\text{ox}} \frac{W_N}{L_N} \frac{1}{I_{\text{SS}}} \frac{V_{\text{SWING}}}{p}}$$

in the derived equation of A_v in Sect. 3 for $I_{\text{SS}} > I_{\text{Low}}$. The width W_N of the PDN transistors is calculated as:

$$W_N = p^2 \left(\frac{A_v}{1 - A_v} \right)^2 \frac{L_{\text{Nmin}} I_{\text{SS}}}{2\mu_n C_{\text{ox}} V_{\text{SWING}}^2} \quad (37)$$

where L_{Nmin} is the minimum length of the NMOS transistor, and all other variables are as previously defined. For the case having $I_{\text{SS}} < I_{\text{Low}}$, the W_N for all the PDN transistors is kept at their minimum value, W_{Nmin} .

References

- Kang, S.-M., Leblebici, Y., & Kim, C. (2014). *CMOS digital integrated circuits: Analysis and design*. (4th ed.). McGraw-Hill Higher Education.
- Kiaei, S., Chee, S. H., & Allstot, D. (1990). CMOS source-coupled logic for mixed-mode VLSI. In *Proceedings—IEEE international symposium on circuits and systems* (Vol. 2, pp. 1608–1611). IEEE.
- Allstot, D. J., Chee, S.-H., Kiaei, S., & Shrivastawa, M. (1993). Folded source-coupled logic vs. CMOS static logic for low-noise mixed-signal ICs. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 40(9), 553–563.
- Musicer, J. M., & Rabaey, J. (2000). MOS current mode logic for low power, low noise CORDIC computation in mixed-signal environments. In *Proceedings of the 2000 international symposium on Low power electronics and design—ISLPED'00* (pp. 102–107). New York, NY, USA: ACM Press.
- Alioto, M., & Palumbo, G. (2005). *Model and design of bipolar and MOS current-mode logic*. Berlin/Heidelberg: Springer.
- Alioto, M., Pancioni, L., Rocchi, S., & Vignoli, V. (2004). Modeling and evaluation of positive-feedback source-coupled logic. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 51(12), 2345–2355.
- Pandey, N., Gupta, K., & Gupta, M. (2014). An efficient triple-tail cell based PFSCl D latch. *Microelectronics Journal*, 45(8), 1001–1007.
- Pandey, N., Gupta, M., & Gupta, K. (2015). A PFSCl based configurable logic block. In *2015 Annual IEEE India conference (INDICON)* (pp. 1–4). IEEE.
- Gupta, K., Shukla, P., & Pandey, N. (2016). On the implementation of PFSCl adders. In *2016 Second international innovative applications of computational intelligence on power, energy and controls with their impact on humanity (CIPECH)* (pp. 287–291). IEEE.
- Gupta, K., Mittal, U., Baghla, R., Shukla, P., & Pandey, N. (2016). On the implementation of PFSCl serializer. In *2016 3rd international conference on signal processing and integrated networks (SPIN)* (pp. 436–440). IEEE.
- Gupta, K., Mittal, U., Baghla, R., & Pandey, N. (2016). Implementation of PFSCl demultiplexer. In *2016 international conference on computational techniques in information and communication technologies (ICCTICT)* (pp. 490–494). IEEE.
- Tyagi, A., Pandey, N., & Gupta, K. (2016). PFSCl based Linear Feedback Shift Register. In *2016 international conference on computational techniques in information and communication technologies (ICCTICT)* (pp. 580–585). IEEE.
- Agrawal, R. K., Pandey, N., & Gupta, K. (2017). Implementation of PFSCl razor flipflop. In *2017 International conference on computing methodologies and communication (ICCMC)* (pp. 6–11). IEEE.
- Cheng, Y., & Hu, C. (2002). *MOSFET modeling & BSIM3 user's guide*. Kluwer Academic Publishers.

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



Ranjana Sivaram received the B.Tech. degree in electronics and communication engineering from Ambedkar Institute of Technology, IP University, New Delhi, India in 2010 and the M.Tech. degree in VLSI & Embedded Systems from Delhi Technological University, New Delhi in 2012. From 2012 onwards, she is working in Department of Telecommunications, Government of India. She is currently pursuing part time Ph.D. in VLSI from Delhi

Technological University, New Delhi. Her current research interests include digital VLSI design.



Kirti Gupta received B.Tech. in Electronics and Communication Engineering from Indira Gandhi Institute of Technology, Delhi in 2002, M. Tech. in Information Technology from School of Information Technology in 2006. She received her Ph.D. in Electronics and Communication Engineering from Delhi Technological University, in 2016. Since 2002, she is with Bharati Vidyapeeth's College of Engineering, New Delhi and is presently serving as Professor in

the same institute. A life member of ISTE, and senior member of IEEE, she has published more than 100 research papers in international, national journals and conferences. Her teaching and research interest is in digital VLSI design.



Neeta Pandey received her M.E. in Microelectronics from Birla Institute of Technology and Sciences, Pilani in 1991 and Ph.D. from Guru Gobind Singh Indraprastha University, Delhi in 2009. She has served in Central Electronics Engineering Research Institute, Pilani, Indian Institute of Technology, Delhi, Priyadarshini College of Computer Science, Noida and Bharati Vidyapeeth's College of Engineering, Delhi in various capacities. At present, she is a

professor in the ECE department, Delhi Technological University.

Her teaching and research interests include analog and digital VLSI design. A life member of ISTE, and senior member of IEEE, USA, she has co-authored over 100 papers in international, national journals of repute and conferences.



Investigation of machining performance in die-sinking electrical discharge machining of pentagonal micro-cavities using cylindrical electrode

Shrikant Vidya^{1,3} · Reeta Wattal¹ · P Venkateswara Rao²

Received: 24 October 2020 / Accepted: 29 April 2021 / Published online: 8 May 2021
© The Brazilian Society of Mechanical Sciences and Engineering 2021

Abstract

The aim of the present article is to fabricate pentagonal micro-cavities and describe the influence of current on the machining performances. Machining of EN-24 alloy steel samples was performed in die-sinking electrical discharge machining (EDM) machine with varying values of current using polygon cycle approach. The machined pentagonal cavities were examined under optical microscope and scanning electron microscope (SEM) to evaluate machining performances in terms of corner error, white layer formation, surface crack distribution, and globule formation. It is found that as the value of current increases, there is more formation of white layer with non-uniform distribution of cracks and the thickness of white layer increased from 6.21 to 8.20 μm with increase in current. On the other hands, surface finish deteriorates when the current value rises. In addition to this, there is an enhancement in tool wear rate with increasing current. At the higher values of current, the spark energy increases which leads to greater melting and evaporation and production of smoke and bubbles on the dielectric surface. This study revealed that die-sinking EDM coupled with short electronic pulses and precise electrode movement is capable of producing microstructures under appropriate operating conditions.

Keywords EDM · micro-cavities · MRR · TWR · White layer · Surface characterization · Globules

1 Introduction

With a rapid growing demand for difficult-to-cut materials having high precision and tolerances in medical, electronic, aerospace, and advanced industrial applications, there becomes a great challenge for manufacturers to explore manufacturing protocols which can generate components at macro as well as micro-level with improved characteristics and reasonable cost. Such areas are very promising for the development of miniaturized devices and structures. EDM has proved to be the most successful and prominent

machining tool as compared to other advanced machining technologies such as ultrasonic machining (USM), electro-chemical machining (ECM), and laser machining due to its inherent behaviour of contactless machining irrespective of hardness. In EDM, the material gets washed away by the thermal energy of the quick and regular spark generated between the electrodes. EDM has been categorised into variants such as wire-cut, die-sinking, dry EDM, powder mixed, and micro-EDM which enable generation of components and devices on macro as well as micro-scale.

In most of the studies, researchers have discussed the influence of different parameters when machining different work materials, use of different tool materials, tool shape and size, and use of different dielectrics. It has been reported by several researchers that the rate of material removal and tool wear shows an enhancement with the rise in current value. Also, the distance between the globules increases, surface finish deteriorates as well as the average globule diameter and formation of globules, micro-cracks and voids increase with the increasing values of current [1, 2]. The influence of operating parameters on the performance measures during the machining of different work materials has

Technical Editor: Lincoln Cardoso Brandao.

✉ Shrikant Vidya
skvrsm@gmail.com; shrikant963_vidya@yahoo.in

¹ Department of Mechanical Engineering, Delhi Technological University, New Delhi, India

² Department of Mechanical Engineering, Indian Institute of Technology, New Delhi, India

³ School of Mechanical Engineering, Galgotias University, Uttar Pradesh, Greater Noida, India

also been investigated. In this context, the surface characteristics of machined Fe–Mn–Al alloy were analyzed by Guu et al. [3] by utilizing atomic force microscopy technology and reported that the pulse-on time contributes more in defining surface texture as compared to pulse current. On similar grounds, Si_3N_4 -TiN ceramic composite was machined using EDM by employing various pulse shapes. It was concluded that the material removal mechanisms and ceramic properties vary for different pulse shapes leading to a significant variation in surface texture [4]. Optimization of process parameters during machining of AlSiTi ceramic composite was performed by Patel et al. [5] to predict surface roughness model using response surface methodology and reported that surface finish gets decreased as the discharge current increases, and however, it is affected dominantly by pulse-on duration. Similarly, the grain size of materials also influences the surface morphology of machined surfaces. In this direction, EDM machining of ultra-fine-grained aluminium was performed by Mahdeih et al. [6] so as to examine process performance. It was concluded that lower value of machining parameters must be selected for better results of surface quality and also reported that the white layer thickness, density of cracks, and heat-affected zones are more prominent in ultra-fine-grained aluminium as compared to coarse-grain aluminium. Phan et al. [7] carried out multi-attribute optimization in vibration-aided EDM of high carbon silicon tool steel and reported that the quality factors got improved using low frequency vibration-added machining and Taguchi–TOPSIS computational approach. Similar works have been reported by researchers over vibration-assisted machining of silicon-based steel and utilization of multi-criteria decision making approaches like Taguchi-data envelopment analysis-based ranking and Taguchi–grey analysis [8, 9]. Another interesting work in this area has been carried out by several academic groups by employing different dielectric fluids. For instance, Amorim et al. [10] studied effects of varying sizes of molybdenum powder particles mixed with dielectric fluid during EDM machining of AISI H13 tool steel. It was reported that the surface properties of steel got modified due to the presence of Mo_2C , Fe–Mo, and Mo crystalline phases leading to increase in hardness. On same grounds, micro-powders were added to the oil as dielectric and their effects were examined for dimensional accuracy, wear on tool, rate of metal removal, and quality of surface. The addition of micro-powders improves the surface finish and stability of the profile machined [11]. Similarly, in the direction towards green electrical discharge machining, tap water was used as dielectric and optimization of parameters was done using the combination of Taguchi method and grey relational analysis by Tang et al. [12]. It was concluded that there is an improvement in MRR, decrease in the rate of electrode wear, and increase in surface finish. To improve the process

performance, several researchers have also worked on utilization of different designs, shapes as well as materials of tool electrodes. In this perspective, Khan et al. [13] examined the role of round, triangular, square, and diamond electrodes on the machining performance at varying values of discharge current. It was revealed that highest MRR and least wear and best surface finish are obtained by utilizing electrodes which were round. The square, triangular, and diamond-shaped electrodes followed the round ones. In the same direction, EDM machining of particle reinforced metal matrix composites using hexagonal electrode having through hole was performed by Lin et al. [14] and reported that the performance improves due to larger discharge debris gap. The machining performance is greatly affected by current and followed by flushing pressure, spindle rotation speed and duty cycle. Similarly, Nair et al. [15] machined Ti6Al4V using negative polarity and brass as electrode material and analysed the role of input factors on performance measures and reported that there is thickening effect in recast layer and an increment in material removal and roughness of surface while enhancing the values of current and time of discharge. On same ground, Khan et al. [16] reported that there is greater occurrence of wear at the tool electrode cross-section as compared to at the length side of electrode. They also concluded that the copper electrode wears lesser than the brass electrode due to higher thermal conductivity and electrodes wear less during machining of aluminium as compared to steel. Machining of Ti6Al4V using bundled electrode having multihole inner flushing technique and investigation of machining performance of die-sinking EDM were carried out by Gu et al. [17]. They reported that this technique enables rough machining at larger areas with lower wear of electrodes and an improved MRR. Also, Singh et al. [18] reported that wear rate of electrode and roughness of machined surface are lower in case of electrical discharge machining assisted with argon gas perforated tool as compared to air assisted and solid rotary tool. Yilmaz et al. [19] conducted machining of Inconel 718 and Ti–6Al–4 V employing brass and copper single-channel as well as multi-channel tool electrodes. They reported that the use of single-channel electrodes produces higher MRR as compared to multi-channel electrodes. There is lower occurrence of wear in copper electrodes than brass electrodes in case of both single as well as multi-channel electrodes. They also concluded that the surface quality becomes better in case of machining through multi-channel electrodes, while the hardness is lower. Towards the approach of machining polygonal shapes, Reuleaux triangle tool path strategy was adopted to fabricate polygons by Ziada et al. [20]. Rotating curvilinear tool electrodes were utilized to machine polygons in die-sinking EDM. It was concluded that there is improvement in the flushing between the discharge gap by the combined effect of tool rotation and translation which leads to

maximum utilization of the front side of the tool electrode. In order to exploit the micro-level variant of EDM named as micro-EDM, several researchers utilized micro-EDM as well as micro-EDM milling to fabricate micro-holes, channels as well as cavities of different shapes to evaluate the physical behaviour involved in the machining and the machining performance. For instance, mist deionized water jet was utilized by Li et al. [21] in micro-EDM drilling which resulted in the machining of deep micro-holes with high accuracy and machining speed as well as improvement in the exhaust of debris. On the same background, Singh et al. [22] employed process parameters as capacitance, voltage, and feed rate to investigate the machining performance as MRR, overcut, and machining time. It was concluded that straight micro-holes having fine edges can be efficiently achieved by micro-EDM. In case of overcut and MRR, capacitance plays a major role, while feed rate becomes a vital factor in determination of machining time. Karthikeyan et al. [23] investigated the shape, form, and surface quality of the channels produced by micro-EDM milling. It was reported that the amount of redeposition is influenced by the rotational motion of tool and there is no redeposition on the tool surface due to the centrifugal force. Vidya et al. [24] fabricated micro-holes, channels, cross-channels, triangular, and square geometries in micro-EDM milling and analyzed the machining performance in terms of shape error, surface integrity, and characterization. It was concluded that circular holes are the best in terms of dimensional accuracy, while triangular cavities have the best surface finish. It was also reported that there is significant effect of tool rotation on the surface characteristics such as globule formation, recast layer formation, and the flow of eroded particles. While no studies have reported the die-sinking EDM machining of pentagonal shapes using cylindrical tool electrode utilizing polygon cycle strategy and influence of parameters on the machining performance during machining of pentagonal geometries.

In this paper, authors investigate the process performance of die-sinking EDM machining of pentagonal micro-cavities by employing polygon cycle approach. All these geometries are machined using cylindrical electrode at varying current values, and effects of current are presented based on the analysis of corner error, material removal rate, tool wear, white layer formation, globule formation, surface crack distribution, and surface roughness.

2 Experimental details

The experiments were performed over a XPERT-1 Electrical Discharge Machine of Electronica Machine Tools which contains e-pulse 50 CNC power supply unit enabling the fabrication of multifaceted tool path trajectories as well as micromachining. 400 μm cylindrical copper

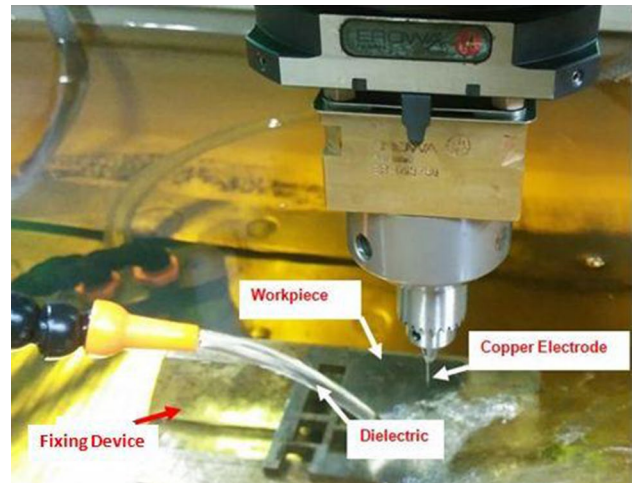


Fig. 1 EDM machining setup

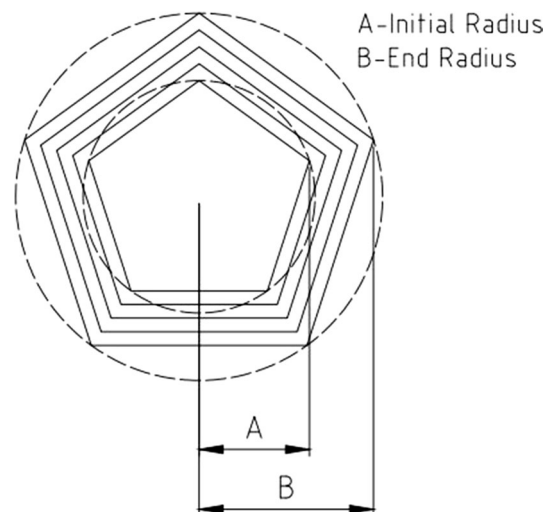


Fig. 2 Tool path strategy adopted

electrode of length 50 mm was selected as tool material, and EN-24 alloy steel of $100 \times 100 \times 15$ mm was chosen as work material. The weight percentage composition of EN-24 alloy steel is: carbon 0.402, silicon 0.340, manganese 0.770, nickel 1.551, chromium 0.900, molybdenum 0.276 and iron 95.61 [25]. Figure 1 portrays the setup utilized for machining. EDM oil – IPOL SEO 450 has been used as a dielectric which is continuously circulated and under which both tool and workpiece are submerged. Similar to conventional machining process, the cylindrical tool electrode is mounted onto the spindle that travels along Z-axis constantly to impart the machining depth and the workpiece is mounted over the X–Y positioning table. The predefined polygon cycle tool path strategy (Fig. 2) ensured by CNC controller is provided to the tool to machine pentagonal cavities of depth of 2000 micron

and keeping tool electrode diameter as reference dimension. The machining was carried out at the varying current values, while other parameters were kept constant as listed in Table 1. There were three experiments performed at each value of current, and the average value of machining performance was recorded for analysis.

The pentagonal cavities machined and their geometries were examined under an optical microscope. These cavities were analysed under SEM and characterized critically to examine the white layer formation, surface crack distribution, and globule formation. Cavities were also examined to ensure profile accuracy in terms of surface roughness measurement.

3 Results and discussion

Figure 3 shows the microscopic images of pentagonal geometries machined at varying current levels. These samples were critically examined to assess the machining performance of EDM on different parameters such as error in shape, formation of recast layer, distribution of cracks, formation of globules, and surface roughness.

Table 1 Parameters selected for machining

Parameters	Values
Current	1 A, 1.25 A, 1.50 A, 1.75 A and 2 A
Pulse-on time	5 μ s
Pulse-off time	16 μ s
Gap voltage	75 V
Feed rate	27 mm / min
Polarity	Tool (+ve)

3.1 Geometry errors

As mentioned in Sect. 2, the tool electrode is fed in a predefined strategy in which it travels parallel with respect to the inner periphery to machine the entire polygonal cavity in a number of steps. Hence, the location and direction of discharge vary according to the geometry to be machined. During the machining of cavities, the secondary discharge also comes into effect in addition to the primary discharge which acts as a key player in material removal, electrode wear, spark jumping, short circuiting, and unstable spark gap. Due to these effects, the pentagonal cavities produced by cylindrical electrode in EDM do not have sharp corners, i.e. the corners get rounded which creates the need of secondary machining and finishing operations. Apart from this, the cavities contain the presence of burrs and distortions at the edges as well as recast layer formation also takes place. As the value of current increases, the discharge energy increases considerably in micromachining through EDM leading to greater chances of occurrence of melting as compared to vapourization of the workpiece as well as tool material. Due to this, the geometry starts losing its shape, and chances of distortion and overcut increase. From Fig. 3, it is evident that the pentagonal cavity machined at 1 A is the most accurate in shape and size while that machined at 2 A is the least accurate in shape and has more overcut.

3.2 Current vs. MRR and TWR

Equations 1 and 2 were employed to calculate the values of MRR and TWR, respectively.

$$\text{MRR} = \frac{\text{Average depth} \times \text{Average width} \times \text{length}}{\text{Time of machining}} \quad (1)$$

$$\text{TWR} = \frac{\pi \times (\text{tool electrode diameter})^2 \times \text{length of eroded tool}}{4 \times \text{Time of machining}} \quad (2)$$

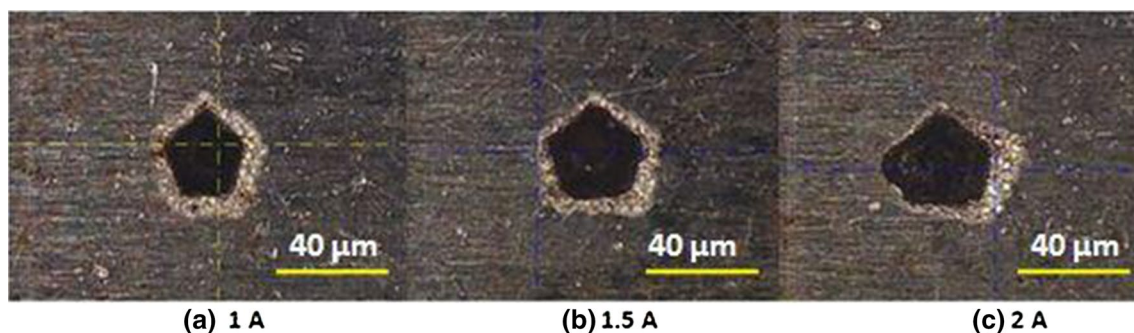


Fig. 3 Shape errors representing corner effect, the presence of burrs, overcut, and deviation in shape

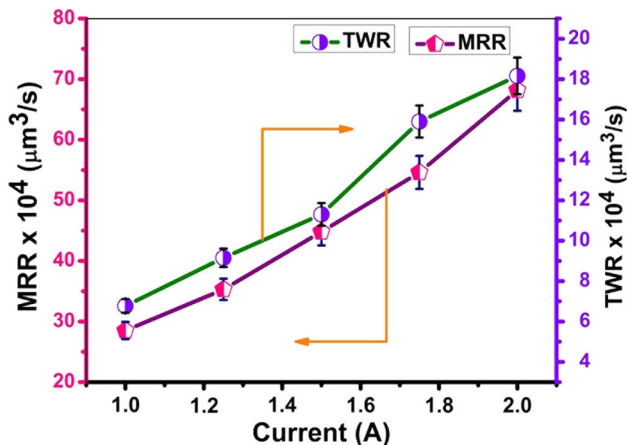


Fig. 4 MRR and TWR as a function of Current

Table 2 MRR and TWR as a function of current

Sl. No	Current (A)	MRR $\times 10^4$ ($\mu\text{m}^3/\text{s}$)	TWR $\times 10^4$ ($\mu\text{m}^3/\text{s}$)
1	1	28.50	6.77
2	1.25	35.30	9.15
3	1.5	44.80	11.30
4	1.75	54.60	15.90
5	2	68.17	18.16

Figure 4 portrays the changes in MRR and TWR with increasing value of current. It is clear from the figure that the value of both MRR and TWR shows an enhancement with the rise in current and it follows a linear trend which suggests that MRR is a function of current where current is variable and on time, off time and gap voltage are kept constant. It happens because of the fact that there is the presence of more thermal energy in the spark at higher values of current leads to more removal of materials from tool as well as workpiece. On the basis of experimental results, it is confirmed that the rate of tool wear and material removal is lower at lower values of current. This is due to the occurrence of less vaporization and melting between the electrodes. It is also clear from Table 2 that MRR values are significantly higher than TWR values which show a good and

capable machining performance of EDM in machining of polygonal cavities. Similar kind of work has been reported by many researchers over various alloys [1, 26 and 27].

3.3 Current vs. surface roughness

In the present article, the influence of current on the surface finish of the pentagonal cavities produced has also been a focus. In this direction, Table 3 presents the values of surface roughness of machined cavities obtained through Zeiss surface roughness measuring machine, SURFCOM FLEX 50A. Figure 5 shows the graphical representation as to how the surface roughness changes with current. It is clear from the figure that with the increment in current, there is more energy in the spark leading to more removal of materials and rougher surfaces. So, with the increment in current, the surface finish of the machined cavities deteriorates. It can be attributed to the fact that there is higher impingement of sparks at the higher values of current leading to greater peak height. Based on the observations from SEM micrographs at higher magnifications, these variations in surface finish and profile errors can be attributed to secondary discharges, inappropriate removal of debris, flushing environment, and recast layer [24]. In the region, when the current increases from 1.25 to 1.5 A, we observed a steeper roughness curve due to rapid increase in surface damage and micro-cracks. However, beyond 1.5 A, there is a less steeper slope due to the fact there is a formation of patterned textured surface over the surface of the material because of the spreading out of redeposited materials [28, 29].

3.4 Surface characterization

The pentagonal cavities machined at different values of current were examined under optical microscope and scanning electron microscope. It is evident from Fig. 3 that cavities machined at all values of current have burrs at their cavity edges which lead to the requirements of secondary operations.

Due to intense heat, local melting, evaporation, and improper removal of molten particles in EDM, white layer (Recast layer) formation takes place as these left-over molten materials cool and solidify. During the course of cooling

Table 3 Measured values of surface roughness

Sl. No	Current (A)	Average surface roughness (μm)				Standard deviation
		Exp. 1	Exp. 2	Exp. 3	Mean	
1	1	3.323	3.654	3.256	3.411	0.17
2	1.25	3.436	3.627	4.223	3.762	0.34
3	1.5	4.423	5.138	4.980	4.847	0.31
4	1.75	4.736	5.724	4.789	5.083	0.45
5	2	5.162	4.842	5.866	5.29	0.43

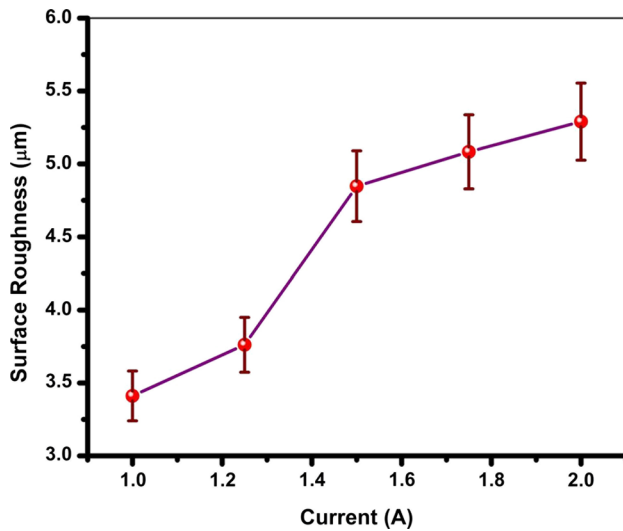


Fig. 5 Current vs. surface roughness

of molten particles, gases evolve which creates pockmarks, voids, and globule formation on the surface of machined cavities [30, 31].

The thickness of formed recast layer was measured at four different locations as portrayed in Fig. 6 with SEM micrographs, and average value was calculated for each value of current. In this direction, Fig. 7 shows that the thickness of the white layer had an increasing trend with an increase in the current. The average thickness of the recast layer at 1 A, 1.5 A, and 2 A was found to be 6.21 μm, 7.46 μm, and 8.20 μm, respectively. The magnitude of the average thickness of white layer is dependent on the pulse energy which is a function of pulse current and pulse duration. Since, the pulse duration has been kept constant throughout the study, the white layer thickness is a function of current which serves as the pivotal factor for pulse energy [32]. Due to the

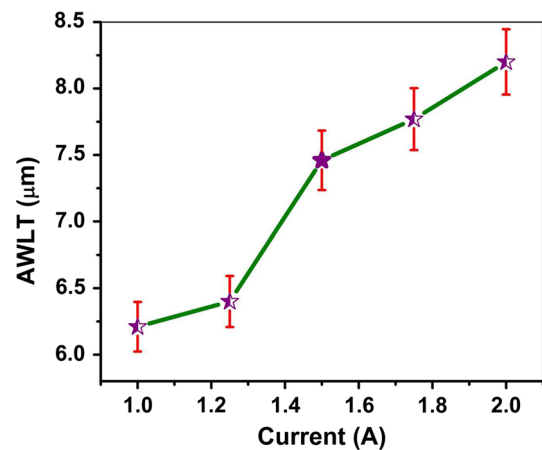


Fig. 7 Current vs. average white layer thickness (AWLT)

recast layer, residual stresses buildup and formation of voids and micro-cracks take place. Hence, formation of micro-cracks and voids cannot be avoided in EDM.

In addition to the formation of recast layer and micro-cracks, there is the presence of globules over the machined surface which plays a major role in determining the surface finish. From SEM micrographs, it is clearly visible that there is even distribution of globules at lower values of current, while there is uneven dispersion of globules along with irregular shapes at higher values of current as portrayed in Fig. 8. On the surface of cavities machined at 1 A, there is uniform distribution of micro-cracks and voids, less formation of recast layer, and regular formation of small globules. In case of cavities machined at 1.5 A, recast layer formation along with micro-cracks and voids is more, globules are bigger as compared to that machined at 1 A and the continuous surface crack distributed all over the surface is clearly visible and seems well adhered to the parent metal. While at the current value of 2 A, the

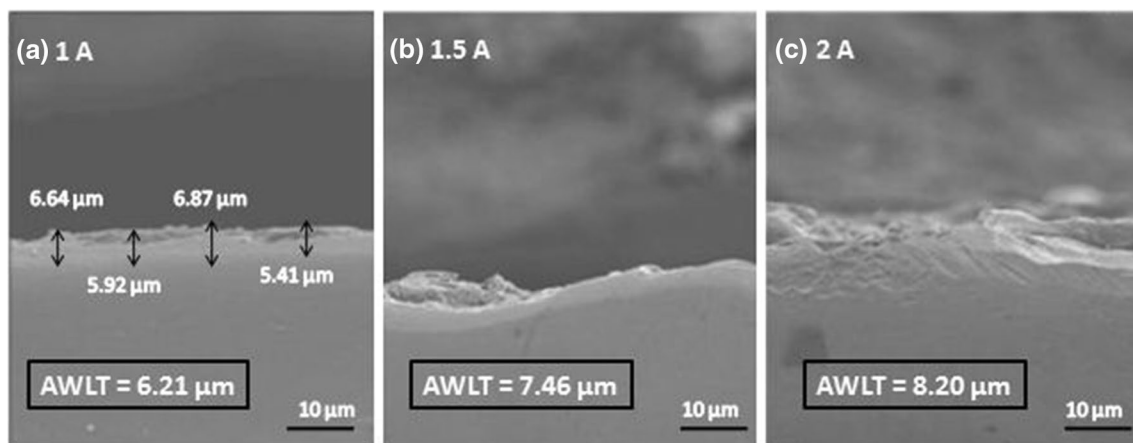


Fig. 6 Average white layer thickness at (a) 1 A (b) 1.5 A and (c) 2 A

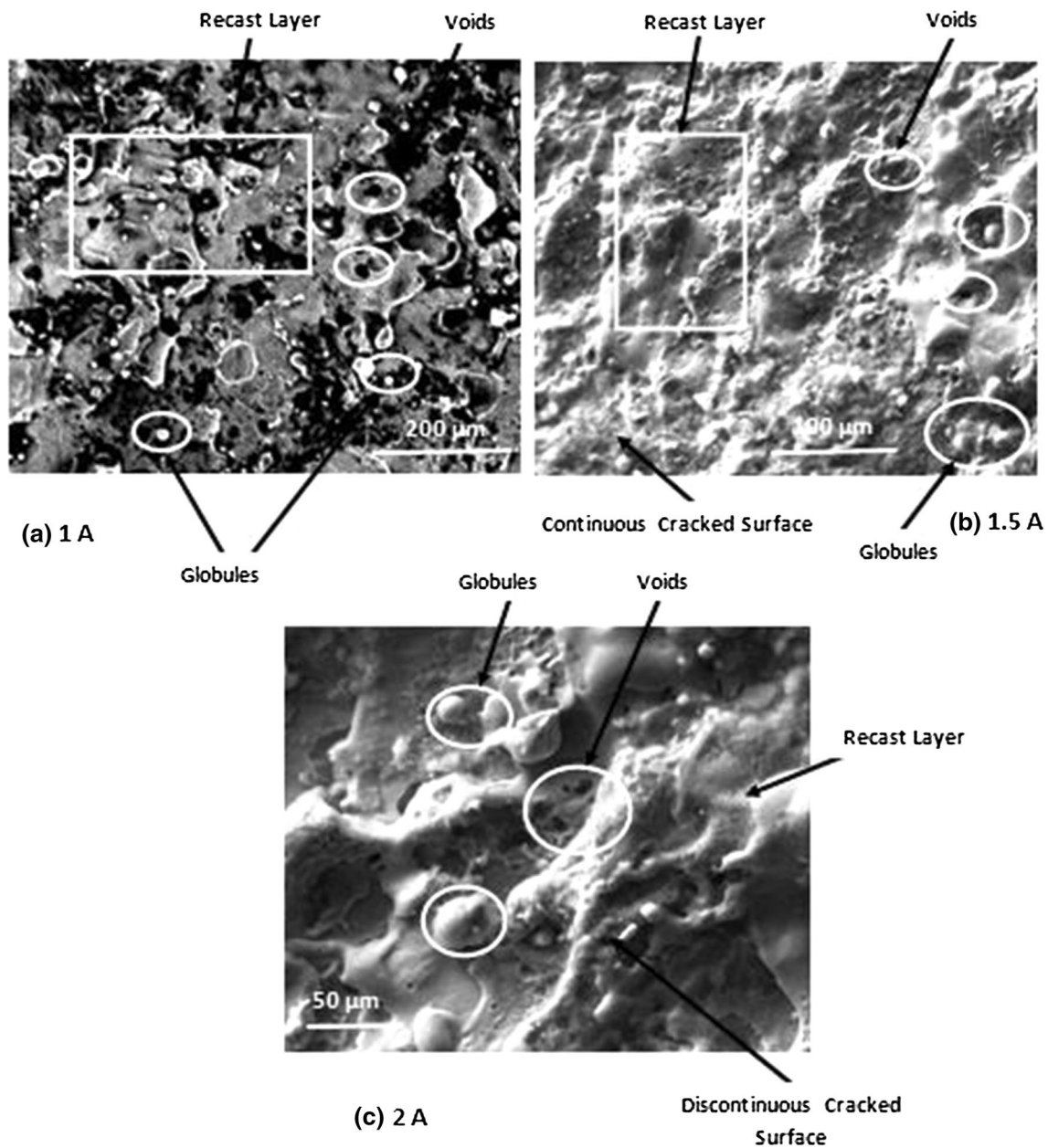


Fig. 8 Surface morphology at (a) 1 A (b) 1.5 A and (c) 2 A current

globules are bigger and distorted in shape, more recast layer, and discontinuous distribution of cracks and voids over the surface.

As the value of current increases, the size of plasma channel increases, and the magnitude of surface tension increases between the liquid and solid metal part. Due to this, more resistive forces come into picture between the two phases, and the formation of globules and recast layer occurs at the greater extent. In a nutshell, it can be attributed that the magnitude of formation of globules and AWLT is a function of the surface tension [32]

4 Conclusions

The fabrication of pentagonal micro-cavities over EN-24 alloy steel by utilizing polygon cycle approach in die-sinking EDM machining was examined for its performance. The influence of current was analyzed on the machining performance in terms of shape error, surface characteristics, MRR, and TWR. It is found that the removal rate of material and wear rate of tool increased when the machining current values were gradually increased. The

calculated values of MRR increased from 28.50 (at 1 A) to $68.17 \times 10^4 \mu\text{m}^3/\text{s}$ (at 2 A), while the TWR increased from 6.77 to $18.16 \times 10^4 \mu\text{m}^3/\text{s}$, respectively. The extent of formation of white layer seems to increase when the current was gradually increased. Thorough examination of machined surface characteristics depicted the existence of globules, cracks, and voids which generally increases, while there is a progressive growth in the current. When micromachined at a value of 2 A, discontinuous cracks and surface damage were also noted which can be attributed to the fact that at higher current, there is a significant rise in the spark energy which ultimately leads to greater melting and evaporation and production of smoke and bubbles on the dielectric surface. As a consequence, cavitation and sudden cooling may occur at the surface which could possibly build up the formation of thicker recast layer and rougher surfaces which results in building up of residual stresses and micro-cracks.

In an attempt to present promising research, authors intend to investigate further to establish process ability to fabricate micro-geometries of different shapes on different work materials in order to be able to facilitate in the development of micro- and nano-scale functional devices, which are of greater interest in the modern era of miniaturization.

References

- Arooj S, Shah M, Sadiq S, Jaffery SHI, Khushnood S (2014) Effect of current in the EDM Machining Of Aluminum 6061 T6 and its effect on the surface morphology. *Arab J Sci Eng* 39(5):4187–4199. <https://doi.org/10.1007/s13369-014-1020-z>
- Nikalje AM, Kumar A, Srinadh KVS (2013) Influence of parameters and optimization of EDM performance measures on MDN 300 steel using Taguchi method. *Int J Adv Manuf Technol* 69:41–49. <https://doi.org/10.1007/s00170-013-5008-8>
- Guu YH, Hou M, Ti K (2007) Effect of machining parameters on surface textures in EDM of Fe–Mn–Al alloy. *Mater Sci Eng A* 466, 61–67. DOI: <https://doi.org/10.1016/j.msea.2007.02.035>
- Liu K, Reynaerts D, Lauwers B (2009) Influence of the pulse shape on the EDM performance of Si_3N_4 -TiN ceramic composite. *CIRP Ann-Manuf Tech* 58, 217–220. DOI: <https://doi.org/10.1016/j.cirp.2009.03.002>
- Patel KM, Pandey PM, Rao PV (2009) Determination of an optimum parametric combination using surface roughness prediction model for EDM of $\text{Al}_2\text{O}_3/\text{SiC}_w/\text{TiC}$ ceramic composite. *Mater Manuf Processes* 24:675–682. <https://doi.org/10.1080/10426910902769319>
- Mahdiah MS, Mahdavinjad R (2016) Recast layer and micro-cracks in electrical discharge machining of ultra-fine-grained aluminum. *Proc IMechE Part B: J Eng Manufact* 1–10, DOI: <https://doi.org/10.1177/0954405416641326>
- Phan NH, Muthuramalingam T (2020) Multi-criteria decision-making of vibration-aided machining for high silicon-carbon tool steel with Taguchi-topsis approach. *SILICON*. <https://doi.org/10.1007/s12633-020-00632-w>
- Phan NH, Muthuramalingam T (2020) Multi criteria decision making of vibration assisted EDM process parameters on machining silicon steel using taguchi-DEAR methodology. *SILICON*. <https://doi.org/10.1007/s12633-020-00573-4>
- Phan NH, Banh TL, Mashood KA, Tran DQ, Pham VD, Muthuramalingam T, Nguyen VD, Nguyen DT (2020) Application of TGRA-based optimisation for machinability of high-chromium tool steel in the EDM process. *Arab J Sci Eng* 45:5555–5562. <https://doi.org/10.1007/s13369-020-04456-z>
- Amorim FL, Dalcin VA, Soares P, Mendes LA (2017) Surface modification of tool steel by electrical discharge machining with molybdenum powder mixed in dielectric fluid. *Int J Adv Manuf Technol* 91:341–350. <https://doi.org/10.1007/s00170-016-9678-x>
- Sahu DR, Mandal A (2020) Critical analysis of surface integrity parameters and dimensional accuracy in powder-mixed EDM. *Mater Manuf Processes* 35:430–441. <https://doi.org/10.1080/10426914.2020.1718695>
- Tang L, Du YT (2014) Experimental study on Green Electrical Discharge Machining in Tap Water of Ti–6Al–4 V and Parameters Optimization. *Int J Adv Manuf Technol* 70:469–475. <https://doi.org/10.1007/s00170-013-5274-5>
- Khan AA, Ali MY, Haque MM (2009) A study of electrode shape configuration on the performance of die sinking EDM. *Int J Mech Mater Eng* 4:19–23
- Lin Z, Guo Z, Jiang S, Liu G, Liu J (2018) Electrical discharge drilling of metal matrix composites with a hollow hexagonal electrode. *Adv Compos Lett* 27(5):193–203. <https://doi.org/10.1177/096369351802700503>
- Nair S, Dutta A, Narayanan R, Giridharan A (2019) Investigation on EDM machining of Ti6Al4V with negative polarity brass electrode. *Mater Manuf Processes* 34:1824–1831. <https://doi.org/10.1080/10426914.2019.1675891>
- Khan AA (2008) Electrode wear and material removal rate during EDM of aluminium and mild steel using copper and brass electrodes. *Int J Adv Manuf Technol* 39:482–487. <https://doi.org/10.1007/s00170-007-1241-3>
- Gu L, Li L, Zhao W, Rajurkar KP (2012) Electrical discharge machining of Ti6Al4V with a bundled electrode. *Int J Mach Tools Manuf* 53:100–106. <https://doi.org/10.1016/j.ijmachtools.2011.10.002>
- Singh NK, Pandey PM, Singh KK (2017) experimental investigations into the performance of EDM using argon gas assisted perforated electrodes. *Mater Manuf Processes* 32:940–951. <https://doi.org/10.1080/10426914.2016.1221079>
- Yilmaz O, Okka MA (2010) Effect of single and multi-channel electrodes application on EDM fast hole drilling performance. *Int J Adv Manuf Technol* 51:185–194. <https://doi.org/10.1007/s00170-010-2625-3>
- Ziada Y, Koshy P (2007) Rotating curvilinear tools for EDM of polygonal shapes with sharp corners. *Annals of the CIRP* 56:221–224. <https://doi.org/10.1016/j.cirp.2007.05.052>
- Guodong L, Wataru N (2020) Realization of micro EDM drilling with high machining speed and accuracy by using mist deionized water jet. *Precision Eng* 61, 136–146. ISSN 0141–6359, <https://doi.org/10.1016/j.precisioneng.2019.09.016>
- Singh AK, Patowari PK, Chandrasekaran M (2020) Experimental study on drilling micro-hole through micro-EDM and optimization of multiple performance characteristics. *J Braz Soc Mech Sci Eng* 42:506. <https://doi.org/10.1007/s40430-020-02595-w>
- Karthikeyan G, Garg AK, Ramkumar J, Dhamodaran S (2012) A Microscopic investigation of machining behaviour in I-ED-milling process. *J Manuf Processes* 14(3):297–306
- Vidya S, Vijay V, Barman S, Chebolu A, Nagahanumaiah N (2015) Effects of different cavity geometries on machining performance in micro-electrical discharge milling. *J Micro Nano Manufact* 3(1)
- Das A, Tirkey N, Patel SK et al (2019) A comparison of machinability in hard turning of EN-24 alloy steel under mist

- cooled and dry cutting environments with a coated cermet tool. *J Fail Anal and Preven* 19:115–130. <https://doi.org/10.1007/s11668-018-0574-6>
26. Baldin V, Baldin CRB, Machado AR et al (2020) Machining of Inconel 718 with a defined geometry tool or by electrical discharge machining. *J Braz Soc Mech Sci Eng* 42:265. <https://doi.org/10.1007/s40430-020-02358-7>
27. Kuo CG, Hsu CY, Chen JH, Lee PW (2017) Discharge current effect on machining characteristics and mechanical properties of aluminum alloy 6061 workpiece produced by electric discharging machining process. *Adv Mech Eng* 9(11):1687814017730756
28. Belloufi A, Mezoudj M, Abdelkrim M et al (2020) Experimental and predictive study by multi-output fuzzy model of electrical discharge machining performances. *Int J Adv Manuf Technol* 109:2065–2093. <https://doi.org/10.1007/s00170-020-05718-8>
29. Straplets JL, Dimievi WL (2007) Textured surface for test sample card and mold for manufacturing the same, European Patent No. A2
30. Ekmekci B (2007) Residual stresses and white layer in electric dischargemachining (EDM). *Appl Surf Sci* 253, 9234–9240
31. Upadhyay C, Datta S, Masanta M et al (2017) An experimental investigation emphasizing surface characteristics of electro-discharge-machined Inconel 601. *J Braz Soc Mech Sci Eng* 39:3051–3066. <https://doi.org/10.1007/s40430-016-0643-2>
32. Ramasawmy H, Blunt L, Rajurkar KP (2005) Investigation of the relationship between the white layer thickness and 3D surface texture parameters in the die sinking. *EDM Process Preci Eng* 29(4), 479–490

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

IP Traffic Classification of 4G Network using Machine Learning Techniques

Rahul

Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
rahul@dtu.ac.in

Anupam Raj

Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
anupamraj1312@gmail.com

Amit Gupta

Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
amitgupta7.official@gmail.com

Mayank Arora

Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
mayankarora66@yahoo.com

Abstract—In today's world, the number of internet services and users is increasing rapidly. This leads to a significant rise in the internet traffic. Thus, the task of classifying IP traffic is essential for internet service providers or ISP, as well as various government and private organizations in order to have better network management and security. IP traffic classification involves identification of user activity using network traffic flowing through the system. This will also help in enhancing the performance of the network. The use of traditional IP traffic classification mechanisms which are based on inspection of packet payload and port numbers has decreased drastically because there are many internet applications nowadays which use port numbers which are dynamic in nature rather than well-known port numbers. Also, there are several encryption techniques nowadays due to which the inspection of packet payload is hindered. Presently, various machine learning techniques are generally used for classifying IP traffic. However, not much research has been conducted for the classification of IP traffic for a 4G network. During this research, we developed a new dataset by capturing packets of real-time internet traffic data of a 4G network using a tool named Wireshark. After that, we extracted the inferred features of the captured packets by using a python script. Then we applied five machine learning models, i.e., Decision Tree, Support Vector Machines, K Nearest Neighbours, Random Forest, and Naive Bayes for classifying IP traffic. It was observed that Random Forest gave the best accuracy of approximately 87%.

Keywords—IP Traffic Classification; Port Number; Deep Packet Inspection; Packet Capturing; Feature Extraction; Machine Learning

I. INTRODUCTION

The significant rise in the number of internet users around the world due to lower internet prices and easier access to mobile phones and other devices and services has led to an exponential increase in the amount of IP traffic that is being transmitted globally. This increase in IP traffic can be

attributed to the usage of various applications by internet users in their everyday lives like Email, World Wide Web, text messaging, audio or video calls, and various other internet applications. Thus, classifying IP traffic is very essential for the internet service providers (ISPs) as well as various government and private organizations. IP traffic classification can assist in several network management activities like analyzing the Quality of Service (QoS) for internet service, diagnosis of any fault in the network, etc. It can also help in several network security activities like intrusion detection [1].

There are several techniques that have been proposed for achieving the task of IP traffic classification [2]–[4]. Traditional IP traffic classification mechanisms are based on port number and inspection of the packet payload [2]. However, the use of these techniques has reduced drastically nowadays. In port number-based technique, the task of IP traffic classification is achieved using well known port numbers. The reason for the ineffectiveness of port number-based technique is that nowadays, there are many internet applications (like P2P) which use port numbers which are dynamic in nature rather than well-known port numbers. In payload-based technique, the packet payload contents are analyzed for the task of IP traffic classification. The reason for the ineffectiveness of payload-based technique is the use of various encryption techniques nowadays for encrypting the packet payload, due to which the direct inspection of the packet payload is hindered.

Presently, several machine learning techniques are generally used for classifying IP traffic [4]–[17]. However, not much research work has been conducted for the classification of IP traffic for a 4G network. Thus, in this research work, we have employed machine learning based approach to classify IP traffic for a 4G network. In machine learning based approach, various statistical features, which are independent of the packet payload, are utilized in training several machine learning

models and then the task of classification of IP traffic is achieved by using these trained machine learning models.

During our research, we developed a new dataset by capturing real time internet traffic of a 4G network using a packet capturing tool called Wireshark [18]. After this, we extracted various statistical features from the captured packets using a python script [19]. Then we applied five machine learning models, i.e., Support Vector Machine, Decision Tree, K Nearest Neighbours, Random Forest, and Naive Bayes to classify Email, Instant Messaging, P2P, VOIP, Web Media, and WWW applications. It was observed in our research work that Random Forest gave the best accuracy of approximately 87%. The final code and the dataset for our research work is available at [20].

II. RELATED WORKS

Numerous research works have been carried out for the task of classifying IP traffic, considering different types of internet applications. Many researchers have proposed various classification techniques in this field to classify IP traffic. The following subsections describe some of these research works:

A. Port Number Based Classification

In this technique, first, the ports of the internet applications are registered in the Internet Assigned Number Authority or IANA. Then, the IP traffic is classified using the IANA's list of registered port numbers [21]. As an example, the port numbers for some of the internet applications as registered in IANA are given in Table I.

TABLE I. IANA ASSIGNED PORT NUMBERS FOR SOME INTERNET APPLICATIONS

Application	Port Number
FTP	21
Telnet	23
SMTP	25
DNS	53
HTTP	80
IRC	194

As discussed in Section I, this technique is ineffective nowadays because there are many internet applications (like P2P applications) which employ port numbers which are dynamic in nature rather than well-known port numbers.

B. Payload Based Classification

This approach is also known as the Deep Packet Inspection (DPI) technique. In this approach, the internet traffic packet payload contents are analyzed, and the exact signature of the known applications is searched. This was the first alternative to the approach that was based on port number. This technique was developed specifically for P2P applications [22]. But this classification technique has many disadvantages, due to which it is not widely accepted. Firstly, this approach is not able

classify the internet traffic for which the signatures are not available. Thus, this method involves the continuous updating of the signature pattern of new applications. Very costly hardware is also required in this approach to search for patterns in a packet payload. Since the entire packet payload needs to be analyzed in this approach, a very high storage capacity and computing power is required. Moreover, nowadays the packet payload is encrypted by using different cryptographic techniques, due to which the inspection of packet payload is inhibited and thus this technique becomes ineffective.

C. Machine Learning Based Classification

As discussed in Section I, this technique is based on training a machine learning model using various statistical features which are independent of the packet payload and then using this trained model to classify IP traffic. A major benefit that this approach provides is that the inspection of packet port number or packet payload is not required. Presently, several machine learning techniques are generally used for achieving the task of IP traffic classification [4]–[17].

In [10], two datasets named HIT and NIMS were combined into a single dataset. They considered the traffic for seven different types of internet applications: WWW, DNS, P2P, FTP, IM, TELNET and MAIL. Then, SVM, ANN and C4.5 decision tree were applied for the task of classification of network traffic. For this research, the maximum accuracy was given by C4.5 decision tree.

In [5], network traffic classification technique was discussed step by step. Moreover, in this research work, a live internet traffic dataset was developed by considering DNS, WWW, P2P, FTP and Telnet applications. Also, they applied Naive Bayes and Bayes Net, C4.5 decision tree and SVM, to classify network traffic. For this research, C4.5 Decision Tree gave the highest accuracy of 78.91%. However, this research work only considered 23 features for the classification process.

In [11] & [12], a live internet traffic dataset was developed with a duration of two seconds and two minutes respectively for packet capturing from each application. This research work considered eight different internet applications: Web media, WWW, instant messaging, FTP data, E-mail, Software Updates, P2P & VOIP. New datasets were also developed by feature reduction done by feature selection using consistency & correlation-based algorithms. Finally, 5 machine learning models were applied for classifying IP traffic in these datasets: C4.5, RBF, MLP, Naïve Bayes and Bayes Net. It was concluded that for these research works Bayes Net had the best performance for accuracy & training times in all the 3 datasets.

In [13], a flow based traffic classification is done. E-mail, WWW, CHAT, FTP, Instant Messaging, VOIP and P2P applications were considered for developing an internet traffic dataset. They created two different types of datasets one with all the features and other with reduced features. Further, for IP-traffic classification they used C4.5, K-Nearest Neighbor, Naive Bayes, RBF and Bayes Net machine learning models. Bayes Net & C4.5 were the most accurate ML algorithms for the task of classifying IP traffic as proposed by this paper.

However, none of the previous research works have used a dataset generated on 4G network. Moreover, most of these research works are focused on only the classification, thus lacking practical applications. The analysis on most prominent features can have various practical applications in areas such as network management and security.

III. PROPOSED APPROACH AND EXPERIMENTAL DESIGN

The steps followed for the implementation of our research work are shown in Fig. 1. All these steps and the performance measures used are discussed in the following subsections.

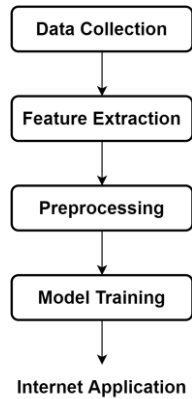


Fig. 1. Steps followed for implementation

A. Internet Traffic Data Collection

For the purpose of our research work, we developed a new dataset by capturing real time internet traffic data of a 4G network using a popular packet capturing tool called Wireshark. The source and destination ports have to be IPv6 addresses in order to determine if the obtained network is 4G. Some other popular packet capturing and analyzing tools available are tcpdump [23] and tshark [24].

We captured internet traffic data for a duration of 30 seconds for six types of applications: Email, Instant Messaging, P2P, VOIP, Web Media and WWW applications by connecting to a 4G network using hotspot from mobile data. In Table II, we have given all the applications that we considered for capturing packets for all the traffic classes mentioned before. The internet packets captured during this step were saved in .pcap format.

TABLE II. APPLICATIONS CONSIDERED FOR EACH TRAFFIC CLASS

Traffic Class	Application
Email	Gmail, Hotmail and Yahoo Mail
Instant Messaging	Whatsapp, Facebook Messenger, Microsoft Teams, Discord, Skype
P2P	UTorrent
VOIP	Google Meet, Zoom, Microsoft Teams, Discord, Skype
Web Media	Youtube, Coursera, Udemy, etc
WWW	Various websites visited using Google Chrome and Mozilla Firefox

B. Feature Extraction

After the packets were captured in a PCAP file format using Wireshark, the packets were categorized into their respective flows. A flow is a series of packets exchange for a single application which can be identified by packets having the same Destination & Source IP address & ports and protocols. Flows are bidirectional in nature and the forward direction is identified by the first packet in the flow. This was done to group packets that were of the same type so that flow features could be extracted which could be used to train the model. The features for each flow were extracted using a python script [19] by using the inherent information present in the captured PCAP file. For example, a feature named flow duration for a particular flow was calculated by measuring the time difference between the first and the last packet for that flow. Similarly, a total of 65 features were extracted for each flow. Some of these features are given in Table III. Features which are labelled to be bidirectional in the table mean that they are obtained for each direction in the communication. Otherwise, only one feature is extracted for each sender and receiver communication.

TABLE III. SOME OF THE EXTRACTED FLOW FEATURES

Feature	Bidirectional
Flow Duration	Yes
Inter-Arrival time for packet	No
Average Packet Size	Yes
Forward Packets per flow	No
Backward packets per flow	No
Download-Upload Ratio	Yes
Max Forward inter-arrival time	No
Total Number of packets per flow	Yes
Max Backward inter-arrival time	No

C. Data Preprocessing

After the extraction of flow features from the PCAP file, static features like Source and Destination IP addresses and ports were removed before training the model. Also flows containing only a single packet were removed as features like inter-arrival time require at least 2 packets in a particular flow for their calculation. Thus, features were calculated for 1899 flows. After this, the dataset was scaled using Standard Scaler to standardize the data so that no feature might dominate over other features while classification of the packets. A train-test split ratio of 3:1 was done for training and testing the model.

D. Model Training

The preprocessed data was now used to train the various machine learning algorithms present in the scikit-learn library [25]. The models used for the purposes of this research are Support Vector Machines, Decision Tree, K Nearest Neighbours, Random Forest and Naive Bayes algorithms. The trained models were now used to classify the packets into their

respective labels to identify the user activity. The final code and the dataset for our research work is available at [20].

E. Performance Measures

The evaluation of the various models were done on the following measures: Accuracy, Precision, Recall, F1-Score and Training Time. A brief description of these performance measures is given below [26], [27]:

- 1) **Accuracy:** Accuracy tells us about how much our prediction is right.

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative} \quad (1)$$

- 2) **Precision:** It is basically the total positive predictions upon total values predicted which are there in the positive class. It gives us the measure of how correct the classifier is.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (2)$$

- 3) **Recall:** It gives us the measure of completeness of a classifier. It is calculated as follows:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3)$$

- 4) **F1-score:** It is given by the following formula:

$$F1\ Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (4)$$

It basically gives us a balance between precision and recall.

- 5) **Training Time:** It is the time required for training the algorithm.

IV. RESULTS & OBSERVATIONS

The accuracy and training time of all the algorithms are shown in Table IV and are represented graphically in Fig. 2 and Fig. 3 respectively.

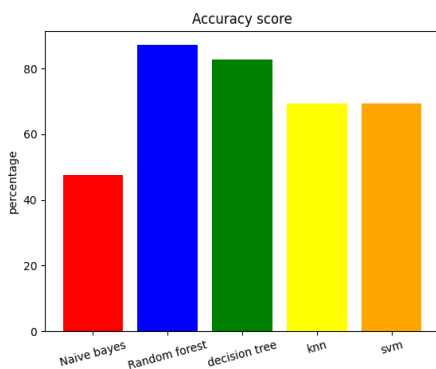


Fig. 2. Accuracy of all algorithms

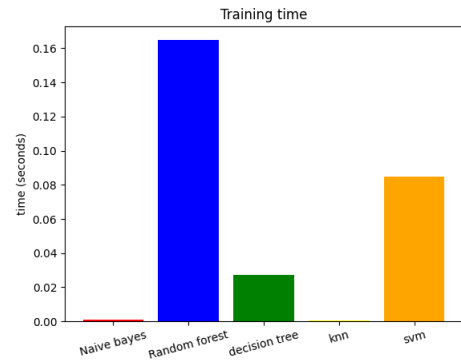


Fig. 3. Training Time of all algorithms

TABLE IV. ACCURACY AND TRAINING TIME OF ALL ALGORITHMS

Algorithm	Accuracy (%)	Training Time (s)
Naive Bayes	47.57	0.0099
Random Forest	87.15	0.17424
Decision Tree	83.00	0.2725
KNN	69.00	0.0005
SVM	69.00	0.08633

From Table IV and Fig. 2, it is clear that the accuracy of Random Forest is highest (i.e. 87.15%) among all the used algorithms. We can see from Table IV and Fig. 3 that the training time of random forest is highest and the training time of KNN is lowest among all the algorithms. Fig. 4, 5 and 6 show the precision, recall and f1-score values respectively that we calculated for different internet applications for the three most accurate algorithms i.e. Random Forest, Decision Tree and SVM. From Fig. 4, 5 and 6 it is evident that Random Forest algorithm gives best precision, recall and f1-score values for most of the internet applications as compared to other algorithms.

The features that were given the most importance were Maximum backward inter-arrival time, Max forward inter-arrival time and Download-Upload ratio for the classification according to the Random Forest classifier.

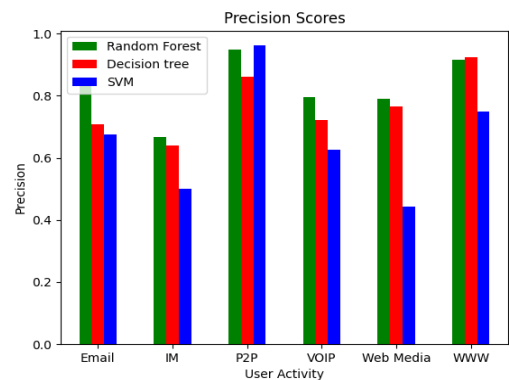


Fig. 4. Precision for three most accurate algorithms for different internet applications

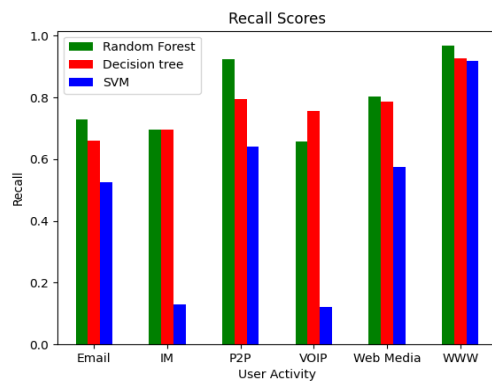


Fig. 5. Recall for three most accurate algorithms for different internet applications

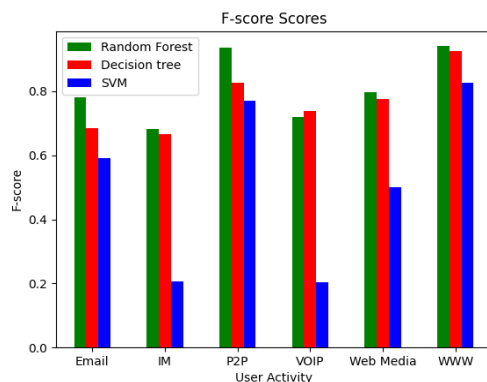


Fig. 6. F1-score for three most accurate algorithms for different internet applications

V. CONCLUSION AND FUTURE EXPANSION

During this research, a new dataset was created from packets captured using Wireshark from various sites on a 4G network. From the captured packets, 65 implicit features were extracted by using a python script by dividing the packets into various flows. After this, the dataset was preprocessed, and this dataset was used to train 5 different ML classifiers. Then these models were used to classify packets for their corresponding user activity.

By comparing all the performance measures, it was observed that the Random forest classifier was the best choice for the task of IP traffic classification among all the classifiers used in this project. Also, the most relevant features used for classifying the IP traffic i.e., maximum forward and backward inter-arrival times and Download-upload ratio, were identified by using the Feature Importance method in Random Forest. These features can have practical applications for future research works related to areas such as network management and security.

During this research, the dataset was generated by capturing packets on only 2 devices. For creating a more relevant dataset, the capturing process can be done at different environments like offices, college campus, residential facilities etc. This research can also be extended to other user activities

other than the ones used in this research. Also more sophisticated algorithms like Neural Networks may provide better results for this classification.

REFERENCES

- [1] D. S. V, "Automatic Spotting of Sceptical Activity with Visualization Using Elastic Cluster for Network Traffic in Educational Campus," *J. Ubiquitous Comput. Commun. Technol.*, vol. 2, no. 2, 2020, doi: 10.36548/jucct.2020.2.004.
- [2] T. T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE Communications Surveys and Tutorials*, vol. 10, no. 4, 2008, doi: 10.1109/SURV.2008.080406.
- [3] P. Foremski, "On different ways to classify Internet traffic : a short review of selected publications," *Theor. Appl. Informatics [Archived site]*, vol. 25, no. 2, pp. 119–136, 2013.
- [4] N. Namdev, S. Agrawal, and S. Silkari, "Recent advancement in machine learning based internet traffic classification," in *Procedia Computer Science*, 2015, vol. 60, no. 1, pp. 784–791, doi: 10.1016/j.procs.2015.08.238.
- [5] M. Shafiq, X. Yu, A. A. Laghari, L. Yao, N. K. Karn, and F. Abdessamia, "Network Traffic Classification techniques and comparative analysis using Machine Learning algorithms," 2017, doi: 10.1109/CompComm.2016.7925139.
- [6] A. A. Mohamed, A. H. Osman, and A. Motwakel, "Classification of unknown Internet traffic applications using Multiple Neural Network algorithm," 2020, doi: 10.1109/ICCIS49240.2020.9257715.
- [7] J. H. Shu, J. Jiang, and J. X. Sun, "Network Traffic Classification Based on Deep Learning," in *Journal of Physics: Conference Series*, 2018, vol. 1087, no. 6, doi: 10.1088/1742-6596/1087/6/062021.
- [8] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of encrypted and VPN traffic using time-related features," 2016, doi: 10.5220/0005740704070414.
- [9] B. Yamansavascular, M. A. Guvensan, A. G. Yavuz, and M. E. Karşilgil, "Application identification via network traffic classification," 2017, doi: 10.1109/ICCNC.2017.7876241.
- [10] M. Shafiq, X. Yu, and D. Wang, "Network traffic classification using machine learning algorithms," in *Advances in Intelligent Systems and Computing*, 2018, vol. 686, doi: 10.1007/978-3-319-69096-4_87.
- [11] K. Singh, S. Agrawal, and B. S. Sohi, "A Near Real-time IP Traffic Classification Using Machine Learning," *Int. J. Intell. Syst. Appl.*, vol. 5, no. 3, 2013, doi: 10.5815/ijisa.2013.03.09.
- [12] K. Singh and S. Agrawal, "Comparative analysis of five machine learning algorithms for IP traffic classification," 2011, doi: 10.1109/ETNCC.2011.5958481.
- [13] A. (Karunya U. Jamuna and V. (Karunya U. Edwards S.E, "Efficient Flow based Network Traffic Classification using Machine Learning," *Int. J. Eng. Res. Appl.*, vol. 3, no. 2, 2013.
- [14] T. S. Tabatabaei, F. Karray, and M. Kamel, "Early internet traffic recognition based on machine learning methods," 2012, doi: 10.1109/CCECE.2012.6335034.
- [15] J. M. Wang, C. L. Qian, C. H. Che, and H. T. He, "Study on process of network traffic classification using machine learning," 2010, doi: 10.1109/ChinaGrid.2010.53.
- [16] D. Qin, J. Yang, J. Wang, and B. Zhang, "IP traffic classification based on machine learning," 2011, doi: 10.1109/ICCT.2011.6158005.
- [17] V. Labayen, E. Magaña, D. Morató, and M. Izal, "Online classification of user activities using machine learning on network traffic," *Comput. Networks*, vol. 181, 2020, doi: 10.1016/j.comnet.2020.107557.
- [18] "Wireshark · Go Deep." <https://www.wireshark.org/> (accessed Feb. 05, 2021).
- [19] "GitHub - anupamraj1312/Flowmeter: A python script for extracting flow features from a PCAP file." <https://github.com/anupamraj1312/Flowmeter> (accessed Feb. 07, 2021).

- [20] "GitHub - anupamraj1312/4G-IP-traffic-classification-using-Machine-learning." <https://github.com/anupamraj1312/4G-IP-traffic-classification-using-Machine-learning> (accessed Feb. 07, 2021).
- [21] "Service Name and Transport Protocol Port Number Registry." <https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml> (accessed Feb. 05, 2021).
- [22] M. Finsterbusch, C. Richter, E. Rocha, J. A. Müller, and K. Hänßgen, "A survey of payload-based traffic classification approaches," *IEEE Commun. Surv. Tutorials*, vol. 16, no. 2, 2014, doi: 10.1109/SURV.2013.100613.00161.
- [23] "TCPDUMP/LIBPCAP public repository." <https://www.tcpdump.org/> (accessed Feb. 05, 2021).
- [24] "tshark - The Wireshark Network Analyzer 3.4.3." <https://www.wireshark.org/docs/man-pages/tshark.html> (accessed Feb. 05, 2021).
- [25] "scikit-learn: machine learning in Python — scikit-learn 0.24.1 documentation." <https://scikit-learn.org/stable/> (accessed Feb. 05, 2021).
- [26] "20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics | by Shervin Minaee | Towards Data Science." <https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce> (accessed Feb. 06, 2021).
- [27] "Machine Learning - Performance Metrics - Tutorialspoint." https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_algorithms_performance_metrics.htm (accessed Feb. 06, 2021).

Justifying Biofield (Aura) Studies as Complementary and Alternative Medicine (Cam)

Ankit Dutta¹, Subnear Kour², Dr. Priyanka Jain³

^{1,2}Dept. of Electronics & Communication Engineering., Delhi Technological University, India

³ Associate Professor, Dept. of Electronics & Communication Engineering., Delhi Technological University, India

ABSTRACT

Human biofield refers to the Electromagnetic (EM) field emitted by the human body. This radiation is a very faint one. This paper studies the various instruments used for its detection, i.e., Dipole Antenna, Lecher Antenna and Frequency Wave Detector (FWD). After establishing FWD as better method of EM radiation measurement, the best configuration of measurement using a FWD is studied. The study is concluded by reviewing EM distribution throughout the human body and previous studies in the Biofield domain is carried out. Reviewing all these factors highlight the capabilities of using human biofield as a complementary and alternative medicine (CAM).

Keywords

Biofield, Aura, Electromagnetic, Frequency

INTRODUCTION

Human body involves physiological and biological interactions within itself and with its surroundings. This leads to exchange in energies amidst environment and humans. This biochemical exchange of energies is termed as 'biofield' [1][2][3]. Further analysis shows the presence of Electromagnetic (EM) field. This radiation is also called aura [4]. These EM radiations are present in different characteristic frequencies and independent intensities [5][6]. Such EM radiations show the presence of electric currents in our bodies. There is a huge scope of use cases for biofield in various medical fields which have been studied in the subsequent sections. It enables us to perceive humans' spiritual, mental and physical states in a more visual manner [7].

ELECTROMAGNETIC ASPECT OF BIOFIELD

The generation of EM fields by the human body requires detailed study of the EM field. There are two categories of EM fields, the first one is high frequency oscillating and coherent EM field and the other one has two aspects: the Frolich field, and the Popp photon field. The Frolich field is a microwave to MHz to a lower frequency range coherence. The second one is visible/near ultra-violet/infrared diffuse fields. The Frolich field has been observed but at lesser frequencies than predicted. The pop field is supported by observations of the statistical coherence of biophotons. Our current examination of biofields proves the need to go beyond classical physics and biology. EMFs or quantum and quantum-like processes [8] and other coherent states may be the carriers of biofields.

INSTRUMENTS USED FOR MEASURING BIOFEILD

The EM field consists of two components, namely the electrical and magnetic parts. It is accompanied by a self-propagating wave. The wave is periodic in fashion and thus has its own

characteristic frequency, amplitude and wavelength. For the purposes of this study, we need to measure most importantly the frequency of EM wave. The three most common devices used for measuring EM frequencies of biofield radiation are, Dipole Antenna, Lecher Antenna and Frequency Wave Detector.

Dipole Antenna

A dipole antenna is the most widely used antenna class. They can be used on their own, and as a part of some other antenna class as well. They can be used in broadcasting, radio communication and in many similar fields.

In ultra-high field MRI, dipole antenna has many advantages over conventional designs. The fractionated dipole antenna is now being used; it is a new device that is used for body imaging at 7 Tesla. In this antenna, the legs of the dipole are split into segments, interconnected by inductors or capacitors [9].

A study has been done on the length of dipole antenna using numerical simulations. In this study, an optimal design has been developed and compared with the previous design, the single side adapted dipole (SSAD).

Lecher Antenna

The lecher antenna is an advanced electronic instrument used for manual measurements in the biofield domain. It allows the use of the biological sensitivity of a man to measure even the most subtle electromagnetic biofields.

During the last forty years, the antenna enabled to discover both the presence of biologically interesting natural electromagnetic fields and the specific vital frequencies nourishing and spread by each human organ [10]. A skilled operator with the antenna can measure accurately their intensity on each polarity separately, obtaining in this way the best information about the biological effects of electromagnetism on the vital processes. With this instrument, electromagnetic field surrounding the human body could be detected.

Frequency Wave Detector

The frequency wave detector is commonly used for the purposes of EM radiation detection of human aura. The human body releases EM radiation which can be easily picked up by the detector [5]. It is a hand-held device which is equipped with a specially tuned antenna and is used as a frequency meter [11].

The tuning is done such that an accurate reading as well as a real-time reading of the aura frequencies at the testing points. Due to the weak nature of aura signals, the frequency wave detector is retrofitted with a highly sensitive synchronous detector and a filter module which is able to block out random noise [11]. It operates in the Gigahertz and Megahertz frequency range. Several studies were conducted using the above-mentioned instrument where the study was conducted in the Gigahertz range for the purposes of frequency measurement [12].

One of the first studies that is reviewed was conducted by Kadir, R. S. S. A., et al. It involved a total of 115 patients who were a part of the National Stroke Association of Malaysia (NASAM) [13]. A total of 16 measurements is taken from the left and right side of the body. The second study reviewed which uses a frequency wave detector involved 41 participants. Out of the forty-one participants, there were a total of thirty-one patients ailing from kidney diseases and the rest were non-kidney disease patients. For the patients with kidney diseases, the measurements were taken twice, before and after the hemo-dialysis. For maintaining accuracy, three readings were taken at each point [14]. The third study which was reviewed, involved taking measurements on

16 points around the human body. There are eight points of measurement, each on the left-side and right-side. Data is also collected from the seven points of the chakra system [15].

Configuration of frequency wave detector

To obtain the proper configuration for measurement of biofield frequencies from an electromagnetic detector, A. Jalil et Al. conducted a study [12]. 10 healthy participants (5 male and 5 female volunteers) in the age group of 26-38 years took part in this study. The participants were placed in an air-conditioned room where they were made to stand in a comfortable spot. To avoid any discrepancies, background frequencies in ambient conditions were noted both after and before the measurements [16] and all the measurements taken from the participants were done at the same place. Measurements were taken from all the seven chakras.

The frequencies were noted at distances ranging from a centimeter to 10 centimeters at each point of measurement. Lengths of the antenna were adjusted from the first to the seventh segment. This selection was shortened to the length of the antenna being in the third, fifth and seventh segment, and a distance of 1 cm, 5 cm and 10 cm was chosen as they signify the lower, center and upper position of the setup. Boxplot analysis was done for the measurements satisfying these parameters.

From the measurements taken, males and females yielded 37 and 22 outliers respectively. Most of them were from distance of 10 cm and the antenna adjusted for the third and fifth segment. Though readings at 1 cm and 5 cm and antenna segments 5 and 7 provide more stable and reliable frequency reading. The lowest frequency standard deviation occurs at 1 cm and the highest occurs at 10 cm. In conclusion, measurements are best performed at distances of 1 cm and 5 cm with the antenna adjusted for the fifth and seventh segment. Thus, making body radiation wave detector a suitable option for detecting human biofield frequencies.

FREQUENCY DISTRIBUTION OF BIOFIELD IN HUMAN BODY

Based on a study conducted by A. Jalil et Al., where a total of 33 participants (17 male and 16 female participants) in the age group of 19-26 years took part. The measurements were taken at the seven chakra points sixteen other points on both the left and right side of the body. The experimental setup involved an anechoic chamber which was temperature-controlled chamber at $23 \pm 2^\circ\text{C}$ where the floor had a stationary ferrite stand. Movement of subject was limited. Background frequencies were measured prior to and after [16] the experiments were conducted for the purposes of generating usable data. All this was done for the purposes of reducing the effect of environmental frequencies.

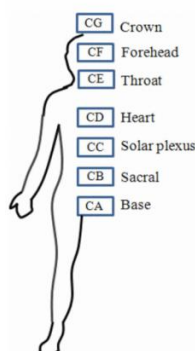


Fig. 1 Chakra positions

The chakra positions of the human body are represented using Fig. 1. The first part of the study [11] involved analyzing human radiation frequency at the seven Chakra points and the abovementioned sixteen points. Boxplot analysis was made use of for this study. Fig. 2,3 and 4 displays distributions of biofield frequencies. Individual analysis of the data shows the frequency distribution of radiation at chakra, left and right side of the bodies of females and males differ. Males have higher frequency ranges than females in the left and right-side boxplot distributions. The maximum difference frequency distribution was observed at R6 and L6 for males and females with a difference of 33.58 MHz and 31.41 MHz respectively. The left side generates a 48% mean frequency distinction while the left side generated 52%.

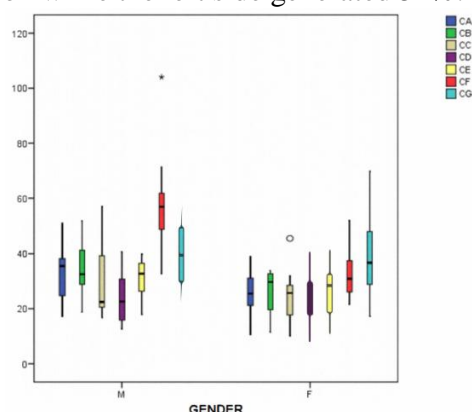


Fig. 2 Chakra

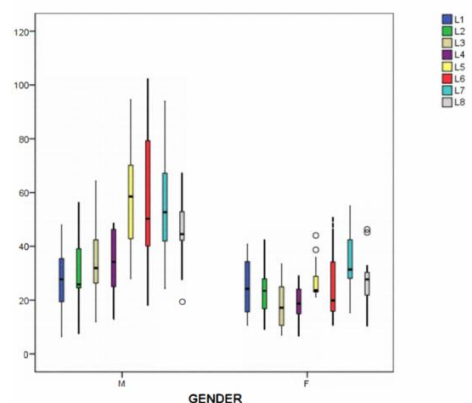


Fig. 3 Left side

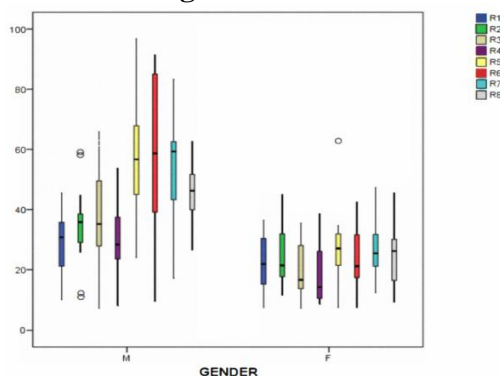


Fig. 4 Right side

Furthermore, a scatterplot analysis was used to analyze the relationship of frequency distribution between females and males. Fig. 5,6 and 7 shows the non-overlapping data for all points of measurement between males and females. For almost all groups on the left and right side has moderate positive correlation, but the CG chakra group displays a curvilinear relationship. Fig. 8 and 9 displays for chakra groups its individual scatterplots for both males and females, this is done to establish the strength of relationship between the variables. Males as whole participant has moderate relationship, meanwhile for females it was a blob-type arrangement (weak linear relationship). The relationship between the chakra groups was found using the Pearson product moment correlation coefficient, which establishes linear correlation. An example of this in this study is the correlation factor between CD and CC for males and females which was $r = 0.63$ and $p=0.007$, and $r = 0.15$ and $p=0.58$ respectively.

Fig. 5 Chakra

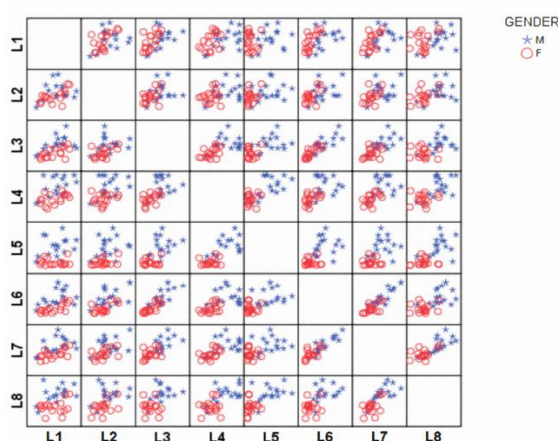


Fig. 6 Left side

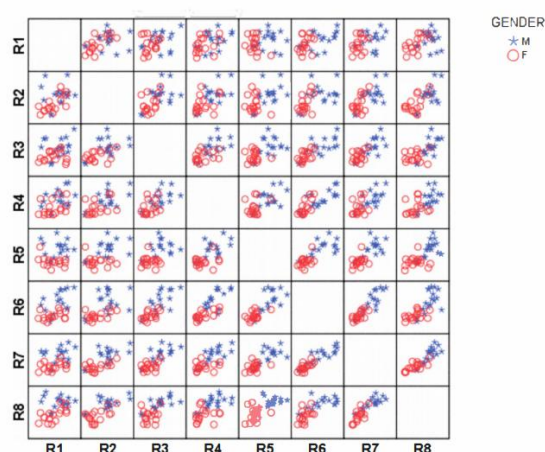


Fig. 7 Right side

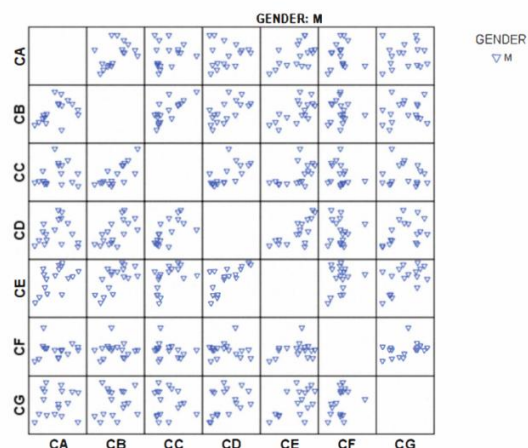


Fig. 8 Male

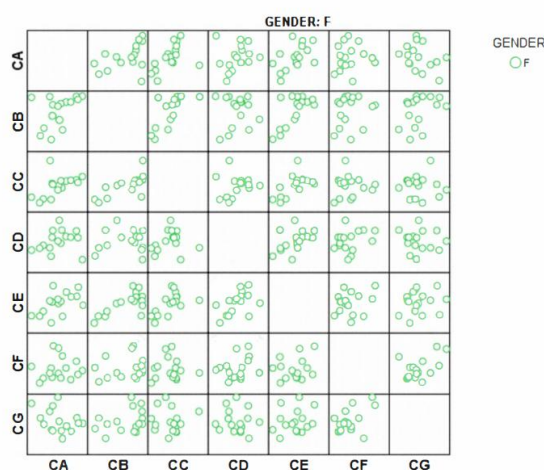


Fig. 9 Female

Thus, males are seen having higher frequencies ranges of biofield than females and, males and females have a difference relationship of frequency emission.

TABLE 1: prior studies on biofeild

References	Number of Applicants	Medical Studies
Alexandrova et al. (2003a) [17]	247 out of the 303 total participants were observed to suffer from bronchial asthma, the remaining 56 patients were healthy.	Bronchial asthma patients participated in experimental analysis.
Gimbut et al. (2004) [18]	The number of applicants in this analysis was 20.	Experimental analysis was done on factors responsible for uterus imbalance.
Alexandrova et al. (2003c) [19]	The number of applicants in this analysis was 43. 23 out of 43 were seen to be having allergic reactions and the remaining 20 were healthy.	Allergic reaction risk was calculated using experimental analysis.

References	Number of Applicants	Medical Studies
Gedevanishvili et al. (2004) [20]	The number of applicants in this analysis was 57. 22 out of the 57 were suffering from lung cancer and the remaining 35 are suffering from breast cancer.	Patients suffering from breast and lung cancer undergoing radiotherapy treatment participated in experimental analysis for their optimal assessment.
Alexandrova et al. (2003b) [21]	The number of applicants in this analysis was 87. 30 out of the 87 were suffering from chronic viral hepatitis while 25 out of 87 were ailing from chole-lithiasis and the remaining 32 out of 87 are ailing with primary Biliary dyskinesia.	Patients suffering from Chronic viral hepatitis undergoes experimental analysis.
Gagua et al. (2004) [22]	The number of applicants in this analysis was 347. 249 out of 347 were cancer patients and rest were in the control group.	The state of a cancer patient is determined using experimental analysis.
Krashenuk et al. (2006) [23]	The number of applicants in this analysis was 21.	Experimental analysis of monitoring the therapeutic effect.
Gagua et al. (2004) [22]	The number of applicants in this analysis was 347. 109 out of the 347 were suffering from lung cancer while 140 out of 347 were ailing from breast cancer and the remaining 98 were in the control group.	Statistical analysis is done on patients suffering from breast and lung cancer and healthy people by analyzing their biofield.

Table I lists the various studies that have been carried out in the domain of biofield analysis.

METHODS FOR BIOFEILD ANALYSIS

Based on the research efforts that have already been conducted in the Biofield domain, there are multiple methods to measure and analyze the biofield of a human or any living object for that matter:

- BiopulsarReflexograph.
- Aura color space visualizer algorithm.
- Quantum resonance magnetic (QRM) analyzer.

BiopulsarReflexograph

It is one of the most accurate ways of identifying problems with bodily functions and organs using aura/biofield as the method of measurement. Its main function is that of energy measurement of the human body which is done by generating results in the form of activities of chakras and various graphs of the organs. It also generates the aura image of the entire body. Different organs of our body, consciousness, subtle energy centers, and meridians have a connection to the different reflex-zones (certain parts of hands). The Biopulsarreflexograph method makes use of this concept and then takes energy readings from each reflex zone. The

frequencies of the energy readings are then represented using various color codes. This helps in visualizing the human biofield/aura.

The instruments involved in this method are useful for giving an overall preview of a human's health on the basis of its built-in software. These readings provide the energy readings of the chakras along with a list of organs that are associated with each chakra. Graphical representation of every organ's energy level is provided by the built-in software functionality. These are then compared to threshold values already mentioned in the software. If the result is lower than the threshold value, it is an indication of present or is a future prediction of illness that might occur within that region (organ).

Comparing these results to already available traditional medical reports one can observe high accuracy of about 85% [24].

Aura color space visualizer algorithm

The electromagnetic radiation emitted from the human aura lies outside the visible range, hence to observe such results we need a method to represent aura in a human interpretable form which will help in analyzing results and also to observe patterns [25].

One way of doing this is by using the aura color space visualizer algorithm which is an image-processing method used for the purposes of human bio-field detection.

Since aura isn't visible to the naked eye, we have to define a new color model. To satisfy this requirement, this algorithm maps the dominating pixel values with visible RGB values which makes it visible. This method is formally known as the pixel manipulation method.

The measurement setup to execute the given method is relatively cheaper as its major components of expenditure are only the camera, software, and a light source. The accuracy of this method has been calculated as a percentage of the number of correct results to total records. After studies carried out by [24] they received 63% accuracy based on their study carried out in static laboratory environments.

Barring the advantages of using this method, we have a few disadvantages too, one is a limitation of being incapable of providing precise results and the lack of a standard scale of measurement [24].

The future scope of improvements lies in applying several image processing techniques such as image enhancement and transformation techniques for obtaining better results.

Quantum Resonance Magnetic Analyzer

The health condition of a person can be also determined through the emission of the EM waves of the human body using a QRM analyzer.

QRM analyzer analyses the health condition of a human body by measuring the EM radiations from the human body. It both measures and analyses the radiations. These radiations are emitted during cell regeneration.

The QRM analyzer used by Chhabra et al. in their studies [24], analyses 36 parameters in 60 seconds. It also has amplification techniques that are built-in along with a microprocessor that calculates input signals to a standard quantum spectrum. The microprocessor makes use of the Fourier series and can thus identify over 30 different functionalities of the human body such as liver, kidney, brain, etc.

The level of condition is represented by a color scale where; red - very low, yellow - low, blue - tolerated and green - well.

CONCLUSION

This paper has successfully studied the various methods of biofield detection, and also established the best configuration for human biofield using FWD based on previous studies. These involved boxplot analyses based on several sets of measurements. Future studies can be done in the field of biofield making use machine learning, such that greater datasets can be analysed at ease and more conclusive studies can be done such that we can use biofield as a complementary and alternative medicine.

REFERENCES

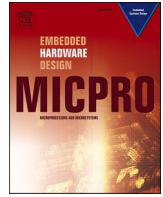
- [1] G. Chhabra, A. Narayanan, S. Samantha, S. Samanta, "Human aura: A new vedic approach in IT," 2013, pp. 5–7.
- [2] K. Korotkov, Energy fieldselectrophotonic analysis in humans and nature. Ebookit.com, 2013.
- [3] B. Rubik, "Measurement of the human biofield and other energetic instruments," Mosby's complementary & alternative medicine: A research-based approach, pp. 61–87, 2009.
- [4] C. T. Tart, "Concerning the scientific study of the human aura," Journal of the Society for Psychical Research, vol. 46, no. 751, pp. 1–21, 1972.
- [5] I. Item, "Technologies and energy medicine. Resonant Field Imaging (RFI)," 2004.
- [6] A. R. Liboff, "Toward an electromagnetic paradigm for biology and medicine," The Journal of Alternative & Complementary Medicine, vol. 10, no. 1, pp. 41–47, 2004.
- [7] D. C. Lewis, Advanced Studies of the Human Aura: How to Charge Your Energy Field with Light and Spiritual Radiance. Meru Press, 2013.
- [8] M. C. Kafatos, G. Chevalier, D. Chopra, J. Hubacher, S. Kak, N. D. Theise, "Biofield Science: Current Physics Perspectives," Global advances in health and medicine, vol. 4, pp. 25–34, 2015.
- [9] A. J. E. Raaijmakers et al., "The fractionated dipole antenna: A new antenna for body imaging at 7 Tesla," Magnetic resonance in medicine, vol. 75, no. 3, pp. 1366–1374, 2016.
- [10] M. Nieri, P. K. Singhania, "BIOENERGETIC LANDSCAPES REDUCE STRESS AND RESTORE HEALTH USING ELECTROMAGNETIC PROPERTIES OF PLANTS," Building Organic Bridges, vol. 3, pp. 701–704, 2014.
- [11] S. Z. A. Jalil, M. N. Taib, H. A. Idris, M. M. Yunus, "Examination of human body frequency radiation," 2010 IEEE Student Conference on Research and Development (SCORed), pp. 4–7, 2010.
- [12] S. Z. A. Jalil, M. Y. M. A. Karim, H. Abdullah, M. N. Taib, "Instrument system setup for human radiation waves measurement," 2009 IEEE Student Conference on Research and Development (SCORed), pp. 523–525, 2009.
- [13] R.S.S.A. Kadir, Z. H. Murat, M.Z. Sulaiman, M. N. Taib, F. A. Hanapiah, W.R.W. Omar, "The preliminary investigation of electromagnetics radiation for the left hemisphere stroke," 2014, pp. 606–610.
- [14] S. Z. Jalil et al., "Investigation of human electromagnetic radiation characteristic for kidney disease patients," International Journal of Engineering & Technology, vol. 7, no. 4.11, pp. 40–43, 2018.
- [15] R.S.S.A. Kadir, Z. H. Murat, M.N. Taib, S. Z. A. Jalil, "Investigation of electromagnetics radiation for stroke patients and non-stroke participants," 2015, pp. 130–134.
- [16] K. J. Hintz, G. L. Yount, I. Kadar, G. Schwartz, R. Hammerschlag, S. Lin, "Bioenergy

- definitions and research guidelines,” *Alternative Therapies in Health and Medicine*, vol. 9, no. 3; SUPP, pp. A13–A30, 2003.
- [17] R. Alexandrova et al., “Analysis of the bioelectrograms of bronchial asthma patients,” 2003, pp. 70–81.
 - [18] V. S. Gimbut, A. V. Chernositov, and E. V. Kostrikina, “GDV parameters of woman in phase dynamics of menstrual cycle,” 2004, pp. 80–82.
 - [19] R. A. Alexandrova, V. I. Trofimov, E. E. Bobrova, and V. K. Parusova, “Comparison of dermal allergology test results and changes of GDV bioelectrograms in case of contact with phytocosmetic substance in test tube,” 2003, pp. 1–4.
 - [20] E. G. Gedevanishvili, L. G. Giorgobiani, and A. Kapanidze, “Estimation of radiotherapy effectiveness with gas discharge visualization (GDV),” 2004, pp. 98–99.
 - [21] R. A. Alexandrova, V. I. Nemtsov, D. V. Koshechkin, and S. U. Ermolev, “Analysis of holeodoron treatment effect on cholestasis syndrome patients,” 2003, pp. 4–6.
 - [22] P. O. Gagua, G. L. G. Ge, and A. Kapanadze, “The GDV technique application to oncology,” *Measuring energy fields: state of the science*, pp. 43–50, 2004.
 - [23] A. I. Krashenuk, A. D. Danilov, and K. G. Korotkov, “Investigation of system optimization of vegetative nervous system work under hirudotherapy impact as a result of comparative analysis of GDV signal and cardiorythm nonlinear analysis,” *Proceedings of X International Scientific Congress on Bioelectrography*, pp. 31–35, 2006.
 - [24] G. Chhabra, A. Prasad, V. Marriboyina, “Comparison and performance evaluation of human bio-field visualization algorithm,” 2019, pp. 1–12.
 - [25] G. Chhabra, A. Prasad, and V. Marriboyina, “Implementation of aura colorspace visualizer to detect human biofield using image processing technique,” *Journal of engineering science and technology*, vol. 14, no. 2, pp. 892–908, 2019.



Contents lists available at ScienceDirect

Microprocessors and Microsystems

journal homepage: www.elsevier.com/locate/micpro

Leakage reduction in dual mode logic through gated leakage transistors

Neetika Yadav^{a,b}, Neeta Pandey^{a,*}, Deva Nand^a^a Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, 110042, India^b Department of Electronics and Communication Engineering, Amity School of Engineering and Technology (affiliated to GGSIPU), Sector 125, Noida 201303, India

ARTICLE INFO

Keywords:

Dual mode logic
Dynamic
Footed diode transistor
GALEOR
Static

ABSTRACT

This contribution proposes a technique for leakage power reduction in Dual Mode Logic (DML) circuits by incorporating Gated Leakage Transistor (GLT). The resulting circuits are named as GALEOR with Dual Mode Logic (GDML). Further, GDML design is extended by including a footed diode transistor, the design so obtained is referred to as GALEOR with Dual Mode Logic with footed diode (GDMLD). The analysis is done using footed type A and type B DML gates, resulting in GDML and GDMLD variants referred to as GDML-TA, GDML-TB, GDMLD-TA and GDMLD-TB. Two input NAND and NOR gates along with a full adder and a 2-bit multiplier circuit are used to investigate the proposed techniques at 90 nm and 45 nm technology nodes in both static and dynamic mode using SymicaDE tool. Analysis of leakage power reveals that its value increases with technology scaling. Average leakage power saving is 44.69%–74.11% for GDML and 67.18%–90.76% for GDMLD in static mode. Similarly, in pre-charge phase of dynamic mode, this value varies from 5.47%–28.22% for GDML and 14.55%–77.51% for GDMLD. For evaluation phase, average leakage power saving of 44.69%–74.11% for GDML and 67.18%–90.76% for GDMLD is achieved. Analysis of delay reveals that both the techniques increase delay of the design while providing significant leakage power saving.

1. Introduction

Aggressive technology scaling in recent years has led to the development of high-performance devices. However, this comes at the cost of increased power in the designs [1]. Power can be divided mainly into two types- Static and Dynamic power [2]. In CMOS designs, different types of leakage constitute static power and dynamic power consists of switching, short circuit and glitching power [1]. To minimize power consumption, efforts are made to propose various techniques [3–30]. These power reduction techniques can be broadly categorized into two categories based on two major principles. The first category is based on the principle of controlling the power supply of the design [3–18] and the second category of techniques employ logical effort approach [19–23]. The adiabatic logic [3–9] falls under first category and uses power clocks in place of constant DC source. It minimizes energy dissipation by slowing down the charging and discharging of a node. It has an additional property of energy recycling from the circuit to the power-supply by using specially designed power-clock generator [3–9]. Multiple supply voltage design [10–13] and subthreshold region operation [14–16] are also representatives of first category which work on employing dual power supply and reduced power supply respectively.

Dual supply voltage or dual VDD [10] technique assigns different supply voltages to different paths in the design depending on the criticality of the path. Low supply voltage is assigned to non-critical paths and high supply voltage to critical paths using algorithms so that the performance is not compromised. Subthreshold transistor operation [14–16] is another technique that involves MOSFET operation in subthreshold region where the supply voltage is reduced to a value less than threshold voltage. Power gating [17–18] is yet another technique of first category wherein the connection to power supply ceases to exist in sleep mode. This technique uses additional transistor(s) and sleep signals for proper operation. The second classification involves optimizing the size of transistors in a design using a mathematical model at circuit level, referred to as logical effort [19–23], which greatly reduces power. Till 90 nm technology node, dynamic power dominates the total power consumption but with the advent of technology scaling, there is an increase in leakage power due to smaller feature size and threshold voltage reduction [2]. So, it necessitates the development of leakage control techniques to address the issue of leakage power. Substantial research has already been done to devise such techniques for different logic families [24–30].

Apart from this, there is an urgent need for alternative logic styles

* Corresponding author.

E-mail address: n66pandey@rediffmail.com (N. Pandey).<https://doi.org/10.1016/j.micpro.2021.104269>

Received 20 July 2020; Received in revised form 10 March 2021; Accepted 21 April 2021

Available online 4 May 2021

0141-9331/© 2021 Elsevier B.V. All rights reserved.

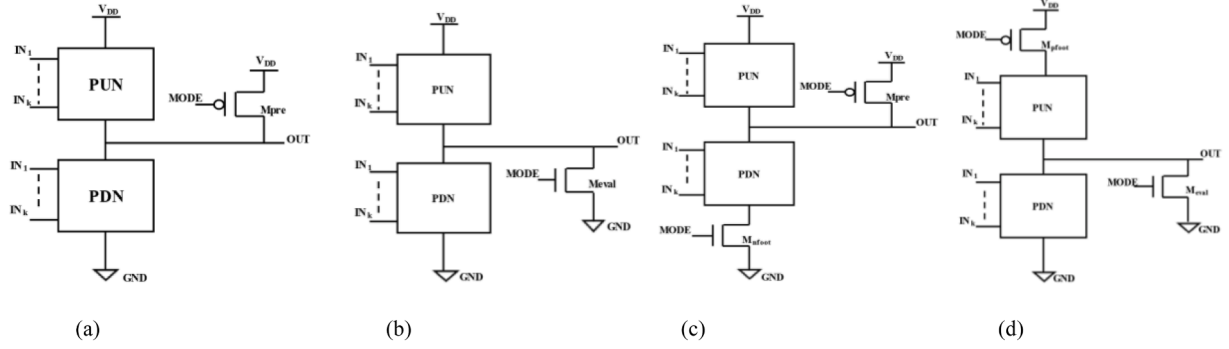


Fig. 1. Dual Mode Logic topologies: (a) Type A [40] (b) Type B [40] (c) Footed Type A [40] (d) Footed Type B [40]

which can satisfy both energy and performance requirements of the design. DML family is one such design alternative which allows two operational modes-static and dynamic in a design [31]. The logic can be implemented using two topologies- type A and type B-both of which can operate in static and dynamic mode. Low power consumption is achieved in static mode and the dynamic mode exhibits high performance [32]. Its structure includes a static CMOS gate with an additional PMOS transistor for type A and NMOS transistor for type B.

Limited study has been conducted to reduce leakage power [33-37] in the context of DML circuits. Two leakage control techniques are mainly used to control total power consumption in DML. First technique

is power gating approaches- sleep, sleepy stack and dual sleep- which have been studied, mainly for basic DML gates- NOT, NAND and NOR and sequential designs in both type A and type B topology [33-36]. Second approach employs multi-threshold concept where the effect of variation of threshold voltage of additional transistor on leakage is investigated in type A NOR and type B NAND gate [37]. In power gating technique, sleep transistor incurs additional overhead of control signals and appropriate timing to generate sleep signal [33]. The multi-threshold concept requires identification of transistors whose threshold is to be modified. To alleviate these drawbacks, this paper introduces GALEOR (Gated Leakage TransistOR) technique to combat

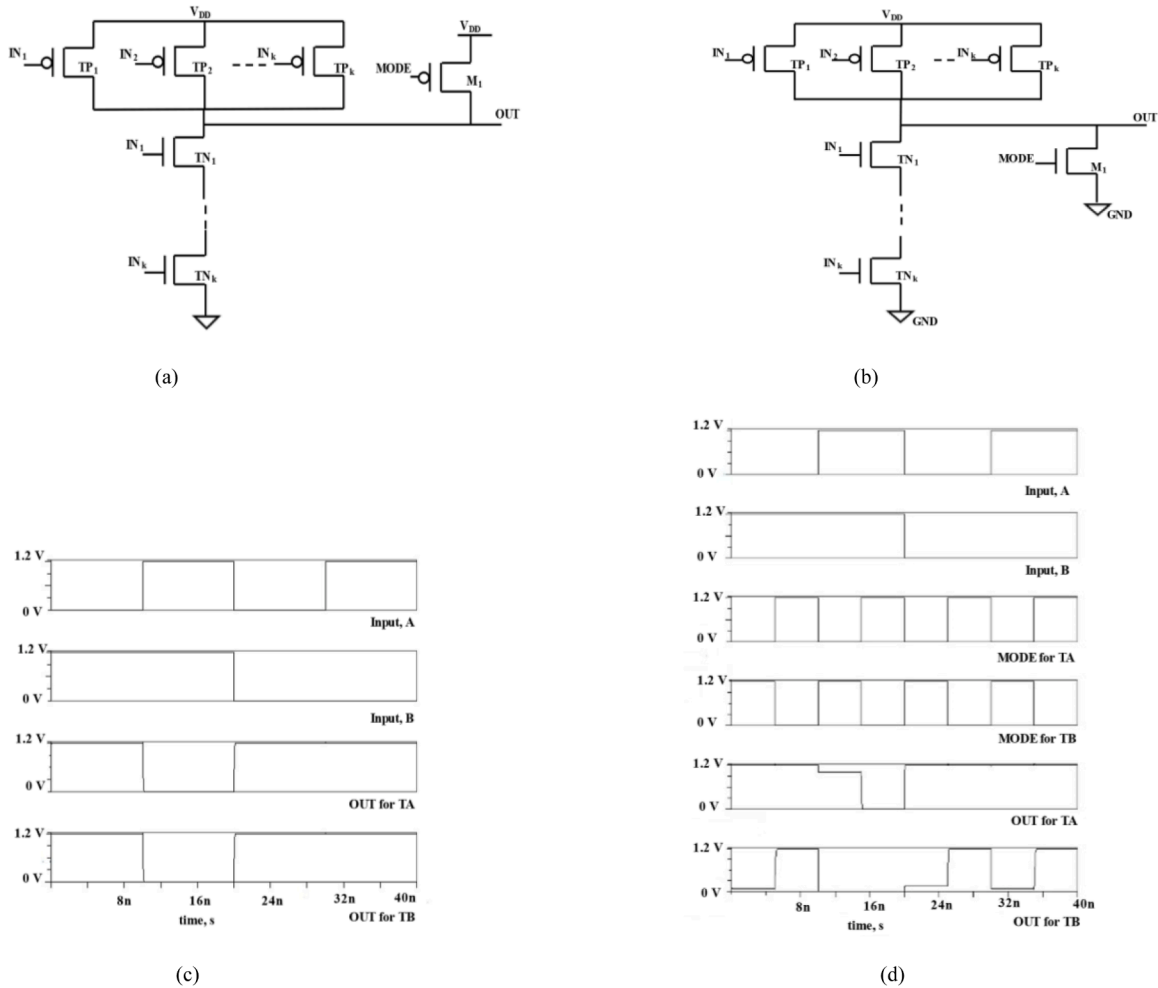
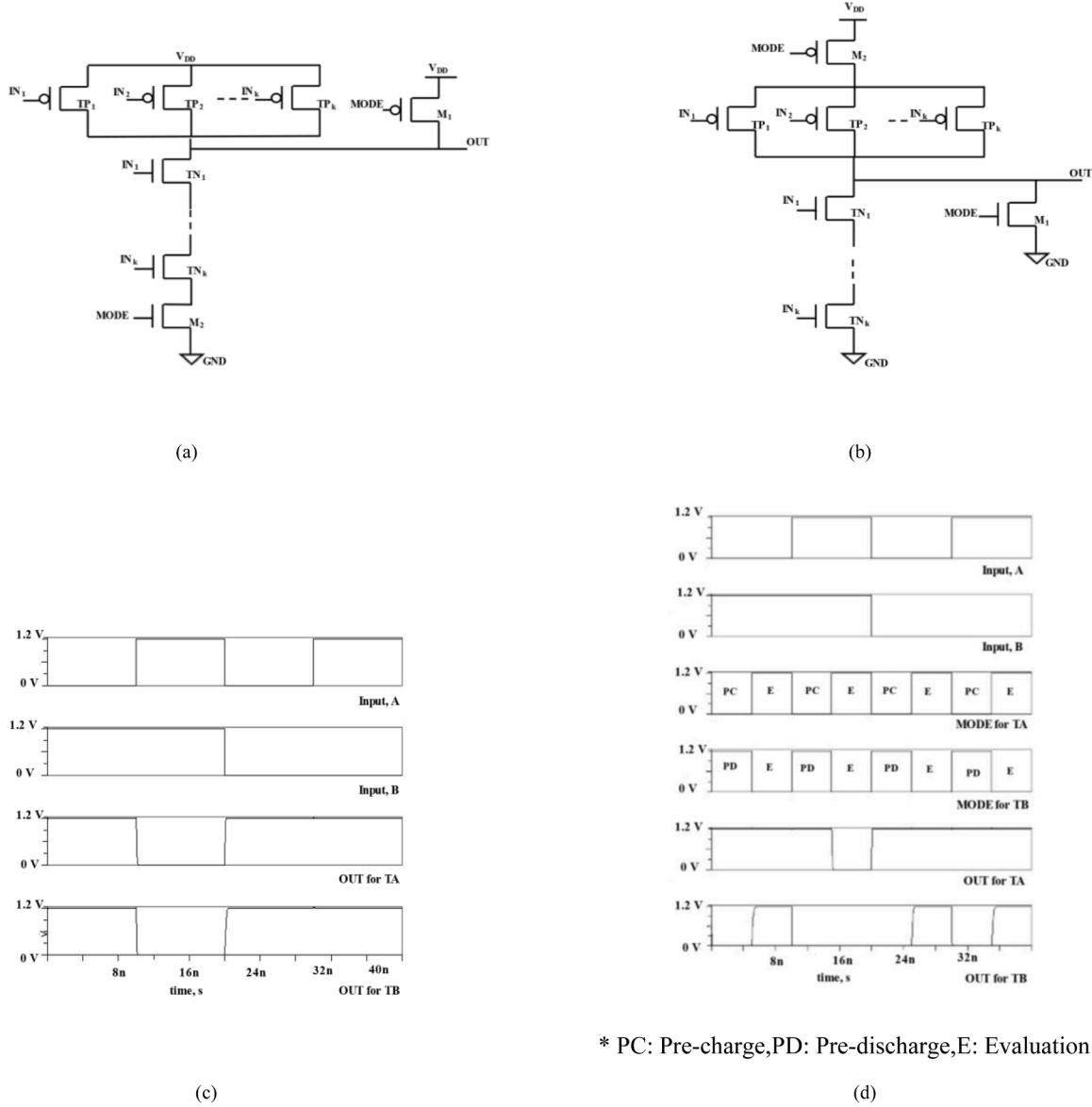


Fig. 2. DML unfooted (a)type A NAND gate (b) type B NAND gate (c) timing waveform of 2-input type A(TA) and type B(TB) NAND gates in static mode (d) timing waveform of 2-input type A(TA) and type B(TB) NAND gates in dynamic mode at 27 °C



* PC: Pre-charge, PD: Pre-discharge, E: Evaluation

Fig. 3. DML footed (a) type A NAND gate (b) type B NAND gate (c) timing waveform of 2-input type A (TA) and type B (TB) NAND gates in static mode (d) timing waveform of 2-input type A (TA) and type B (TB) NAND gates in dynamic mode at 27 °C

leakage in footed DML circuits [38]. GALEOR is a self-controlled technique which employs GLTs to reduce leakage and it avoids any additional circuitry to control GLTs [30].

The organisation of the paper is as follows: section 2 gives an overview of DML architecture, types of DML designs and the associated leakage mechanism. In section 3, the proposed GALEOR based DML logic and GALEOR with footed diode [39] approach along with associated leakage equations is explained. Section 4 states the simulation results for 2-input GALEOR with DML type A and type B NAND and NOR gate (GDML-TA-NAND2, GDML-TA-NOR2, GDML-TB-NAND2, GDML-TB-NOR2). This section further includes delay analysis of proposed approaches, simulation results for 2-input GALEOR with footed diode transistor for type A and type B NAND and NOR gates (GDMLD-TA-NAND2, GDMLD-TA-NOR2, GDMLD-TB-NAND2, GDMLD-TB-NOR2), effect of load capacitance, full adder and 2-bit multiplier design and subsequently, the conclusion of the paper is placed in section 5.

2. Overview of DML architecture

Dual mode logic family, proposed in [32], allows two modes of operation with the help of a mode signal-Static and Dynamic mode. The static mode provides the benefit of energy saving and dynamic mode assures high performance. A conventional DML gate is obtained by attaching a pre-charge (M_{pre}) / pre-discharge (M_{eval}) transistor to a static CMOS gate. This logic family has two variants-unfooted and footed structure, as shown in Fig. 1.

2.1. Unfooted DML design

The unfooted DML design consists of PUN and PDN with an extra pre-charge/pre-discharge transistor. It can be implemented via two topologies- Type A and Type B as shown in Fig. 1(a-b). An additional pre-charge transistor (M_{pre}) and pre-discharge (M_{eval}) transistor is added at the output of a conventional CMOS gate for type A and type B topology respectively. DML based unfooted type A NAND and type B NAND gates are shown in Fig. 2 (a-b). The working of unfooted type A NAND and type B NAND is illustrated through timing waveforms

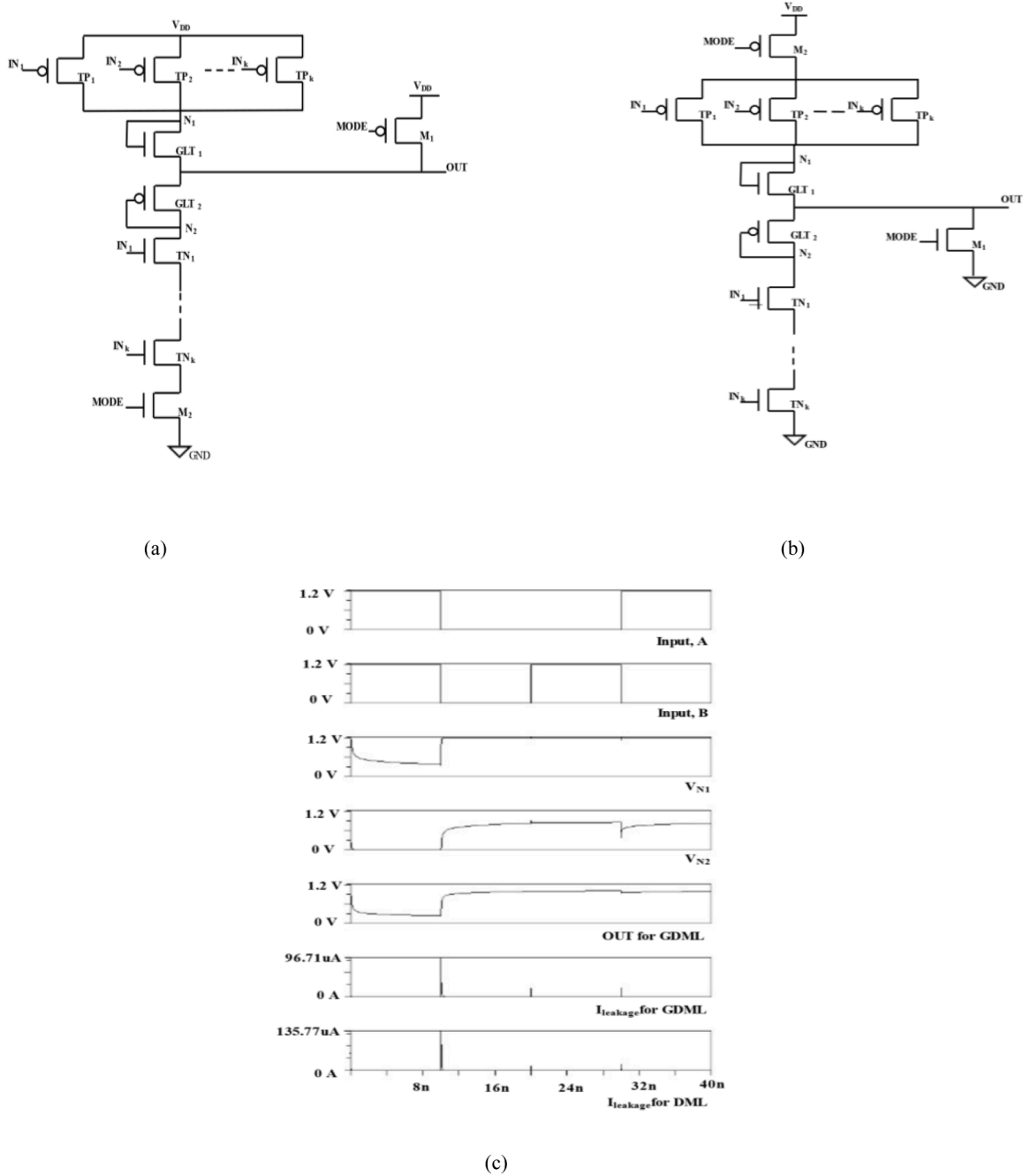


Fig. 4. GDML footed (a) type A NAND gate (b) type B NAND gate (c) timing diagram of proposed approach in 2-input NAND gate at 27 °C

obtained in SYMICA DE using PTM BSIM 90 nm technology parameters. The output in static mode is depicted in Fig. 2 (c) where the mode signal is kept at logic HIGH (LOW) for type A (type B). It may be observed that the output matches with the functionality of NAND gate. For dynamic mode, a clock signal is used as $MODE$ signal which permits two phases of operation -pre-charge (pre-discharge) and evaluation for type A (type B). The output node is charged to V_{DD} (discharged to ground) using M_{pre} (M_{eval}) transistor in type A (type B) topology in pre-charge (pre-discharge) phase. In evaluation phase of dynamic mode, the applied inputs decide the state of output in both type A and type B designs. The output in dynamic mode is placed in Fig. 2 (d) where a clock signal is applied so that evaluation in both type A and type B designs is performed simultaneously. It may be noted that the outputs for both the circuits are same in evaluation phase and conform with NAND functionality. Thus, DML, as its name suggests, has the capability of operating in two modes

by varying the $MODE$ signal in two different topologies [31].

2.2. Footed DML design

Footed DML gates are implemented by using footer transistor in type A and header transistor in type B along with the pre-charge and pre-discharge transistor as shown in Fig. 1(c-d). The footed design for type A consists of an additional NMOS transistor (M_{nfoot}), placed between Pull Down Network (PDN) and ground. Similarly, for type B, this structure includes an additional PMOS transistor (M_{pfoot}) between Pull Up Network (PUN) and supply voltage [40]. DML based unfooted type A NAND and type B NAND gates are shown in Fig. 3 (a-b). The working of footed type A NAND and type B NAND is illustrated through timing waveforms obtained in SYMICA DE using PTM BSIM 90 nm technology parameters. The output in static mode is depicted in Fig. 3(c) where the

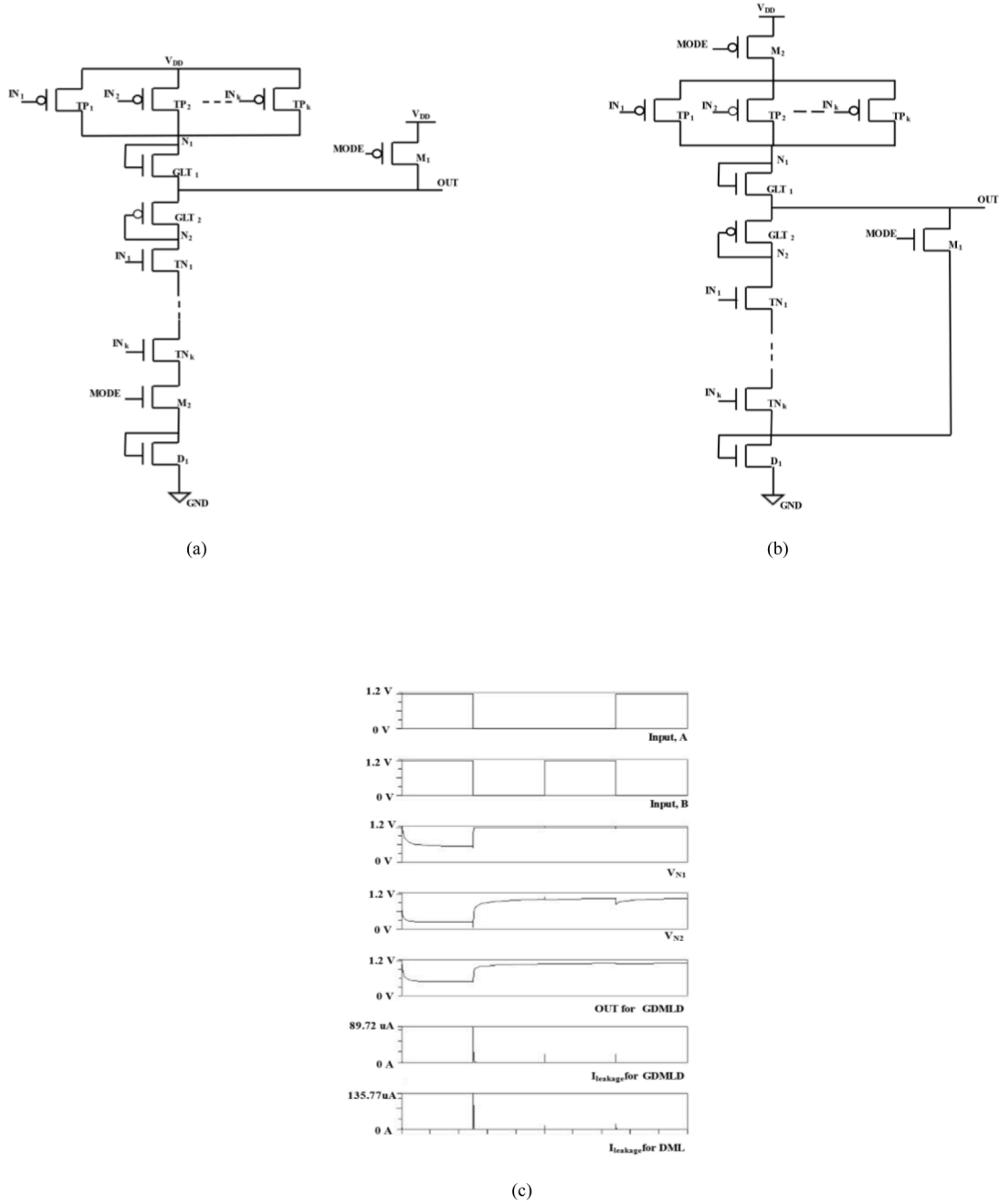


Fig. 5. GDMLD footed (a) type A NAND gate (b) type B NAND gate (c) timing diagram of proposed approach in 2-input NAND gate at 27 °C

mode signal is kept at logic HIGH (LOW) for type A (type B). The additional footer(header) transistor is turned on in type A (type B) in static mode. It may be observed that the output matches with the functionality of NAND gate. For dynamic mode, a clock signal is used as $MODE$ signal which permits two phases of operation -pre-charge (pre-discharge) and evaluation for type A (type B). The output node is charged to V_{DD} (discharged to ground) using M_{pre} (M_{eval}) transistor in type A (type B) topology in pre-charge (pre-discharge) phase. In evaluation phase of dynamic mode, the applied inputs decide the state of output in both type A and type B designs. The output in dynamic mode is placed in Fig. 3 (d) where a clock signal is applied so that evaluation in both type A and type B designs is performed simultaneously. It may be noted that the outputs for both the circuits are same in evaluation phase and conform with NAND functionality. Thus, DML, as its name suggests, has the capability of operating in two modes by varying the $MODE$ signal

in two different topologies [31].

2.3. DML leakage current analysis

This section investigates the presence of leakage current in footed DML designs by considering DML type A and type B NAND gate as shown in Fig. 3(a-b). In footed type A NAND gate, the pre-charge transistor (M_1) is in off for static mode. Depending on the inputs applied, some amount of leakage current exists in the off transistors and also in the off pre-charge transistor (M_1). This creates a path for a current to flow from supply to ground. Similarly, in pre-charge phase of dynamic mode, transistor M_1 is used to charge the output to V_{DD} irrespective of the inputs applied. Ideally there should not be any current flowing in PDN as transistor M_2 is off but due to leakage in off transistors, leakage current flows from supply to ground. During evaluation phase, the leakage

current exists in off transistors and is dependent on the inputs applied. Similar analysis can be done for footed type-B NAND gate in static and dynamic mode. The only difference is that in static and dynamic mode, additional leakage exists because of the pre-discharge transistor (M_1).

3. Proposed approach

In this section, leakage reduction GALEOR technique is proposed for designing DML footed circuits, termed as GDML. Further a diode footed transistor is incorporated in GDML design and is referred to as GDMLD.

3.1. GALEOR based DML Logic

The proposed GDML based footed type A NAND gate is illustrated in Fig. 4(a), here two gated leakage transistors (GLTs)- GLT1 and GLT2 are inserted between PUN and PDN. The drain and gate terminals of each GLT are connected. The output is obtained from the connected source terminals of the two GLTs, and a pre-charge transistor (M_1) is attached at the output node. These two transistors reduce leakage by introducing stacking effect in the design. When the MODE signal is HIGH (static mode), transistor M_1 is off and M_2 is on. The switching of GLT1 and GLT2 is dependent on the voltages at node N1 and N2. To elucidate the operation, the timing waveforms are shown in Fig. 4(c). When all inputs are HIGH, all PMOS transistors (TP_1 - TP_k) are turned off, resulting in low potential at node N1 (V_{N1}). This makes transistors GLT1 off which leads to an increase in the resistance in the path from supply to ground and eventually, leakage current reduces. If all the inputs are LOW, all the NMOS transistors except M_2 are turned off. Now the voltage at node N2 (V_{N2}) is closer to supply voltage which makes transistor GLT2 off and decreases leakage current.

Further, the leakage currents for DML and GDML are also shown in Fig. 4(c). The leakage current is observed as 0.37nA (5.06 nA) and 0.53 nA (11.72 nA) for all inputs low (high) for GDML and DML gates which confirms the proposition. The leakage current is dependent on the inputs applied. In pre-charge phase of dynamic mode (MODE=LOW), the output node is charged to HIGH level. The applied inputs don't affect the output but the leakage current is dependent on the inputs applied. When the MODE signal goes HIGH or during evaluation phase, the output node will be charged or discharged depending on the applied inputs. Consider when all inputs are LOW, the output node will be charged. One of the GLT (GLT2) would be in off state which would increase the resistance of supply to ground path. Similarly, for inputs HIGH case, the low voltage potential at node N_1 makes GLT1 off, as a result the number of OFF transistors from supply to ground is increased therefore more leakage reduction is achieved. The proposed GDML based type B NAND is shown in Fig. 4(b) and the leakage mechanism is same as that of GDML type-A NAND gate.

3.2. GALEOR based DML Logic with footed diode

GDMLD circuits are obtained by placing a footed diode transistor above the ground terminal in both type A and type B topology designs. The GDML circuit of Fig. 4(a-b) is modified by adding an NMOS diode transistor (D_1), with its gate and drain terminals connected above the ground terminal. The resulting type A and type B GDMLD circuits are shown in Fig. 5. This configuration results in further leakage reduction by introducing stacking effect [30]. Here the mechanism behind leakage reduction is same as in the case of GDML circuits of type A and type B topology, both in static and dynamic mode. Further, the leakage currents for DML and GDMLD are also shown in Fig. 5(c).

3.3. Equations for leakage power

Leakage power is prominently due to subthreshold current [2]. Leakage power, P_{leakage} due to subthreshold current is given by equation 1 [2].

Table 1

Delay values for 2-input NAND and NOR gates at 90 nm and 45 nm at 27 °C

Delay (ns), 90 NM				Delay (ns), 45 NM		
	DML	GDML	GDMLD	DML	GDML	GDMLD
TA-NAND-2	0.07	0.09	0.64	0.05	0.07	0.6
TA-NOR-2	0.09	0.12	0.67	0.07	0.09	0.61
TB-NAND-2	0.09	0.11	0.63	0.07	0.08	0.59
TB-NOR-2	0.11	0.14	0.67	0.08	0.1	0.61

$$P_{\text{leakage}} = I_{\text{subthreshold}} * V_{DD} \quad (1)$$

where V_{DD} is power supply and $I_{\text{subthreshold}}$ is leakage current given by equation 2 [2]

$$I = I_0 \exp \left((V_{gs} - V_t) / \alpha V_{th} \right) \quad (2)$$

where V_t and V_{th} correspond to the device threshold voltage and thermal voltage ($V_{th} = 25.9\text{mV}$ at room temperature); and I_0 is the current when $V_{gs} = V_t$. The parameter α is a constant depending on the device fabrication process, ranging from 1.0 to 2.5.

It can be observed from equation 2 that frequency and load capacitance do not affect the leakage current. Static power depends on leakage current [41]. Also, static power is independent of frequency [41]. Therefore, it can be inferred that leakage power is independent of frequency.

4. Simulation results

The circuits are simulated in a footed DML, proposed GDML and GDMLD circuits are simulated for 2-input NAND and NOR circuits using 45 nm and 90 nm PTM models for CMOS technology. Both type A and type B designs are compared in static and dynamic mode using SymicaDE tool at 1.2 V power supply and load capacitance of 5fF. High threshold voltage transistors are used as GLTs for both GDML and GDMLD circuits. SymSpice is the SPICE simulator to document the function of the proposed circuits. For power analysis, SymProbe tool is used. The (W/L) of all NMOS and PMOS except GLTs and footed diode transistor are kept at (120nm/90nm) and (120nm/45nm) for 90 nm and 45 nm technology nodes respectively. The width of the footed diode transistor is taken as 240nm.

To examine the effect of GLTs on voltage headroom in the proposed topologies, simulations are carried out by varying aspect ratios of GLT. It is observed that voltage headroom improves while the delay deteriorates therefore the aspect ratio of GLTs is taken as four times of their respective type. The proposed circuits are analysed for leakage power dissipation at 90 nm and 45nm. The total leakage power over all inputs for 2-input NAND and NOR gate in both topologies is considered. The average percentage power saving of proposed circuits is calculated by taking footed DML circuit as reference. This section consists of five subsections namely impact on delay, four subsections for comprehending average leakage power saving in static and dynamic mode of proposed DML designs, adder and multiplier implementation.

4.1. Impact on delay

To investigate the impact of proposed techniques on delay, the analysis is done for 2-input GDML and GDMLD NAND and NOR gates in static mode. The simulated delay values are enlisted in table 1 for 90 nm and 45nm. It may be observed that GDMLD variant has largest delay while GDML variants show a maximum increase of 25% in delay with respect to corresponding DML variants. However, there is improvement in delay if the width of footed diode transistor is increased. Similar results are obtained from simulations at 45 nm technology node. Further, the variation in delay is also observed by varying load capacitor at 90 nm and 45 nm as depicted in Fig. 6. It is found that the effect is more severe

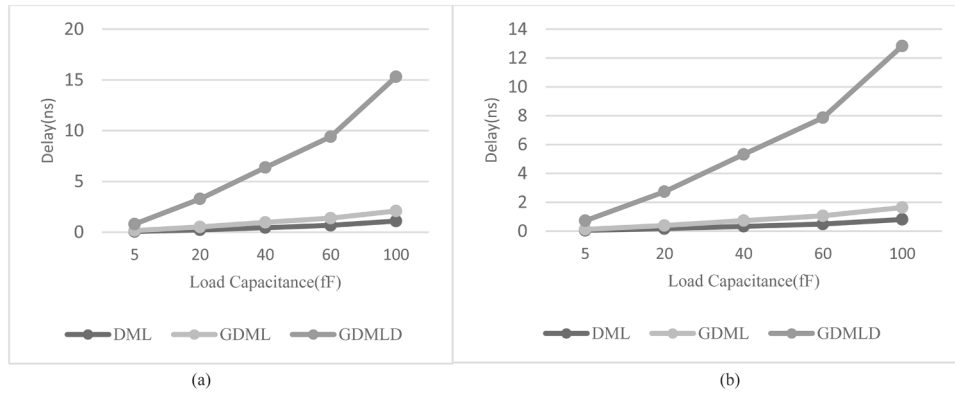


Fig. 6. Variation of delay with load capacitance for 2-input NAND type A gate (a) 90 nm (b) 45 nm at 27 °C

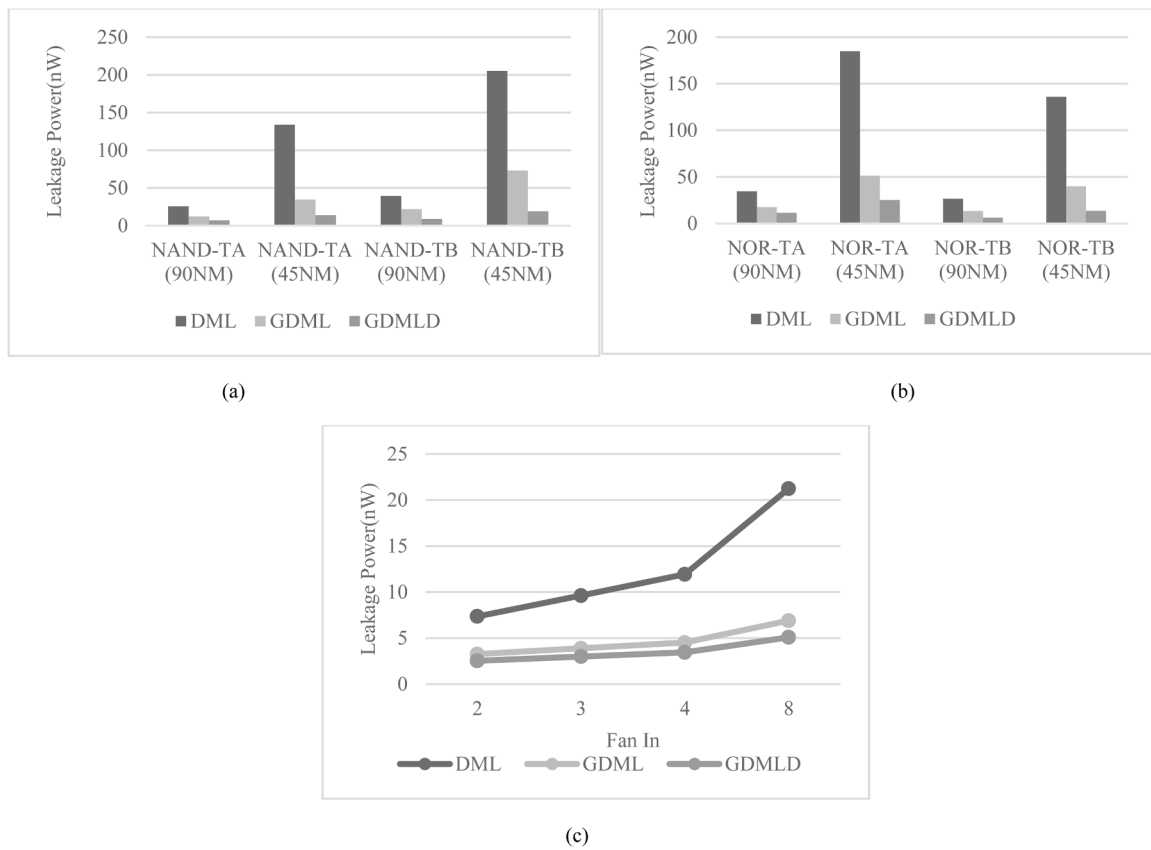


Fig. 7. (a) Total leakage power comparison for type A and type B NAND gate using DML, GDML and GDMLD technique in static mode at 90 nm and 45 nm at 27 °C (b) Total leakage power comparison for type A and type B NOR gate using DML, GDML and GDMLD technique in static mode at 90 nm and 45 nm at 27 °C (c) Average leakage power comparison for NAND type A in static mode for different fan in at 90 nm at 27 °C

in case of GDMLD as compared to GDML technique.

4.2. Static mode leakage power and leakage energy analysis

The total leakage power of the existing and proposed NAND and NOR DML circuits in static mode for 90 nm and 45 nm technology nodes at 27 °C is given in Fig. 7(a-b), it can be observed that the proposed GDML and GDMLD approach effectively reduces the total leakage power of standard footed DML circuits. Average leakage power variation with fan-in is also examined at 90 nm and 45 nm and observations for TA-NAND2 at 90 nm are summarized in Fig. 7(c). Further, the average leakage power is also examined for fan-in of 2,3,4 for NAND and NOR gates at 90 nm and 45 nm technology nodes and the trend is found to be

Table 2

Average percentage leakage power saving for two input GDML and GDMLD NAND and NOR type A and type B topology in static mode at 45 nm and 90 nm at 27 °C

	Average Percentage leakage power saving, Static Mode			
	GDML-TA-NAND2	GDML-TA-NOR2	GDML-TB-NAND2	GDML-TB-NOR2
90nm	51.97	49.56	44.69	50.1
45nm	74.11	70.67	64.29	72.24
	GDMLD-TA-NAND2			
	GDMLD-TA-NOR2	GDMLD-TB-NAND2	GDMLD-TB-NOR2	
90nm	73.38	67.15	77.97	76.59
45nm	89.69	86.46	90.76	90.08

Table 3

Leakage energy values for 2-input NAND and NOR gates at 90 nm and 45 nm at 27 °C

	Leakage Energy(aJ), 90 nm			Leakage Energy(aJ), 45 nm		
	DML	GDML	GDMLD	DML	GDML	GDMLD
TA-NAND-2	0.45	0.28	1.09	1.67	0.61	2.07
TA-NOR-2	0.77	0.52	1.9	3.59	1.65	2.89
TB-NAND-2	0.89	0.6	1.37	3.24	1.03	3.69
TB-NOR-2	0.73	0.46	1.04	2.72	1	2.06

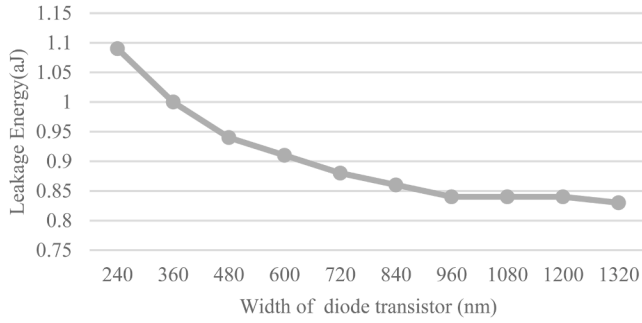


Fig. 8. Variation of leakage energy in GDMLD with footed diode transistor width variation at 27 °C.

similar to those observed for Fig. 7(c). Also, there is an increase in leakage power as technology is scaled i.e., from 90 nm to 45nm. Average percentage leakage power saving achieved by the proposed circuit techniques is enlisted in table 2. Maximum average power saving of 51.9 % is observed in TA-NAND2 circuit in GDML and 77.9% in TB-NAND2 in GDMLD at 90nm. At 45nm, this value is 74.1% for GDML-TA-NAND2 and 90.8% for GDMLD-TB-NAND2 circuit. It can be inferred that proposed GDMLD approach is more efficient than proposed GDML counterpart for leakage reduction. Further, the impact of footed diode transistor width on leakage power is also examined and it is observed that the leakage power saving reduces with increasing width.

Further the leakage power-delay-product (PDP) i.e., the leakage energy, is evaluated for 2-input NAND and NOR gates at 90 nm and 45 nm and is placed in the table 3. It is observed that there is significant decrease in leakage energy using proposed GDML approach; however, in proposed GDMLD approach, the value of leakage energy increases as compared to DML.

Further, the impact of footed diode transistor width on the leakage energy is examined and the observations are shown in Fig. 8. It is found that leakage energy reduces initially but becomes nearly constant after 840 nm.

4.3. Dynamic mode leakage power analysis

In dynamic mode, the leakage power of the proposed and existing DML technique in pre-charge and evaluation phase is observed by taking MODE signal frequency as 100 MHz and the results are depicted in Fig. 9-10 respectively. It shows that the total leakage power is reduced

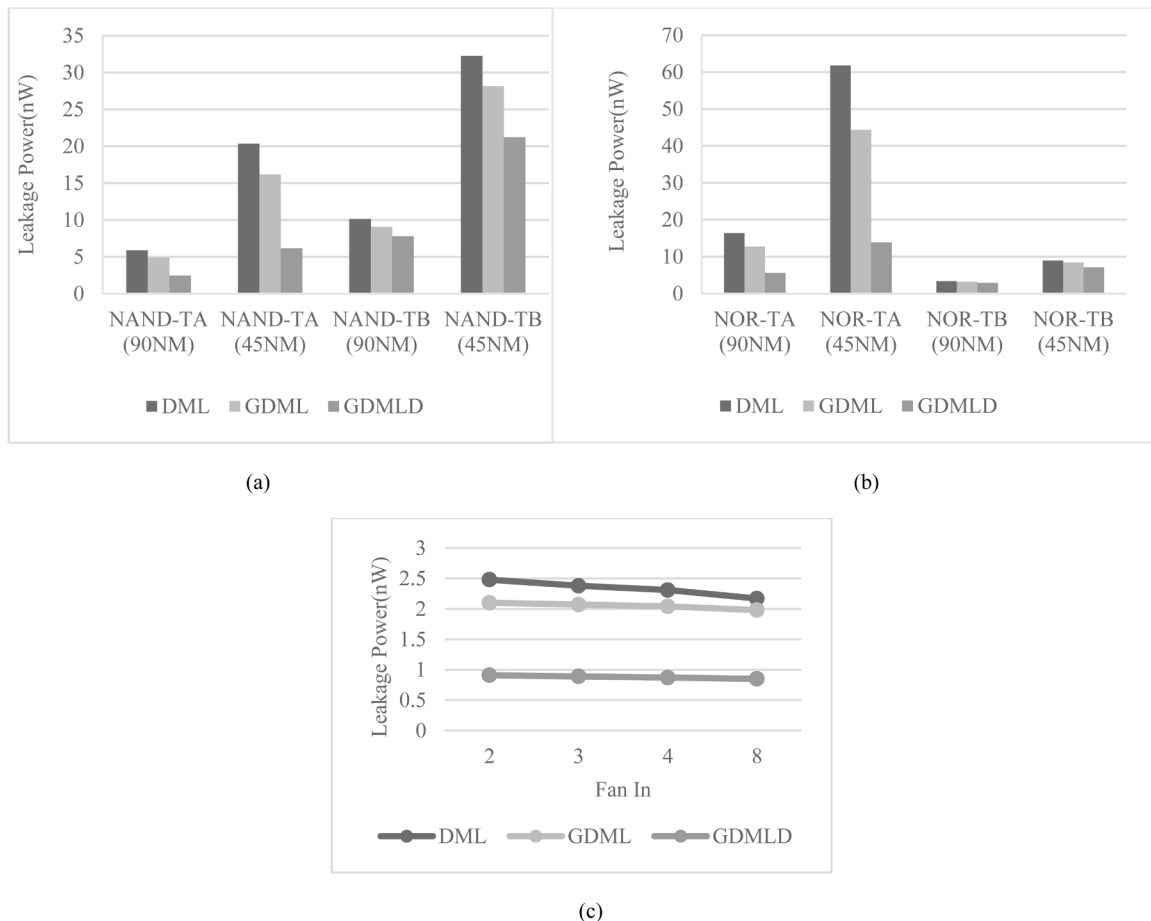


Fig. 9. (a) Total leakage power comparison for type A and type B NAND gate using DML, GDML and GDMLD technique in pre-charge phase at 90 nm and 45 nm at 27 °C (b) Total leakage power comparison for type A and type B NOR gate using DML, GDML and GDMLD technique in pre-charge phase at 90 nm and 45 nm at 27 °C (c) Average leakage power comparison for NAND type A in pre-charge phase for different fanin at 90 nm at 27 °C

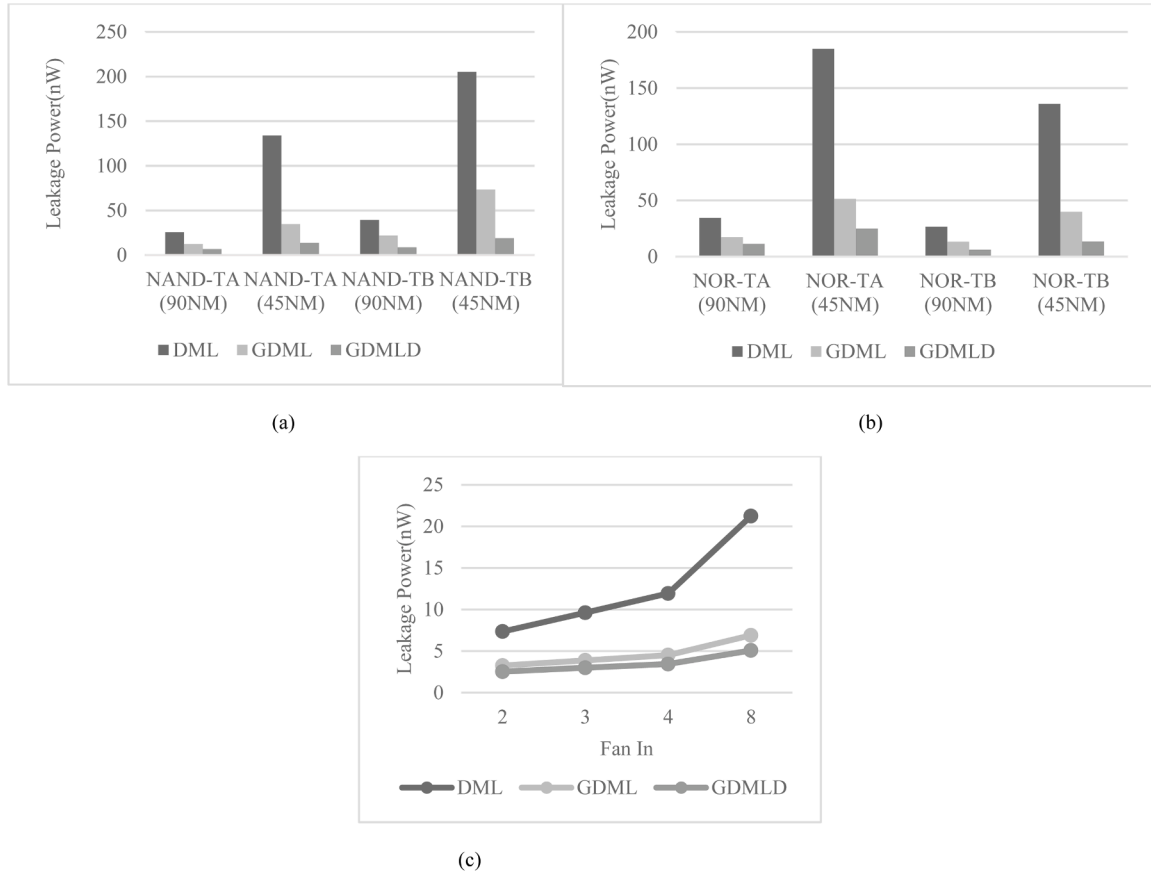


Fig. 10. (a) Total leakage power comparison for type A and type B NAND gate using DML, GDML and GDMLD technique in evaluation phase at 90 nm and 45 nm at 27°C (b) Total leakage power comparison for type A and type B NOR gate using DML, GDML and GDMLD technique in evaluation phase at 90 nm and 45 nm at 27°C (c) Average leakage power comparison for NAND type A in evaluation phase for different fanin at 90 nm at 27 °C

Table 4

Average percentage leakage power saving for two input GDML and GDMLD NAND and NOR type A and type B topology in dynamic mode at 45 and 90 nm at 27 °C

Average Percentage leakage Power Saving, Pre-charge				
	GDML-TA-NAND2	GDML-TA-NOR2	GDML-TB-NAND2	GDML-TB-NOR2
90nm	16.38	22.23	10.46	5.47
45nm	20.61	28.22	12.7	5.67
Average Percentage leakage Power Saving, Evaluation				
	GDML-TA-NAND2	GDML-TA-NOR2	GDML-TB-NAND2	GDML-TB-NOR2
90nm	51.96	49.6	44.69	50.1
45nm	74.11	70.67	64.28	72.24
	GDMLD-TA-NAND2	GDMLD-TA-NOR2	GDMLD-TB-NAND2	GDMLD-TB-NOR2
90nm	58.4	66.02	23.06	14.55
45nm	69.84	77.51	34.22	20.74
	GDMLD-TA-NAND2	GDMLD-TA-NOR2	GDMLD-TB-NAND2	GDMLD-TB-NOR2
90nm	73.38	67.18	77.97	76.59
45nm	89.69	86.46	90.76	90.08

with the incorporation of the proposed designs, both in pre-charge and evaluation phase. Total leakage power variation for NAND and NOR gates for pre-charge and evaluation phase is depicted in Fig. 9(a-b) and Fig. 10(a-b) respectively. Average leakage power variation with fan-in is also examined for pre-charge phase and evaluation phase at 90 nm and 45 nm and observations for TA-NAND2 in pre-charge and evaluation phase at 90 nm is summarized in Fig. 9(c) and Fig. 10(c) respectively.

Further, the average leakage power is also examined for fan-in of for

Table 5

Delay values for 2-input NAND type A gate at 90 nm in dynamic mode at 27 °C

Frequency (MHz)	Delay(nsec)		
	DML	GDML	GDMLD
1	0.03	0.07	0.65
10	0.03	0.07	0.67
100	0.03	0.07	0.67

2,3,4 for NAND and NOR gates at 90 nm and 45 nm technology nodes for both pre-charge and evaluation phase and the trend is found to be similar to those observed for Fig. 9(c) and Fig. 10(c) respectively. Table 4 enlists the average percentage leakage power saving offered by the GDML and GDMLD designs in dynamic mode. In both pre-charge and evaluation phase, an upward trend is observed in leakage power from 90 nm to 45nm. In pre-charge, the GDML approach offers a maximum average percentage saving of 22.2% in TA-NOR2 for 90 nm and 28.2% in TA-NOR2 for 45nm. Similar observations for GDMLD are 66% in TA-NOR2 for 90 nm and 77.5% in TA-NOR2 for 45nm. In evaluation phase, a comparison of average leakage power saving reveals that GDML and GDMLD offers maximum leakage power saving of 51.9% in TA-NAND2 and 77.9% in TB-NAND2 respectively at 90nm. At 45nm, this value increases to 74.1% in TA-NAND2 for GDML and 90.8 % in TB-NAND2 for GDMLD. It is analysed that the GDMLD approach is more adept at leakage power saving than GDML in dynamic mode.

Also, DML, GDML and GDMLD designs are simulated at different mode (clock) frequency of 1MHz, 10 MHz and 100 MHz for 2-input NAND-TA at 90 nm and the observations are enlisted in table 5. It can be observed that there is negligible change in delay value.

Table 6

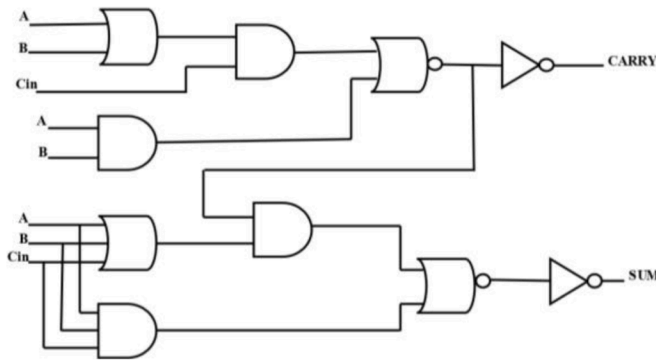
Leakage power and average power values for 2-input type A NAND in static mode for different load capacitance at 90 nm at 27 °C

Leakage Power(nW)						
Capacitance (fF)	5	20	40	60	80	100
DML	6.40	6.40	6.40	6.40	6.40	6.40
GDML	3.07	3.07	3.07	3.07	3.07	3.07
GDMLD	1.70	1.70	1.70	1.70	1.70	1.70
Average Power(uW)						
Capacitance (fF)	5	20	40	60	80	100
DML	6.96	9.66	13.24	16.56	19.21	21.09
GDML	4.2	5.31	6.35	6.97	7.26	7.29
GDMLD	1.59	1.94	2.95	3.57	4.31	4.51

Table 7

Leakage PDP and average PDP values for 2-input type A NAND in static mode for load capacitance of 100fF at 90 nm at 27 °C

	Leakage PDP (aJ)	Average PDP (fJ)
DML	7.04	23.2
GDML	3.04	7.22
GDMLD	22.52	59.62

**Fig. 11.** Schematic of a full-adder design [43]

4.4. Effect of load capacitance

The simulations are carried out to observe leakage and average power for different load capacitance(5fF-100fF) for a 2-input NAND type A gate in static mode at 90 nm and the findings are enlisted in [table 6](#). It may be observed that with increase in load capacitance value average power increases while leakage power remains constant. Further, there is significant reduction in both leakage and average power in proposed GDML and GDMLD designs as compared to DML design.

Also, the leakage PDP and average PDP for proposed designs and DML design at load capacitance of 100fF is enlisted in [table 7](#). Though the PDP is more for GDMLD design, the leakage minimization is prime concern in battery operated devices as it drains battery when the device is in idle state [42]. The proposed GDMLD design significantly reduces leakage power.

4.5. Full adder design

A full adder circuit is also implemented using the schematic of the [Fig. 11](#) [43]. This schematic is realized using DML and proposed GDML and GDMLD approaches. The sum block is implemented using type B topology and the carry block is implemented using type A topology. The simulated timing waveforms for sum, carry and inputs are shown in [Fig. 12](#), which match with the theoretical results of full adder circuit. The average leakage power for DML, GDML and GDMLD are enlisted in [table 8](#). As observed, the proposed approaches-GDML and GDMLD provides a maximum power saving of 26.91% and 44.59%.

4.6. 2-bit multiplier design

A 2-bit multiplier circuit with 2-bit inputs A and B and 4-bit output (O3, O2, O1, O0) is realized using DML and proposed GDML and GDMLD approaches. The simulated timing waveforms for 2-bit inputs at 90 nm are shown in [Fig. 13](#) which match with the theoretical results of 2-bit multiplier circuit. The average leakage power values for DML, GDML and GDMLD are enlisted in [table 9](#). As observed, the proposed approaches-GDML and GDMLD provides a maximum power saving of 62.3% and 65.33%.

A summary of observations based on the data enlisted in [tables 1-9](#) is as follows:

- There is a surge in the leakage power of the existing and proposed design with technology scaling i.e. from 90 nm to 45nm.
- The proposed GDML and GDMLD techniques show better performance in terms of power saving at lower technology node i.e., the percentage power saving is enhanced at 45 nm as compared with that of 90nm.
- The GDMLD approach proves to be more efficient at power saving in all circuits i.e. type A and type B NAND and NOR gates than the GDML approach.
- The reason behind better GDMLD performance in terms of power saving is mainly due to the presence of footed diode transistor which provides more stacking effect.
- Delay of the DML designs increase due to incorporation of GLTs in the proposed designs. The increase is more for GDMLD approach as compared to the GDML approach. Further, delay also increases as we increase the load capacitor value. The effect is more severe in case of GDMLD as compared to GDML technique.
- Leakage energy i.e., the product of delay and leakage power for 2-input DML, GDML and GDMLD gates (TA-NAND-2, TA-NOR-2, TB-NAND-2 and TB-NOR-2) is computed. It is observed that there is significant improvement in leakage energy using proposed GDML approach; however, in proposed GDMLD approach, the leakage energy deteriorates as compared to DML. However, the leakage energy in GDMLD improves with increasing footed diode transistor width.
- For a full adder and a 2-bit multiplier design, significant leakage power saving is witnessed using the proposed approaches. Similar to the smaller designs, the proposed GDMLD technique is more effective in combating leakage power as compared to GDML technique.

5. Conclusion

This paper has presented a novel GALEOR based leakage power reduction technique for footed DML circuits. The proposed designs use GLTs with and without a footed diode transistor, which efficiently reduces the leakage power in footed DML designs. The simulative investigations for type A and type B NAND and NOR circuits along with a full adder circuit at 90 nm and 45 nm shows that the leakage power increases at lower technology node, hence the efficiency of the proposed techniques also improves. The proposed novel GALEOR techniques exhibit impressive results for both type A and type B topology. In static mode, using GDML approach, the maximum average power saving of 74.1% and 70.7% is observed for type A and type B topology respectively. Corresponding values using GDMLD approach are (89.7%, 90.8%) for (type A, type B) topology. Similarly, in pre-charge phase, GDML achieves a maximum power saving of 28.2% and 12.7% for type A and type B respectively. Similar observations for GDMLD are (77.5%, 34.2%) for (type A, type B) topology. In evaluation phase, maximum average power saving of (74.1%, 70.7%) for (type A, type B) using GDML and (89.7%, 90.8%) for (type A, type B) topology using GDMLD approach. The proposed techniques prove to be effective in combating leakage power in a full adder and a 2-bit multiplier circuit also, with GDMLD technique providing more leakage power reduction than GDML technique. For proposed GDML approach, the leakage power

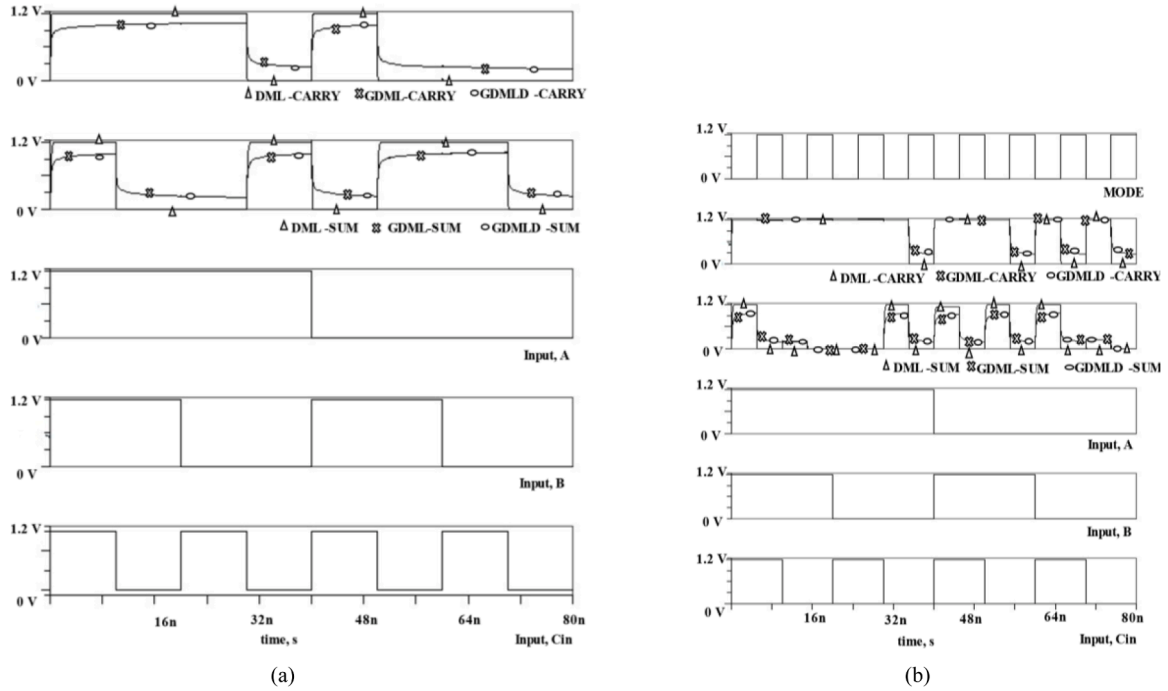


Fig. 12. Full Adder (a) timing waveform in static mode (b) timing waveform in dynamic mode

Table 8

Average leakage power for full adder circuit at 90 nm at 27 °C

Average Leakage Power(90nm)			
	DML	GDML	GDMLD
Static	35.5	25.94	19.67
Pre-charge	7.44	5.76	5.18
Evaluation	35.5	25.94	19.67

and leakage energy decrease significantly at the cost of deteriorated output voltage swing and delay. Proposed GDMLD approach offers significant leakage power reduction with increased leakage energy and

delay values and decreased output voltage swing. So, there is a trade-off between leakage power, leakage energy, output voltage swing and delay.

Table 9

Average leakage power for 2-bit multiplier circuit at 27 °C

Average Leakage Power (nW) (90nm)			
	DML	GDML	GDMLD
Static	215.47	81.23	74.71
Pre-charge	36.43	34.18	24.62
Evaluation	215.47	81.23	74.71

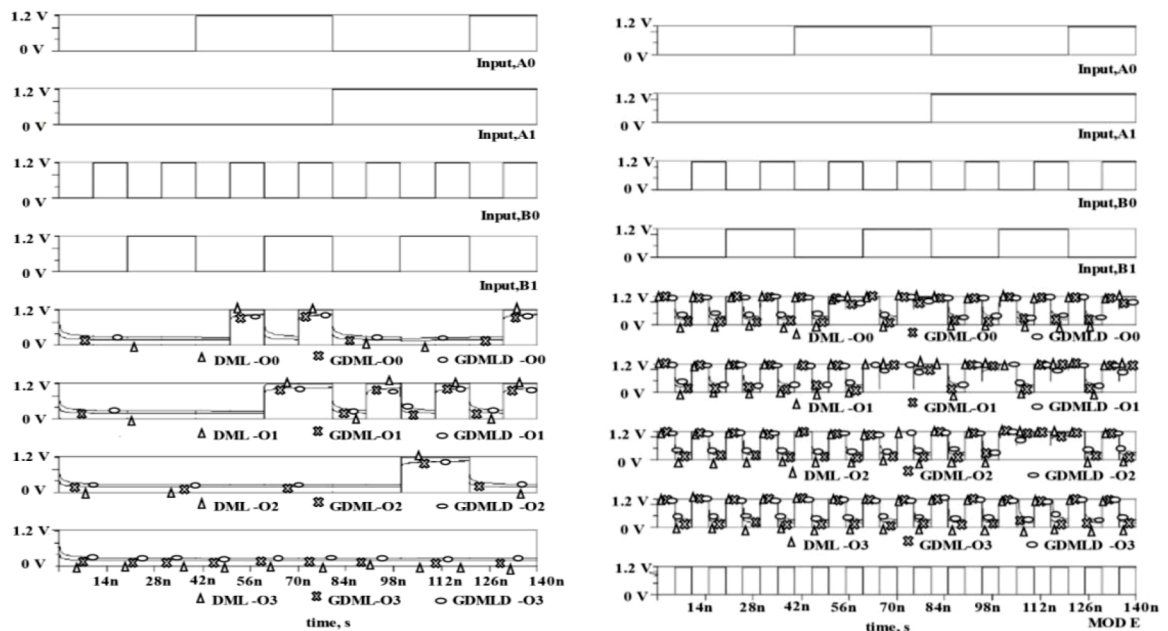


Fig. 13. 2-bit Multiplier (a) timing waveform in static mode (b) timing waveform in dynamic mode

Declaration of competing interest

The authors have no conflict of interest.

References

- [1] SG Narendra, AP. Chandrakasan, Leakage in Nanometer CMOS Technologies, Springer Science & Business Media, 2006, <https://doi.org/10.1007/0-387-28133-9>.
- [2] G.K. Yeap, Practical Low Power Digital VLSI Design, Springer Science & Business Media, 2012, <https://doi.org/10.1007/978-1-4615-6065-4>.
- [3] M Sanadhya, MV. Kumar, Recent development in efficient adiabatic logic circuits and power analysis with CMOS logic, Procedia Comput. Sci. 57 (2015) 1299–1307, <https://doi.org/10.1016/j.procs.2015.07.439>.
- [4] D Kumar, M. Kumar, Comparative analysis of adiabatic logic challenges for low power CMOS circuit designs, Microprocess. Microsyst. 60 (2018) 107–121, <https://doi.org/10.1016/j.micpro.2018.04.008>.
- [5] Gupta K, Gosain V, Pandey N. Adiabatic Differential Cascode Voltage Switch Logic (A-DCVSL) for low power applications. J. King Saud Univ.-Eng. Sci.. 2020. 10.1016/j.jksues.2020.09.018.
- [6] K Murugan, S. Baulkani, VLSI implementation of ultra power optimized adiabatic logic based full adder cell, Microprocess. Microsyst. 70 (2019) 15–20, <https://doi.org/10.1016/j.micpro.2019.07.001>.
- [7] HS Raghav, VA. Bartlett, Investigating the influence of adiabatic load on the 4-phase adiabatic system design, Integration 75 (2020) 150–157, <https://doi.org/10.1016/j.vlsi.2020.06.007>.
- [8] WC Athas, LJ Svensson, JG Koller, N Tzartzanis, EY. Chou, Low-power digital systems based on adiabatic-switching principles, IEEE Trans. Very Large Scale Integrat. (VLSI) Syst. 2 (4) (1994) 398–407, <https://doi.org/10.1109/92.335009>.
- [9] P. Teichmann, Adiabatic Logic: Future Trend and System Level Perspective, Springer Science & Business Media, 2011, <https://doi.org/10.1007/978-94-007-2345-0>.
- [10] Sundararajan V, Parhi KK. Synthesis of low power CMOS VLSI circuits using dual supply voltages. In Proceedings of the 36th annual ACM/IEEE Design Automation Conference. 1999: 72–75. 10.1109/DAC.1999.781234.
- [11] Jung SO, Kim KW, Kang SM. Low-swing clock domino logic incorporating dual supply and dual threshold voltages. In Proceedings of the 39th annual Design Automation Conference. 2002: 467–472. 10.1109/DAC.2002.1012670.
- [12] S.H. Kulkarni, D Sylvester, High performance level conversion for dual V/sub DD/ design, IEEE Trans. Very Large Scale Integr. (VLSI) Syst. 12 (9) (2004) 926–936, <https://doi.org/10.1109/TVLSI.2004.833667>.
- [13] S. Kulkarni, A. Srivastava, D. Sylvester, D. Blaauw, Power optimization using multiple supply voltages. Closing the Power Gap Between ASIC & Custom, Springer, 2007, https://doi.org/10.1007/978-0-387-68953-1_8.
- [14] SM. Sharroush, Analysis of the subthreshold CMOS logic inverter, Ain Shams Eng. J. 9 (4) (2018) 1001–1017, <https://doi.org/10.1016/j.asej.2016.05.005>.
- [15] MA Eldeeb, YH Ghallab, Y Ismail, H. Elghitani, Low-voltage subthreshold CMOS current mode circuits: design and applications, AEU-Int. J. Electron. Commun. 82 (2017) 251–264, <https://doi.org/10.1016/j.aue.2017.08.049>.
- [16] HS Raghav, S Maheshwari, BP. Singh, Performance analysis of subthreshold 32-bit kogge-stone adder for Worst-Case-Delay and power in sub-micron technology. VLSI Design and Test, 2013, pp. 100–107, https://doi.org/10.1007/978-3-642-42024-5_13.
- [17] J Kao, A. Chandrakasan, MTCMOS sequential circuits. Proceedings of the 27th European Solid-State Circuits Conference, IEEE, 2001, pp. 317–320.
- [18] JJ Johannah, R Korah, M. Kalavathy, Standby and dynamic power minimization using enhanced hybrid power gating structure for deep-submicron CMOS VLSI, Microelectron. J. 62 (2017) 137–145, <https://doi.org/10.1016/j.mejo.2017.02.003>.
- [19] A. Kabbani, Logical effort based dynamic power estimation and optimization of static CMOS circuits, Integration 43 (3) (2010) 279–288, <https://doi.org/10.1016/j.vlsi.2010.02.002>.
- [20] S Maheshwari, J Patel, SK Nirmalkar, A. Gupta, Logical effort based power-delay-product optimization. In 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2014, pp. 565–569, <https://doi.org/10.1109/ICACCI.2014.6968530>.
- [21] Maheshwari S, Raghav HS, Gupta A. Characterization of Logical Effort for Improved Delay. In VLSI Design and Test 2013;108–117. 10.1007/978-3-642-42024-5_14.
- [22] T. Nikoubin, Hybrid logical effort for hybrid logic style full adders in multistage structures, IEEE Trans. Very Large Scale Integrat. (VLSI) Syst. 27 (5) (2019) 1138–1147, <https://doi.org/10.1109/TVLSI.2018.2889833>.
- [23] HS Raghav, S Maheshwari, A. Gupta, A comparative analysis of power and delay optimise digital logic families for high performance system design, Int. J. Signal Imag. Syst. Eng. 7 (1) (2014) 12–20, <https://doi.org/10.1504/IJSISE.2014.057934>.
- [24] N Hanchate, N. Ranganathan, LECTOR: a technique for leakage reduction in CMOS circuits, IEEE Trans. Very Large Scale Integr. (VLSI) Syst 12 (2) (2004) 196–205, <https://doi.org/10.1109/TVLSI.2003.821547>.
- [25] TK Gupta, K. Khare, Lector with footed-diode inverter: a technique for leakage reduction in domino circuits, Circ. Syst. Signal Process 32 (6) (2013) 2707–2722, <https://doi.org/10.1007/s00034-013-9615-2>.
- [26] TK Gupta, K. Khare, A new dual-threshold technique for leakage reduction in 65 nm footerless domino circuits, Int. J. Comput. Appl. 61 (5) (2013) 14–20, <https://doi.org/10.5120/9923-4544>.
- [27] P Bajpai, N Pandey, K Gupta, J. Panda, LECTOR incorporated differential cascode voltage swing logic (L-DCVSL), Analog Integr. Circuit. Signal Process. 100 (1) (2019) 221–234, <https://doi.org/10.1007/s10470-019-01466-2>.
- [28] S Singhal, A. Mehra, A novel technique for static leakage reduction in 16 nm CMOS design, Int. J. Electron. Lett 7 (4) (2018) 434–447, <https://doi.org/10.1080/21681724.2018.1525765>.
- [29] R Lorenzo, S. Chaudhury, Review of Circuit Level Leakage Minimization Techniques in CMOS VLSI Circuits, IETE Tech. Rev. (Institution Electron Telecommun Eng India) 34 (2) (2017) 165–187, <https://doi.org/10.1080/02564602.2016.1162116>.
- [30] Katrue S, Kudithipudi D. GALEOR: Leakage reduction for CMOS circuits. 15th IEEE International Conference on Electronics, Circuits and Systems. 2008:574–577. 10.1109/ICECS.2008.4674918.
- [31] V Yuzhaninov, I Levi, A. Fish, Design flow and characterization methodology for dual mode logic, IEEE Access 3 (2016) 3089–3101, <https://doi.org/10.1109/ACCESS.2016.2514398>.
- [32] A Kaizerman, S Fisher, A. Fish, Subthreshold dual mode logic, IEEE Trans Very Large Scale Integr. Syst 21 (5) (2013) 979–983, <https://doi.org/10.1109/TVLSI.2012.2198678>.
- [33] CE Jose, B. Kousalya, Power reduction in CMOS sub-threshold dual mode logic circuits by power gating, IOSR J. VLSI Signal Process. 5 (2) (2015) 60–67.
- [34] PV Lakshmisree, MC. Raghu, Design of subthreshold DML logic gates with power gating techniques, Int. J. Res. Eng. Technol. 3 (4) (2014) 174–180, <https://doi.org/10.15623/ijret.2014.0304032>.
- [35] SN Singh, R. Madhu, Power analysis for CMOS based dual mode logic gates using power gating techniques, Int. J. Sci. Eng. Technol. Res. 4 (12) (2015) 4067–4072.
- [36] SS Patil, SS Pathak, RR Kathar, DS. Patil, Low power based dual mode logic gates using power gating technique, Int. Res. J. Eng. Technol. 4 (5) (2017) 1462–1467.
- [37] Bikki P, Karuppanan P. Low power and high performance multi-Vth dual mode logic design. 11th Int. Conf. Ind. Inf. Syst. 2016:463–468. 10.1109/ICIINF.2016.8262985.
- [38] I Levi, A. Fish, Dual mode logic—Design for energy efficiency and high performance, IEEE access 1 (2013) 258–265, <https://doi.org/10.1109/ACCESS.2013.2262015>.
- [39] TF Bogart, JS Beasley, G. Rico, Electronic Devices and Circuits, Pearson/Prentice Hall, New Jersey, 2004.
- [40] I Levi, A Kaizerman, A. Fish, Low voltage dual mode logic: Model analysis and parameter extraction, Microelectron. J. 44 (6) (2013) 553–560, <https://doi.org/10.1016/j.mejo.2013.03.005>.
- [41] M McCool, J Reinders, A. Robison, Structured Parallel Programming: Patterns for Efficient Computation, Elsevier, 2012.
- [42] Z Chen, M Johnson, L Wei, W. Roy, Estimation of standby leakage power in CMOS circuit considering accurate modeling of transistor stacks. Proceedings International Symposium on Low Power Electronics and Design, IEEE, 1998, pp. 239–244, <https://doi.org/10.1145/280756.280917>.
- [43] SM Kang, Y. Leblebici, CMOS Digital Integrated Circuits, Tata McGraw-Hill Education, 2003.



Neetika Yadav received her B.Tech in Electronics and Communication Engineering from Guru Gobind Singh Indraprastha University (GGSIU), Delhi, India and M.Tech in VLSI Design from Guru Gobind Singh Indraprastha University (GGSIU), Delhi, India. She is currently pursuing Ph.D. from Delhi Technological University (DTU), Delhi, India and working as assistant professor in department of ECE, Amity School of Engineering and Technology, affiliated to GGSIU, Amity University, Noida. Her area of research is low power VLSI.



Neeta Pandey received her M.E. in Microelectronics from Birla Institute of Technology and Sciences, Pilani and Ph.D. from Guru Gobind Singh Indraprastha University Delhi. She has served in Central Electronics Engineering Research Institute, Pilani, Indian Institute of Technology, Delhi, Priyadarshini College of Computer Science, Noida and Bharati Vidyapeeth's College of Engineering, Delhi in various capacities. At present, she is Professor in ECE Department, Delhi Technological University. A life member of ISTE, and senior member of IEEE, USA. She has published papers in international, national journals of repute and conferences. Her research interests are in analog and digital VLSI Design.



Deva Nand received his B.Tech., M.Tech. degrees in Electronic and Communication Engineering from Kurukshetra University, Kurukshetra, India. He received his Ph.D. from Delhi Technological University, Delhi, India. At present he is Assistant Professor in Department of ECE, DTU, Delhi, India. A life member of ISTE, member of IAENG and IEEE. He has published papers in International, National Journals of repute and conferences. His research interests include analog and digital VLSI design.

Mining Tourists' Opinions on Popular Indian Tourism Hotspots using Sentiment Analysis and Topic Modeling

Shefali Singh

Department of Information
Technology

Delhi Technological University
New Delhi, India

shefalisingh_2k17it113@dtu.ac.in

Tureen Chauhan

Department of Information
Technology

Delhi Technological University
New Delhi, India

tureenchauhan_2k17it121@dtu.ac.in

Vibhas Wahi

Department of Information
Technology

Delhi Technological University
New Delhi, India

vibhaswahi_2k17it125@dtu.ac.in

Priyanka Meel

Department of Information
Technology

Delhi Technological University
New Delhi, India

priyankameel@dtu.ac.in

Abstract— User-generated content is an exploration area of interest with regards to web 2.0. The development of social networks and community-based websites have changed the manner in which individuals utilize the Internet. It makes individuals no longer restricted to pursuing the data given by professional channels, but to making individual profiles, producing personalized content, or sharing photographs, recordings, blogs, and so forth. This sort of data comprises the current online user-generated content. With the continuous development of the travel industry, the quantity of online travel review websites has also increased. Indian Tourism is popular for its rich culture and diversity and hence Government of India has increased the number of new tourist destinations to expand their popularity and presence. Researchers have proposed various studies to increase tourism network using Big Data. Techniques of Sentiment Analysis along with Topic Modelling have been used to unearth patterns and observations from online reviews. This paper aims to mine reviews of 10 popular travel destinations in India. Using sentiment analysis technique, the proposed research work has explored the polarity of various reviews extracted from TripAdvisor. Data collection was done by using the web framework Scrapy to acquire more than 10,000 reviews for these destinations. This paper also analyzes the result of doing Topic Modeling on reviews for individual destinations. Results conclude that Joy is the most common emotion in all the visitor's experiences. Indian tourism decision quality can be improved by the help of the results from this study.

Keywords—Sentiment Analysis, Topic Modeling, Indian Tourist Destinations, Unsupervised Learning, AFINN lexicon.

I. INTRODUCTION

Online users at this point do not just investigate content from an online site, they additionally contribute to the online data without any geological limitations. Online user-generated content has become an information sharing tool for people who like to interact with others (Duan et al., 2013). This helps information to accumulate on the internet. The most

popular example is *TripAdvisor*, which is one of the largest travel websites helping travelers across the globe to plan an outing and offer their insight. According to *TripAdvisor*, they as of now, have around 70 million enlisted users with 630 million reviews and opinions posted on the site. There are a total of around 7.5 million hotels, restaurants and attractions listed on their website across around 136,000 destinations around the globe.

The proposed research work deals with TripAdvisor since it is considered as a superior approximation of public sentiment rather than regular web articles and web sites. The reaction on TripAdvisor is more brief and broader. The first technique utilized to mine the contents of TripAdvisor is Sentiment Analysis. Sentiment analysis turned into a mainstream research issue in 2001, this is because of rapid development and advancement of the Internet and the increasing popularity of the review-based sites (Pang & Lee, 2008). The major technology for sentiment analysis is classification and the process includes two errands – training and classifying. The training process is to train datasets and identify model parameters and the classifying process is used to examine documents, whether positive or negative. The cycle of sentiment analysis incorporates analyzing, handling, summing up, and constructing text with emotional terms.

The unsupervised learning-based sentiment analysis is utilized to consider traveler opinions expressed in online surveys on TripAdvisor. A Lexicon-based approach is used to summarize the polarities of single words or phrases based on a sentiment dictionary (Devika et al., 2016). The initial step focuses on finding whether a word has a positive or negative sentimental value. This step primarily relies upon the sentiment dictionary. Also, Topic Modeling is utilized as a method for performing text analysis in order to group all the related keywords together. Online reviews can help the international or local

traveler to make a decision about choosing a right travel destination (Gao et al., 2015) . We want to know what tourists think about their travelling experience in India as this information would be quite beneficial to the Indian tourism industry.

II. RELATED WORKS

NLP Techniques have been used to determine sentiments by using a Sentiment Analyzer that extracts sentiments automatically (Joshi & Tekchandani, 2016) . Sentiment Analysis have been done on platforms such as Twitter using various unsupervised learning approaches (Pandarachalil et al., 2015) . Researchers have also done Sentiment Analysis using some deep learning architectures owing to their high performance and increased speed (Yadav et al., 2020). Numerous approaches like lexicon-based, SVM, Naïve Bayes classifier have been used (Agarwal et al., 2015) . Movie reviews have been analyzed by doing sentiment analysis using machine-learning algorithms like regression tree and support vector regression (Hur et al., 2016) . Hybrid techniques have also been used with performance similar to machine-learning techniques while being enough reliable as the lexicon-based methods (Mudinas et al., 2012) . Researchers have also studied the associated rule mining methods to extract the features in order to compare various negative or positive reviews (Zhang et al., 2010) . Some researchers have also presented some methods to increase the accuracy of sentiment score based on applying aspect-based sentiment analysis to TripAdvisor (Farhadloo & Rolland, 2013) .

III. LITERATURE REVIEW

This section will examine some basic sentiment analysis and topic modeling approaches and general issues while using sentiment analysis with social sites, for example, TripAdvisor, Twitter, and so forth. The main goal of sentiment analysis is to find out a sentiment running throughout the text as per the characteristics of the language. (García et al., 2012) . So, for what reason is sentiment analysis essential to analyzing online social media? Social media is the most widely recognized tool that individuals use every day, in addition to being a data source as well which may give insights into marketing strategy and consumer service. We can benefit from this data by utilizing the correct sentiment analysis approach. First, we will briefly explain Sentiment Analysis and then its three methodologies – the lexicon-based methodology, machine learning approach, and rule-based methodology. Then, we will move to Topic Modeling and its popular approaches.

A. Sentiment Analysis

Sentiment analysis means technology utilization such as natural language processing or machine learning to directly analyze subjective perspectives, feelings, and sentiments. This paper utilizes sentiment analysis based on lexicons with a dictionary approach.

a. Lexicon-based Sentiment Analysis

Sentiment analysis based on lexicon approach comes from text analysis dependent on grammar rules (Amiri et al., 2015) . The strategy is moderately basic, and it principally relies upon what sentiment dictionary utilized. In the study of sentiment analysis, the content should be pre-processed, such as eliminating stop words. The purpose is to decrease the feature selection dimension, decrease the number of calculations, and improve the efficiency of the results. The stop words normally incorporate articles, prepositions, numerals, interjections, and so on, these words are commonly used in the English language and excluding them won't impact the result of the analysis. For instance, "a/an", "the" and "of/off" are the normal stop words utilized in English content which can be ignored.

Using this approach, it's harder to detect negative sentiment compared to positive sentiment because negative sentiments are often expressed using sarcasm. Also, results obtained can depend on the lexicon data used. The disadvantage of using a dictionary-based lexicon is that it is generic and it doesn't account for the specifics of any domain (Feldman, 2013) . If a new word shows up on the Internet and this word is absent from the lexicon data, it can't be measured by this method. Also, it is based on simple text classification and does not have enough emotional words (Collomb et al., 2013) .

b. Machine Learning Sentiment Analysis

This method is more accurate compared to the lexicon-based methodology, since dictionary matching could cause bigger blunders if the sentiment of the word is too high and hence cannot be captured by the dictionary. It doesn't have to dive into the terms, sentences and language structure like the lexicon-based methodology, as the machine learning approach simply calculates the emotional words in the content and gets their scores of emotional tendencies. The machine learning technique chooses a piece of the text to express positive and negative sentiments, individually and afterward, trains the content to get an emotion classifier. The last classification is to give the content a category of 0 or 1 i.e., probability value. For instance, we could say that "the positive probability of this content is 90%, and the negative probability is 10%".

One short-coming of the present sentiment classifiers is that they can't recognize what reviewers like or dislike about the item if they only judge the review on a positive or negative polarity. A reviewer may like the majority of the item but dislike a few of deformities. To overcome this, object recognition techniques need to be applied to identify the parts that a reviewer likes or dislikes.

c. Rule-based Sentiment Analysis

Rule-based sentiment analysis is an unsupervised machine learning technique. This method utilizes rule learning to extract features. Utilizing rule-based sentiment analysis can extract item features from a particular product review (Yang & Shih, 2012) . Furthermore, it is announced that product features are basically characteristics of a product or service and this can help in attracting possible purchasers, and furthermore it can be used to build up a product marketing strategy. Generally, Researchers combine both machine learning and rule-based ways to come up with an algorithm which has a better efficiency of computational results.

B. Topic Modeling

Topic Modeling is the process of extraction of the central idea that is expressed in a document. In this machine learning technique, analysis of data is done to identify the cluster of words which basically represent the central idea of the whole text. It is considered as an unsupervised machine learning technique because there is no requirement of a predefined list of topics and there is no training requirement as well, therefore it is an easy way to analyze the data. Topic Models are helpful for clustering of documents, putting together blocks of data.

a. Probabilistic Topic Modeling

This type of Topic Modeling is considered as a statistical technique for processing text documents. Here, the complete document is viewed as a combination of small number of topics. This approach views the documents as a sum of a few of topics. The document gets a probability score based on the constituting topics of that document. A popular example of this kind of approach is Latent Dirichlet Allocation (LDA).

Latent Dirichlet Allocation

Each document comprises of different words and each word additionally has different topics that it belongs to. Latent Dirichlet Allocation works on the principle of discovering the different topics that the document belongs to, using the constituting words in the document.

LDA forms its basis on the Bag of Words model. It has the assumption that words are exchangeable and thus does not take into account sentence structure. Also the number of topics need to be determined beforehand.

b. Matrix Factorization Topic Modeling

This approach applies processes from algebra to deconstruct a bigger matrix into some smaller matrices. A popular example is Non-negative Matrix Factorization (NMF).

Non-Negative Matrix Factorization

It consists of methods in which the conversion of a matrix V to two matrices W and H happens, with the property that none of the matrix have components in negation. Given a collection of documents, NMF identifies topics and simultaneously classifies the document among these different topics. NMF is a NP hard problem in general thus heuristics have to be used which provide solutions only in special cases.

IV. RESEARCH METHODOLOGY

This research uses a quantitative exploration technique. Here, we changed review information scratched from TripAdvisor into a usable format for Sentiment Analysis and Topic modeling. We should specify that TripAdvisor will have an alternate number of reviews dependent on various domains. For instance, similar review surveys on TripAdvisor India and TripAdvisor Australia will have different quantities of online reviews. For this research, the survey information we utilized is gathered from TripAdvisor India

(<https://www.tripadvisor.in/>). The strategy that was applied to this study was lexicon-based sentiment analysis. We utilize this sentiment analysis technique to quantify the sentiment in the review information. For topic modeling, we will be using Non-Negative Matrix Factorization Technique (NMF).

A. Sentiment Analysis

Sentiment analysis is the most popular application of NLP.

There are a lot of instructional courses and exercises available that are related to analyzing datasets on a wide range of topics such as movie reviews. The trivial part of sentiment analysis is to break down an assemblage of text for understanding the emotion derived from it. Commonly, we evaluate this assessment with a positive or negative worth, called polarity.

The general sentiment is construed as +ve, neutral, or -ve from the indication of the polarity score.

We utilize a lexicon-based sentiment analysis strategy in this study. The lexicon that we have used is AFINN Lexicon. This lexicon basically represents a bag of words with a value assigned to each word and a score given between -5 and 5. A score of +5 indicates an extremely positive word. The lexicon that we have used has one of the least complexities and is very popular lexicon. The latest rendition is AFINN-en-165.txt. It has 3000+ words with assigned polarities to each word. It is created and maintained by Finn Årup Nielsen.

We selected the 10 most popular attractions in India (based on the number of reviews on TripAdvisor). These were:

1. Bandra-Worli Sea Link, Mumbai
2. Taj Mahal, Agra
3. Gurudwara Bangla Sahib, New Delhi
4. Agra Fort, Agra
5. The Golden Temple, Amritsar
6. Siddhivinayak Temple, Mumbai
7. Qutub Minar, New Delhi
8. Swaminarayan Akshardham, New Delhi
9. Amber Fort, Jaipur
10. Mehrangarh Fort, Jodhpur

The following steps summarizes the dataset creation process.

1. We created a Python program using the web crawling framework Scrapy and Selenium for scraping the reviews available on TripAdvisor for these attractions.
2. TripAdvisor's website heavily uses JavaScript. At most 5 reviews are displayed on the page at a time. A user can page through these reviews using the pagination widget given at the end of the webpage. The 'Read more' link needs to be clicked in order to display the full review.
3. Since both the pagination widget and "Read more" link are JavaScript based we used the Selenium WebDriver to simulate user clicks. Selenium WebDriver accepts commands from the Selenium Client API (we used the Client API for Python).

4. After a page loads, first we click (we simulate the click using Selenium WebDriver) the “Read more” link. Then we extract the review and the stars the reviewer gave from each of the 5 reviews present on the page. To extract these reviews, we use Scrapy Selectors. Selectors are Scrapy’s own mechanism for extracting data from HTML webpages. Scrapy Selectors were used instead of BeautifulSoup because they are faster.
5. After we have extracted the data we need, we click (simulated using Selenium WebDriver) the ‘Next’ button present at the bottom of the page to load the next 5 reviews. We repeat steps 4, 5, and 6 till the time the ‘Next’ button is disabled i.e., we run out of reviews to scrape.
6. We scraped all the English language reviews available on TripAdvisor (for the above-mentioned tourist attractions). We got a total of 104,109 individual reviews.
7. The greatest number of reviews were available for Taj Mahal because it is one of the Seven Wonders of the World.

In our study, Sentiment analysis is performed using python. The required packages which were imported are pandas, matplotlib, nltk, afinn. After the dataset is imported, the next step was to clean the text reviews of the dataset. The cleaning process has steps like:

- Converting the text into all lower case
- Removing punctuation
- Removing double spacing
- Removing numbers from text reviews

After the dataset is cleaned, the word reviews are tokenized. Since, we don't have the comfort of a named training dataset, Henceforth, we utilized unsupervised strategies for predicting the sentiment by utilizing information bases, ontologies, information bases, and vocabularies that have point by point data, curated and arranged only for sentiment analysis. Hence, we used a lexicon. Lexicon is basically a vocabulary, these vocabularies have a list of positive and negative polar words with some score related to them and using different procedures like encompassing words, context, grammatical features, phrases, scores were allocated to each review for which we need to compute the sentiment. Two labels are given to each review

- *sentiment score*
- *sentiment category*

The AFINN object has a method called *score()*, which gets a sentence as information and returns a score as yield. The score might be either neutral, negative, or positive. We figure the score of a review, basically by adding all the scores of all the words of the review. Since the score of afinn shifts from -5 to 5, it appoints a score to each review from -5 to 5 where -5 suggests a very negative sentiment and +5 infers an

amazingly positive sentiment. The workflow of the process is given below

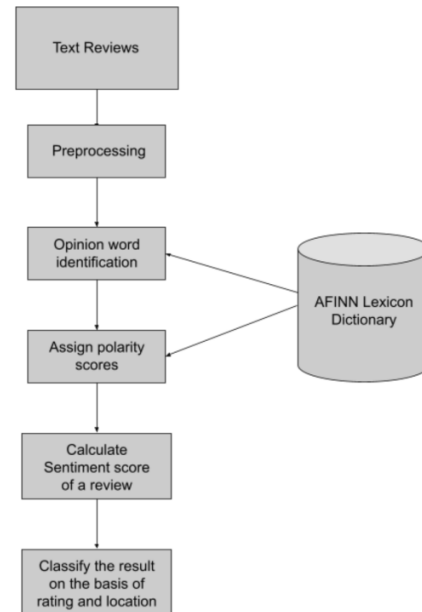


Fig 1: Workflow of sentiment analysis using AFINN lexicon

B. Topic Modeling

Topic Modeling is performed using the Non-Negative Matrix Factorization Technique (NMF). The first step in analyzing the unstructured documents is Tokenization which is splitting text into small tokens. The list of all the reviews is kept in a Document. Each document is represented as a term-vector. Each entry in the term vector represents the number of times that word appears in the review. The illustration is shown below:

Review: A great place to visit in summers, great views.

The corresponding term vector will be:

a	great	place	to	visit	in	summers	views
1	2	1	1	1	1	1	1

Fig 2: The term vector generated from the given review

Then each of these vectors are stacked and the stacking of vectors creates a document-term matrix (A). Further text processing is done which includes steps of minimum and maximum frequency filtering which removes the minimum and maximum frequency words, Stop-word filtering which removes the words that do not contribute to the final topic, etc.

The NMF methods takes the input of the Document-term Matrix (A) and the number of topics (k) and produces two matrices W and H. The H matrix represents the relation of words and the individual topics. The relation is depicted in the form of rows and columns where a row indicates a topic and

all the words in the dictionary are presented in the form of columns. The W matrix represents the document relation with all the topics in the dictionary. Here, the row represents all the documents and the column represents the topics.

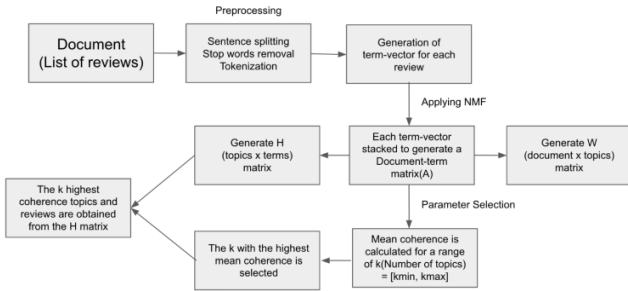


Fig. 3. Illustration of NMF

The next step is parameter selection, Topic Modeling is done by choosing a k (Number of topics) value. A common approach is to find the topic coherence of the results derived by taking different values of k and select the k with the highest coherence value. So, NMF is applied for a range of $k=[k_{\min}, k_{\max}]$ and the mean coherence is calculated for all the different values of k . The k for which the mean coherence has the highest value is selected as the best_ k . The further list of all the topics is extracted from the H matrix and it is stored in the k th (best_ k) row of the matrix. This final result contains k (best_ k) topics and all the terms that are present in those topics.

V. RESULT AND FINDINGS

A. Sentiment Analysis

• Classification on the basis of Star-rating

We will initially introduce our outcomes for the positive/negative classifications based on star ratings given by users alongside the content review. These outcomes go about as the initial step of our classification approach.

star_rating	sentiment_score							
	count	mean	std	min	25%	50%	75%	max
1	481.0	0.883576	5.589700	-28.0	-2.0	1.0	4.0	24.0
2	644.0	2.388199	5.341371	-19.0	-1.0	2.0	5.0	32.0
3	4099.0	3.568431	4.528193	-20.0	1.0	3.0	6.0	36.0
4	23097.0	4.294324	4.400643	-16.0	1.0	4.0	6.0	53.0
5	75787.0	4.277884	4.726368	-19.0	1.0	3.0	6.0	256.0

Fig. 4. Detailed relation between star rating and sentiment score

The above table indicates the sentimental score analysis of reviews differentiated on the basis of star rating. For example, if we consider the last row of the above table i.e., star rating is 5, we will see that the total count of the reviews with rating 5 are **75787**. The mean sentiment score of these ratings is **4.277884**. Similarly, the Minimum score is **-19**.

A more visualized and compact view of the above data is presented below which conveys the same information except all the metadata like min, max, mean.

As can be concluded from the below graph that the star rating 5 was the most assigned rating for all the attractions in the dataset

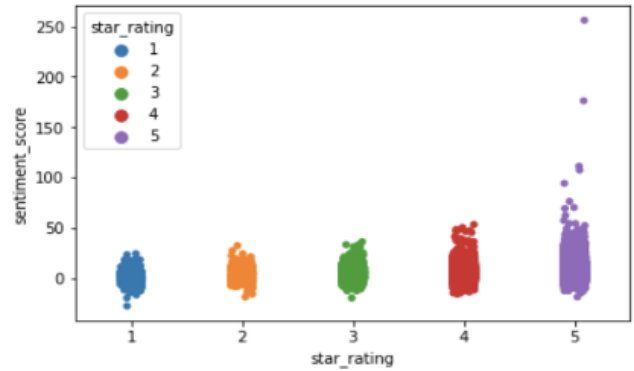


Fig. 5. Strip plot of star rating v/s sentiment score

The above shown graph is a result of the strip plot feature of matplotlib, the box plot of the same data is shown below.

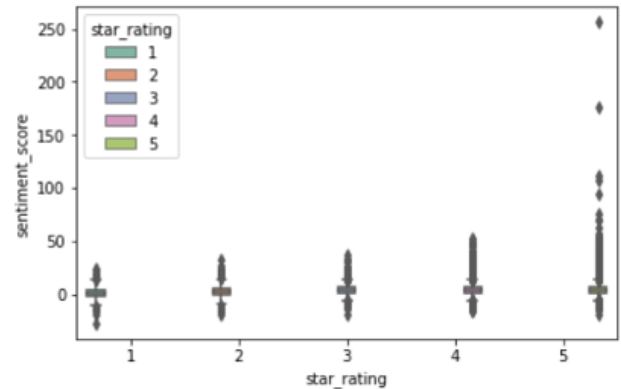


Fig. 6. Box plot of star rating v/s sentiment score

The above graphs show the classification of all the stars given by the user and plot the relationship of the star ratings and the sentiment score calculated by afinn lexicon.

Another plot which shows the relationship of the count of reviews and the star rating is shown below:

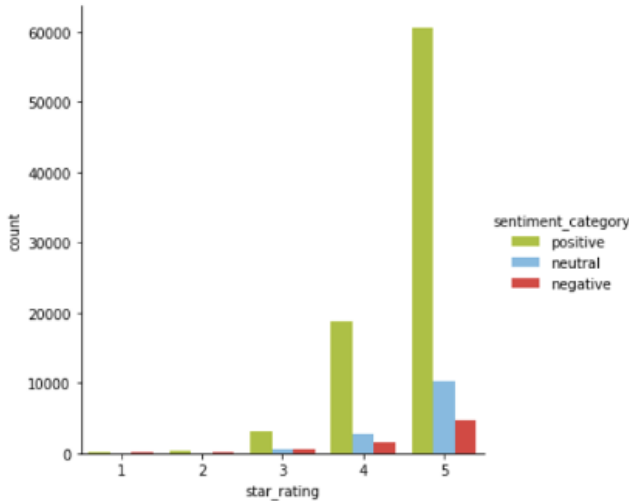


Fig. 7. Bar graph of star rating v/s count of reviews

- *Classification on the basis of location*

The next classification of the reviews is on the basis of the Location of the hotspot which is referred to as the column name “address” in the dataset.

address	sentiment_score							
	count	mean	std	min	25%	50%	75%	max
Agra Fort	8814.0	3.833220	4.026420	-17.0	1.0	3.0	6.0	46.0
Amber Fort	14677.0	4.381617	4.541754	-19.0	1.0	4.0	7.0	57.0
Bandra-Worli Sea Link	9307.0	3.804233	3.558049	-12.0	1.0	3.0	6.0	36.0
Golden Temple	8430.0	4.271293	4.909364	-13.0	1.0	3.0	6.0	176.0
Gurudwara Bangla Sahib	5714.0	4.248687	5.160908	-6.0	2.0	3.0	6.0	256.0
Mehrangarh Fort	6909.0	5.152120	4.655925	-13.0	2.0	4.0	7.0	47.0
Qutub Minar	10705.0	3.620645	3.888276	-15.0	1.0	3.0	6.0	40.0
Shree Siddhivinayak	5520.0	3.386413	3.704495	-9.0	1.0	3.0	5.0	37.0
Swaminarayan Akshardham	10099.0	4.948015	5.082501	-10.0	2.0	4.0	7.0	107.0
Taj Mahal	23933.0	4.311244	5.321624	-28.0	1.0	3.0	7.0	111.0

Fig. 8. Detailed relation between location of the hotspot and sentiment score

The above table indicates the analysis of sentimental score differentiated on the basis of address of location. For example, if we consider the first row of the above table i.e. hotspot is Agra Fort, we will see that the total count of the reviews with address “Agra Fort” are **8814**. The mean sentiment score of these reviews is **3.833220**. Similarly, the Minimum score is **-17**.

The highest number of reviews are given to the location “Taj Mahal” which depicts that it is the most visited location by tourists.

- *Finding the extreme reviews by the users*

By creating a positive index for the positive review for a particular location and doing the same for the negative review, we can find the extreme positive review for that location, for

example, if the location that we are considering is Agra Fort then the most positive and the most negative review is shown by the following piece of code:

```
print ('Most Negative Review for Agra Fort:',
yelp.iloc[neg_idx][['review']][0])
print ()
print ('Most Positive Review for Agra Fort:',
yelp.iloc[pos_idx][['review']][0])
```

Most Negative Review for Agra Fort: *Though the Agra Fort is a fabulous monument, the public are being cheated in the sound and light show, inasmuch as, there is neither sound nor light. Extremely boring and the dialogues can barely be heard. When we wished to complain, we were told that there is no one there to take responsibility, we have to contact Lucknow Tourism. The public was leaving the show within 15 minutes of the start of the show. The staff on premises admitted that the public was being cheated, but said that when they tried to pass on the complaints of the public to their higher-ups, they were asked to shut up as this is a major money-making machine. Please pass on this information so that people are not cheated of their precious time and money.*

Most Positive Review for Agra Fort: *A real day to remember visiting this part of India. Stunning buildings in a great setting with great views. A must see...*

B. Topic Modeling

- *Taj Mahal, Agra*

Topic 01: taj, mahal, visited, beauty, visiting, experience, view
Topic 02: marble, built, white, wife, shah, jahan, mumtaz, mughal
Topic 03: place, nice, good, great, wonderful, awesome, family, view
Topic 04: see, pictures, really, justice, life, magnificent, india
Topic 05: take, get, inside, ticket, gate, people, entrance, water

Fig. 9. Taj Mahal, Agra Topic

As can be seen from Fig. 8, the topics obtained are majorly revolving around the beauty of Taj Mahal. The initial most topic suggests the monument to be as beautiful as a dream. The second topic highlights the history of the monument and how Shah Jahan and Mumtaz are related to the history of Taj Mahal. The subsequent topics suggest that overall Taj Mahal is an amazing piece of architecture with a lot of positive topics.

- *Amber Fort, Jaipur*

Topic 01: elephant, ride, take, get, guide, fort, elephants, top, view
Topic 02: place, visit, history, jaipur, good, nice, guide, great, am
Topic 03: fort, amber, jaipur, palace, beautiful, amer, city, mahal,
Topic 04: show, light, sound, evening, history, night, hindi, english

Fig. 10. Amber Fort Topic

Based on the Fig. 9, the first topic suggests that the Amber Fort is famous of elephant rides and city views. The second topic consists of the historical relevance of the Amber Fort. The third and fourth topics are related to the beauty of the

monument and the things to do like light and sound show, palace visit, etc.

- The Golden Temple, Amritsar

Topic 01: temple, golden, amritsar, visited, visiting, beauty, gold, india, c
 Topic 02: place, holy, peaceful, world, religious, great, love, worship, spir
 Topic 03: sahib, harmandir, guru, sri, granth, darbar, gurudwara, ji, holy, k
 Topic 04: food, langar, free, served, guru, ka, eat, kitchen, community, deli
 Topic 05: visit, amritsar, try, worth, lifetime, complete, everyone, least, v
 Topic 06: inside, head, take, temple, shoes, complex, get, cover, queue, wate

Fig. 11. Golden Temple Topic

In Fig. 10, the topics are mainly related to the serenity, calmness and tranquil vibe of Golden Temple. The topics suggest that Golden Temple is the one of the most visited places in India and is one of the holy places that offer great religious significance and bring a divine feeling to their visitors. The first, second and third topics reveal the same. The fourth topic highlights the free food which is known as langar which is enjoyed equally by the rich and the poor. The subsequent topics reveal the neat and well-maintained architecture and the management and hospitality which is famous world-wide.

- Qutub Minar, New Delhi

Topic 01: delhi, monument, historical, visiting, best, maintained, sout
 Topic 02: tower, built, hindu, mosque, ruins, impressive, temples, ston
 Topic 03: guide, interesting, tour, site, take, really, worth, get, aud
 Topic 04: minar, qutub, qutab, complex, visited, qutb, aibak, monuments

Fig. 12. Qutub Minar Topic

Based on the Fig. 11, the major topics related to the monument Qutub Minar are related to the historical relevance of the beautiful piece of architecture and a great place for family picnics. The initial topics suggest that the monument has solid great historical background and is a complex architectural piece with rustic charm. The subsequent topics reveal that Qutub Minar is used as a picnic spot by many families in the city as it has good food around along with a rich history.

- Gurudwara Bangla Sahib, New Delhi

Topic 01: place, religious, holy, great, worship, people, clean, amazing,
 Topic 02: temple, sikh, people, kitchen, see, day, religion, free, amazin
 Topic 03: sahib, gurudwara, bangla, visiting, inside, visited, love, gold

Fig. 13. Gurudwara Bangla Sahib Topic

From the Fig. 12, the results are related to the religious, calm and holy atmosphere at Gurudwara Bangla Sahib. The topics suggest that Bangla Sahib is the one of the most visited places in India and is known for the peace obtained in visiting the holy place. The fifth topic highlights the free food which is known as langar which is enjoyed equally by the rich and the poor. ther topics reveal the neat and well-maintained architecture and the management and hospitality which is famous world-wide.

VI. PERFORMANCE EVALUATION

This performance is evaluated for the AFINN lexicon. The performance evaluation consists of taking the data set for the polarity classification task, then performing the classification task by applying lexicon. The original dataset extracted from *tripadvisor* contains columns like “reviews”, “star_ratings”, “location”. In order to calculate the accuracy of the lexicon, there should be a review sentiment available for each review to compare with the sentiment extracted from the lexicon, hence this base sentiment is calculated by the “star_ratings” given by the user. If the rating is greater than or equal to 3(≥ 3) then the sentiment is “positive” otherwise “negative”. This list of sentiments is further compared with the sentiments extracted from AFINN and hence the accuracy is calculated.

A. Evaluation Criterion

We have used the following measures for evaluation of classification parameters:

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

Fraction predicted correctly

Where, TP = the number of positive values classified properly,

TN = the number of values negatively classified properly,

FP = the number of positive values classified incorrectly,

FN = the number of negative values classified incorrectly,

This kind of accuracy can be obtained by scikit-learn which will take both the calculated sentiments as well as the predicted sentiments from the AFINN lexicon. The function used for calculating the accuracy is “accuracy_score”. A small implementation is given below:

accuracy = accuracy_score(actual_sentiments, predicted_sentiments)

VII. CONCLUSION

The motive of Sentiment Analysis is to unravel emotions portrayed by travelers by means of text relating to their encounters and experiences. We have focused on a few central points of interest in sentiment analysis, including the categorization of emotional information, and the retrieval of sentiment data. A portion of the reviews are posted by acclaimed bloggers or experienced experts, who are notable and have a high standing in the industry. When these people post an online review about a specific place, it is recognized as a genuine review by other people. This makes people more willing to go to places which have more positive reviews. The investigation shows that if an online review contains positive data as well as the commentator's identity, it can help increase sales.

To summarize, the study of text sentiment analysis has produces various kinds of emotions. 'Joy' emotion is an indicator of tourist satisfaction. While on the other hand, some reviews also consisted of 'Surprised' and 'Sadness' emotions. These emotions conclude that the services provided at some destinations needs to be improved. For the emotion of "surprise", some tourists claim that they were quite surprised by the amount of entry passes that was required for the destinations. For the sad emotions, tourists felt disappointed by the garbage that was carelessly thrown and was diminishing the beauty of those destinations. These results can be utilized to improve the beauty of the destination and the quality of service.

Topic modeling successfully categorized popular topics in all the locations taken in this study. The outcomes are diverse for each destination. The most discussed points are the views of the location, services provided, and activities. The topics obtained with respect to the destinations can be utilized as data to assess the tourist locations and furthermore discover what points are intriguing.

VIII. LIMITATIONS AND FUTURE WORKS

This study just uses TripAdvisor as a survey asset to research how travelers express their feelings through reviews. This study could likewise gather information from other famous travel sites, such as Booking (www.booking.com), Expedia (www.expedia.co.in), Goibibo (www.goibibo.com) and so forth. Secondly, when we manage sentiment analysis, we ought to likewise know about various implications of sarcasm words, and a sentence describing realities however with no sentiment words. Furthermore, computers, in contrast to people, experience a difficulty in dealing with sarcastic words or sentences, and it will diminish the accuracy of sentiment analysis results. It could be that some positive words have been identified in the review; however, the genuine review is negative.

Thirdly, this study could utilize diverse sentiment analysis strategies (such as through questionnaires or analysis of numerical scores) to explore sentiment expression in TripAdvisor.

REFERENCES

- [1] Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., & Hussain, A. (2015). Concept-Level Sentiment Analysis with Dependency-Based Semantic Parsing: A Novel Approach. *Cognitive Computation*. <https://doi.org/10.1007/s12559-014-9316-6>
- [2] Amiri, F., Scerri, S., & Khodashahi, M. H. (2015). Lexicon-based sentiment analysis for Persian text. *International Conference Recent Advances in Natural Language Processing, RANLP*.
- [3] Collomb, A., Costea, C., Joyeux, D., Hasan, O., & Brunie, L. (2013). A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation. In *Research report RR-LIRIS-2014-002*.
- [4] Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment Analysis: A Comparative Study on Different Approaches. *Procedia Computer Science*. <https://doi.org/10.1016/j.procs.2016.05.124>
- [5] Duan, W., Cao, Q., Yu, Y., & Levy, S. (2013). Mining online user-generated content: Using sentiment analysis technique to study hotel service quality. *Proceedings of the Annual Hawaii International Conference on System Sciences*. <https://doi.org/10.1109/HICSS.2013.400>
- [6] Farhadloo, M., & Rolland, E. (2013). Multi-class sentiment analysis with clustering and score representation. *Proceedings - IEEE 13th International Conference on Data Mining Workshops, ICDMW 2013*. <https://doi.org/10.1109/ICDMW.2013.63>
- [7] Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*. <https://doi.org/10.1145/2436256.2436274>
- [8] Gao, S., Hao, J., & Fu, Y. (2015). The application and comparison of web services for sentiment analysis in tourism. *2015 12th International Conference on Service Systems and Service Management, ICSSSM 2015*. <https://doi.org/10.1109/ICSSSM.2015.7170341>
- [9] García, A., Gaines, S., & Linaza, M. T. (2012). A Lexicon based sentiment analysis retrieval system for tourism domain. *E-Review of Tourism Research*.
- [10] Hur, M., Kang, P., & Cho, S. (2016). Box-office forecasting based on sentiments of movie reviews and Independent subspace method. *Information Sciences*. <https://doi.org/10.1016/j.ins.2016.08.027>
- [11] Joshi, R., & Tekchandani, R. (2016). Comparative analysis of twitter data using supervised classifiers. *Proceedings of the International Conference on Inventive Computation Technologies, ICICT 2016*. <https://doi.org/10.1109/INVENTIVE.2016.7830089>
- [12] Mudinas, A., Zhang, D., & Levene, M. (2012). Combining lexicon and learning based approaches for concept-level sentiment analysis. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. <https://doi.org/10.1145/2346676.2346681>
- [13] Pandarachalil, R., Sendhilkumar, S., & Mahalakshmi, G. S. (2015). Twitter Sentiment Analysis for Large-Scale Data: An Unsupervised Approach. *Cognitive Computation*. <https://doi.org/10.1007/s12559-014-9310-z>
- [14] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*. <https://doi.org/10.1561/15000000011>
- [15] Yang, C. S., & Shih, H. P. (2012). A rule-based approach for effective sentiment analysis. *Proceedings - Pacific Asia Conference on Information Systems, PACIS 2012*.
- [16] Zhang, L., Lim, S. H., Liu, B., & O'Brien-Strain, E. (2010). Extracting and ranking product features in opinion documents. *Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference*.
- [17] Mitra, Ayushi. "Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset)." *Journal of Ubiquitous Computing and Communication Technologies (UCCT)* 2, no. 03 (2020): 145-152.
- [18] Yadav, A., Vishwakarma, D.K. Sentiment analysis using deep learning architectures: a review. *Artif Intell Rev* 53, 4335-4385 (2020). <https://doi.org/10.1007/s10462-019-09794-5>.

Modeling and Analysis of High-Performance Triple Hole Block Layer Organic LED Based Light Sensor for Detection of Ovarian Cancer

Shubham Negi, *Member, IEEE*, Poornima Mittal[✉], *Member, IEEE*, and Brijesh Kumar, *Member, IEEE*

Abstract—In this paper a novel triple hole block layer (HBL) structure of the OLED is proposed that depicts an enhanced luminescence of 25285 cd/m^2 with an improvement of 47% over multilayered OLED architecture. It also owes 74% improvement in luminous power efficiency. An in-depth numerical analysis based on Poisson and drift diffusion equation is undertaken and validated against the internal device analysis. The analysis results highlight an enhanced recombination rate within the proposed device. High electron injection and efficient hole blocking contributes to improved recombination rate. Triple HBL OLED is therefore used for diagnosis of ovarian cancer. The device illustrated good response towards varying wavelengths generating a maximum photo current value of 93 mA. A healthy person can be differentiated from an oncological cancer patient based on fluorescence produced by their urine. The fluorescence values for healthy person and oncological cancer patient are in the range of 420 and 440 nm, correspondingly. The cathode current produced by OLED corresponding to these two wavelengths are 5 and 1 mA respectively. Hence, the proposed device can successfully diagnose the ovarian cancer patient. Further, the methodology proposed for diagnosis of ovarian cancer can help in developing a portable, flexible low cost biomedical sensor.

Index Terms—Biomedical sensor, hole block layer (HBL), Langevin's recombination rate, organic light emitting diode (OLED), organic semiconductors (OSC), ovarian cancer.

I. INTRODUCTION

ORGANIC light emitting diode (OLED) is a highly developed device technology in the field of display applications. In the previous decade these devices have shown a significant performance improvement and as a result present day displays are dominated by OLED. Organic LED based displays are preferred due to their superior color quality owing to a wide color variation in a limited spectrum [1]. Furthermore, contrast is highly improved as ambient lighting is not required [2]. These devices are cost effective due to low

temperature fabrication processes [3]: inkjet printing [4], [5], spin coating [6], screen printing [7], etc. Additionally, these processes facilitate fabrication over unconventional substrate (plastic, paper, etc.) resulting in flexible devices [2], [3]. However, OLED has multiple utilization and these devices are actively used for applications such as sensors [8], [9], imaging [10], and visual light communication (VLC) [11]–[13] as well.

Yet these fields have not illustrated a growth similar to OLED display. Hence, application specific OLEDs need to be developed. This is possible through device architectural changes [14] and material development. Therefore, the article presents a novel triple hole block layer (HBL) architecture for the OLED and its utilization as a light detector for diagnosis of ovarian cancer. The novel architecture results in an improved device performance owing to higher charge carrier injection and an enhanced recombination. Consequently, an ameliorated luminescence, current density, and efficiency are observed.

The article is divided in seven sections that include this introduction as a part of section I. Section II discusses the experimental setup and the models utilized. Thereafter, in section III, triple HBL architecture is discussed along with its analysis results. Its internal analysis utilizing the cutline analysis are undertaken in section IV followed by analytical analysis using Poisson's and drift diffusion equation in section V. Section VI highlights the role of novel OLED architecture in diagnosis of ovarian cancer, depicting its utilization as light detector. Finally, Section VII concludes the paper with discussion of important results.

II. EXPERIMENTAL SETUP

The analysis of the novel OLED architecture is conducted using ATLAS 2-dimensional device simulator. It incorporates inbuilt models and numerical equations for depth examination of the internal physics of the device [15]. Even new materials can be used by defining their properties. The complete device dimensions are defined and the individual regions are identified by specifying their properties: energy levels, work-function, carrier concentration, etc. Depending upon the biasing conditions different parameters can be obtained: current density, electric field, electron and hole concentration, recombination rate, etc. The model applied for the examining the organic device are discussed briefly next.

Manuscript received October 13, 2020; revised January 12, 2021 and March 21, 2021; accepted May 5, 2021. This article was recommended by Associate Editor L. Ye. (Corresponding author: Poornima Mittal.)

Shubham Negi is with the Department of Electronics and Communication Engineering, Graphic Era (Deemed to be University), Dehradun 248002, India.

Poornima Mittal is with the Department of Electronics and Communication, Delhi Technological University, New Delhi 110042, India (e-mail: poornimamittal@dtu.ac.in).

Brijesh Kumar is with the Department of Electronics and Communication Engineering, Madan Mohan Malaviya University of Technology, Gorakhpur 273010, India.

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TCSI.2021.3078510>.

Digital Object Identifier 10.1109/TCSI.2021.3078510

1549-8328 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.

See <https://www.ieee.org/publications/rights/index.html> for more information.

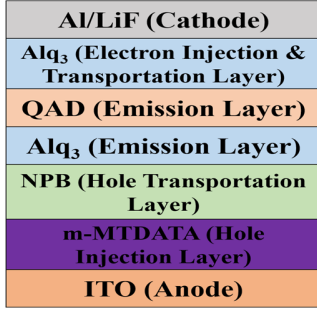


Fig. 1. Structure of multilayered OLED, Device A.

A. Poole Frenkel Mobility Model

Organic semiconductor (OSC) exhibits the hopping carrier transport [12], [13]. As a result, mobility and electric field are modeled through Poole Frenkel mobility model illustrated as:

$$\mu(E) = \mu_0 \exp \left[-\frac{\Delta}{KT} + \left(\frac{\beta}{KT} - \alpha \right) \sqrt{E} \right] \quad (1)$$

The various parameters in (1) are: $\mu(E)$ is field related mobility; μ_0 is zero biased field mobility; Δ the activation energy for zero bias; α - parameter for curve fitting; constants K represents the Boltzmann constant in J/K; whereas T is temperature in K; E is external electric potential applied at OLED and β stands for the Poole Frenkel factor. Further, β is calculated as (where all parameters have standard meaning):

$$\beta = q \sqrt{q / \pi \epsilon_r \epsilon_0} \quad (2)$$

B. Langevin's Recombination Model

Langevin's recombination model is best suited for investigation of recombination process related to OSC owing to their lower mobility [12], [13]. Langevin's model is specifically developed to govern the recombination process in these semiconductors and is defined as

$$R_L(n, p) = r_l(x, y, t) (np - n_i^2) \quad (3)$$

Referring to (3), r_l is coefficient for recombination rate in Langevin's form; and n_i is carrier concentration (intrinsic). The electron and hole concentration are correspondingly, represented by p and n . Further, recombination rate in Langevin's form is obtained from following:

$$r_l(x, y, t) = \frac{q\mu(E)}{\epsilon_r \epsilon_0} \quad (4)$$

These models are validated against the reported fabrication results for the multilayered OLED by Yang *et al.* [16] and the results are tabulated in Table I. Its structure is depicted in Fig. 1.

Table I illustrates that the maximum value of both the parameters for reported experimental and simulated device are close to each other with a relatively low error percentage. The calculated error percentage for current density is 3.05% and that for luminescence is 1.6%. The multilayered OLED taken here is named Device A.

TABLE I
COMPARATIVE RESULTS FOR REPORTED EXPERIMENTAL [16]
DATA AGAINST SIMULATED DEVICE RESULTS

Parameters	Experimental Device	Simulated Device	Deviation
Current Density (mA/cm ²)	459.86	445.79	3.05%
Luminescence (cd/m ²)	16916	17190	1.6%

TABLE II
COMPARISON OF MULTILAYERED, SINGLE AND DOUBLE HBL OLED [14]

Parameter Device Name	Current Density (mA/cm ²)	Luminescence (cd/m ²)	Improvement in Luminescence
Multi-Layered OLED	445.79	1.72×10 ⁴	---
Single HBL OLED	299.77	1.94×10 ⁴	13.11%
Double HBL OLED	354.01	2.37×10 ⁴	37.995

III. NOVEL TRIPLE HOLE BLOCK LAYER OLED

The architecture of the OLED is very simple as compared to other organic devices. Each layer is utilized for a specific purpose: charge injection, transport, emission, etc. In case of organic materials, hole mobility is higher as compared to electron mobility for materials discovered so far. Therefore, these charge carriers have a major influence over the device characteristics. In the conventional architecture of organic LED, hole injection was a major issue. To solve this problem hole injection and transport layers were introduced [17]. Further, with the advancements in fabrication processes such as screen printing [7] and inkjet printing [5], these devices depicted enhanced performance. However, still there are a few shortcomings that can be improved. According to the drawbacks of the particular architecture new layers: charge generation layer (CGL) [18], charge transport and carrier layer (CTCL) [19], mixed interlayer (MI) [20], charge block layer [14], [16], etc., are included. Charge block layers are incorporated in the device architecture to restrict the movement of particular charge carrier. This improves the recombination rate. However, their judicious selection also enhances injection of other type of charge carriers [14]. Consequently, a high charge balance is achieved that enhances the recombination rate. Until now single and double HBL architecture have been analyzed resulting in an augmented device luminescence. Their performance [14] is compared with the multilayered OLED in Table II.

Both these devices have the same architecture as multi-layered OLED with the addition of HBLs; BAq for single HBL and BAq + BPhen for double HBL OLED architectures. Table depicts that luminescence performance is highest for double HBL architecture with a reported value of 23,722 cd/m². This is followed by single HBL and multilayered OLED with values: 19444 and 17190 cd/m², respectively. However, the current density shows a slightly opposite trend. Its highest value is 445.79 mA/cm² for multilayered device, followed by 354.01 and 299.77 mA/cm² for double and single HBL devices, correspondingly.

Al/ LiF (Cathode)
Alq₃ (Electron Injection Layer)
BPhen (Hole Block Layer)
BAIq (Hole Block Layer)
TPBi/ CBP (Hole Block Layer)
Alq₃ (Electron transport Layer)
QAD (Emission layer)
Alq₃ (Emission Layer)
NPB (Hole transport Layer)
m-MTDATA (Hole Injection layer)
ITO (Anode)

Fig. 2. Schematic view of triple hole block layer OLED.

TABLE III
DIMENSIONS OF DIFFERENT LAYERS USED IN DEVICE B AND C

Layers' Stack	Material	Device B (All dimensions in nm)	Device C (All dimensions in nm)
Cathode	Al/LiF	50/1	50/1
Electron Injection Layer	Alq ₃	44	44
Hole Block Layer-1	BPhen	6	6
Hole Block Layer-2	BAIq	6	6
Hole Block Layer-3	CBP	---	6
Hole Block Layer-3	TPBi	6	---
Electron Transport Layer	Alq ₃	10	10
Emission Layer	QAD	0.1	0.1
Emission Support Layer	Alq ₃	5	5
Hole Transport Layer	NPB	10	10
Hole Injection Layer	m-MTDATA	45	45
Cathode	ITO	50	50

A low current density for OLED with HBL(s) is justified as movement of holes is restricted within the device. Moreover, double HBL illustrates higher current density compared to single HBL device. It is result of judicious hole block layer selection in the device. These layers possess LUMO levels close to that of adjacent electron transport and injection layer facilitating their injection. Further, double HBL blocks a higher number of holes and as a result their accumulation increases the positive bias within the device that further attracts more electrons [14].

These results illustrate the effectiveness of HBLs to improve the device performance. Therefore, the present article proposes a novel OLED architecture consisting of three HBLs and its analysis. Two OLED devices with third HBL as TPBi 2,2',2''-(1,3,5-Benzinetriyl)-tris(1-phenyl-1-H-benzimidazole) and CBP: 4,4' -Bis(N-carbazolyl)-1,1' -biphenyl, are analysed and named Device B and Device C, respectively. The third HBL is selected such that this can work in tandem with the other HBLs of the OLED. The structure of the organic LED with triple HBL device is shown in Fig. 2 and the dimension of each layer discussed is enlisted in Table III. Similar to double HBL device, on using three HBLs, the dimension of each hole block layer is further reduced to 6 nm, that increased the device dimensions by 2 nm only [14].

The HOMO and LUMO levels are prime consideration while selecting the third HBL. The concept is explained utilizing the energy band diagram in the following sub-section.

A. Energy Level Diagram of OLED

The energy levels; HOMO and LUMO, of the different incorporating layer in the architecture of organic light emitting diode directly impact its performance parameters. This can be explained based on the energy level diagrams of different OLED devices as depicted in Fig. 3. Fig. 3 (a) illustrates the band diagram for the multilayered OLED, Device A. It is observed that the holes are injected from ITO anode into the m-MTDATA layer. One by one holes traverse NPB, Alq₃, QAD, ALq₃ layers and finally reaches the cathode Al: LiF. Similarly, the electrons enter from the cathode and traverse the same path in the reverse direction.

It is observed that the holes do not encounter any barrier while passing from the emission layer QAD until it reaches the cathode. Since, the hole mobility is higher as compared to that of electron mobility, therefore, these charge carriers reach cathode much before electrons reach the emission layer. As a consequence, phenomenon of carrier quenching take place [14]. Resultantly, the recombination rate decreases and so does the device luminescence. Enhanced device luminescence can be achieved by improving the recombination rate within the emission layer. Therefore, the holes need to be restricted within the emission layer. Hole block layers are utilized for this purpose.

Fig. 3 (b) represents the energy band diagram of double hole block layer OLED. It is observed that with the inclusion of double HBLs: BALq and BPhen, the HOMO levels increase substantially. This creates a barrier for the movement of holes and as a result most of these charge carriers are restricted within the emission layer. On the contrary, the barrier force for the electrons is not substantial to prevent their movement within the EML. Moreover, the holes that are restricted within the emission layer increase the positive bias thereby attracting more electrons. Consequently, electron injection also increases. As a result, the overall charge carrier concentration within the emission layer increases, thereby, improving the luminescence.

The process of blocking the holes is further improved with the inclusion of the third HBL: TPBi as illustrated in Fig. 3 (c). The third HBL is selected such that its HOMO level acts as an effective energy barrier for restricting the movement of holes. On the contrary, the LUMO level is almost similar to adjacent layers, thereby, facilitating the electron injection. Utilizing the present architecture, the charge carrier concentration within the emission layer improves, consequently resulting in higher recombination rate and an enhanced luminescence performance. Results pertaining to impact of triple hole block layer are discussed in the succeeding sections.

B. Triple HBL OLED: Results

Results pertaining to the luminescence and current density of triple hole block layers is discussed in the present section. Fig. 4 shows the narrow recombination region in structure

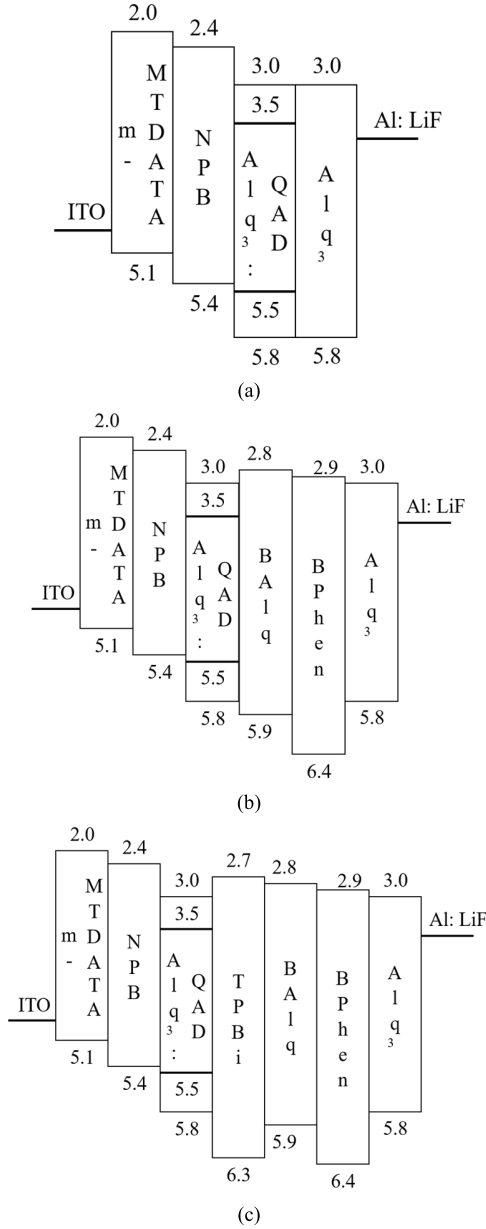


Fig. 3. Energy band diagram of (a) Multilayered OLED, (b) Double HBL OLED and (c) Triple HBL OLED.

of triple HBL OLED as observed from ATLAS at 123 nm. This narrow recombination region is not visible in Device A. Further, the zoomed version in the inset depicts that the emission region covers both sides of the QAD (at 123.1 nm). Thus, these two figures illustrate a good charge balance that exists within the device. This is achieved due to both efficient hole blocking and a well-organized electron injection.

The luminescence and current density result for these two devices are shown in Fig. 5. The result illustrates a significant improvement in the luminescence for Device B and C as compared to Device A. Their luminescence values are 25285 and 24204 cd/m^2 respectively, at an anode potential of 18V. These values are correspondingly 47% and 40% improvement over the multilayered OLED. Even compared to double HBL device the luminescence enhancement is 6.58% and 2.03%

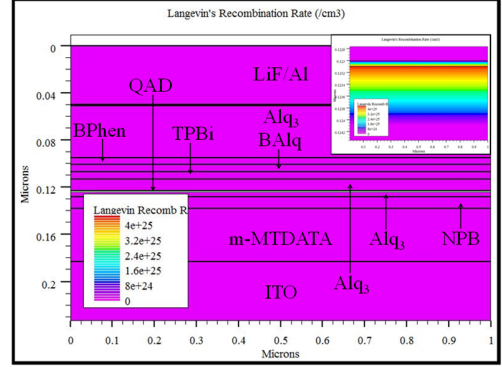


Fig. 4. Langevin's recombination rate as observed in triple HBL OLED.

TABLE IV
COMPARATIVE ANALYSIS OF PARAMETERS FOR DEVICE A, B, AND C

Device Name Parameter	Device A	Device B	Device C
Current Density (mA/cm^2)	445.79	376.34	360.20
Luminescence (cd/m^2)	17190	25285	24204
Improvement in Luminescence with respect to Device A	---	47.09%	40.08%
Luminescence Power Efficiency (lm/W)	6.73	11.73	11.73

for Device B and C, in same order. Moreover, current density values of 376.34 and 360.30 mA/cm^2 for Device B and C, respectively are much closer to Device A as compared to previous devices incorporating HBLs. Luminescence power efficiency [14] is also calculated for these devices based on (5)

$$\eta_P = \frac{L\pi}{JV} \quad (5)$$

where L is the luminescence, J : the current density, and V is operating voltage. Based on (5), power efficiency values for Device A, B, and C are 6.73, 11.73, and 11.73 lm/W . The values for Device B and C also show an improvement of 74.29% over multilayered OLED, Device A. Even though Device B and C have similar architecture, still there is a slight difference in their performance. The reason behind the performance variation is individual properties of these two layers. TPBi has a higher HOMO level (6.3 eV) as compared to CBP (6 eV). This high HOMO level leads for blocking of more holes in Device B, resulting in an improvement for its luminescence performance.

Additionally, the mobility of CBP and TPBi also plays an important role in dictating these characteristics. The electron mobility of TPBi ($5.6 \times 10^{-6} \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$) [21] and CBP ($0.5 \times 10^{-6} \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$) [22] are in the similar range. These electron mobility values correspond closely to mobility of BALq (which is less than $1 \times 10^{-5} \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$) [23] and BPhen ($5.2 \times 10^{-4} \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$) [24]. As electrons are injected from cathode, slowly the mobility value decreases as it reaches emission layer. This will facilitate in higher electron injection but at the same time improving the recombination rate. Since the mobility of CBP is lower as compared to TPBi, hence its current density is adversely affected. The complete comparative result for these devices is tabulated in Table IV.

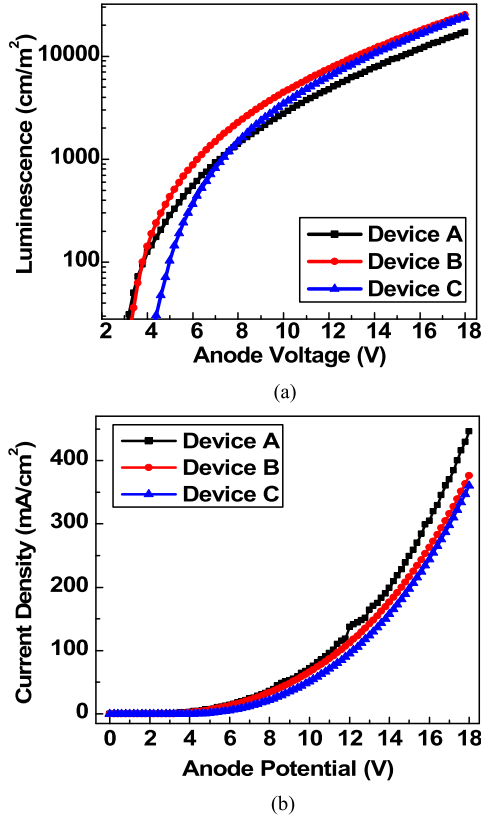


Fig. 5. Characteristics plot for Device A, B and C: (a) Luminescence and (b) Current Density with respect to anode potential.

Thereby, these results highlight the effectiveness of novel device architecture to improve organic LED performance especially in terms of luminescence. The reason for the improved device performance is the higher recombination rate. The reason for higher recombination is a balanced charge carrier injection as shown in the following section.

IV. INTERNAL ANALYSIS OF NOVEL TRIPLE HBL OLED

Internal device analysis gives an insight into the working physics of the device [25], [26]. It is undertaken by utilizing the cutline methodology in ATLAS as depicted in Fig. 6. Various parameters for instance electric field, electron/hole concentration ('e'/'h' concentration), Langevin's recombination rate, etc., can be extracted along this line [27]. The deviation of parameters along the cutline indicates the variation encountered within the different layers of the device. The cutline analysis is implemented on Device A, B, and C with the results illustrated in Fig. 7.

The results depict 'h' and 'e' concentration variation within the different layers of the device in Fig. 7 (a) and (b). Their highest values are almost same, and yet their luminescence characteristic varies by a lot. This is due to the 'h' and 'e' concentration in the proximity of the emission layer. Device A, (emission layer at 111.1 nm) shows a very low 'h' and 'e' concentration within the emission layer as observed in Fig. 7 (a) and (b). The highest 'h' concentration (in vicinity of emission layer) is at 110 nm (towards the cathode). Thus, the recombination also occurs at 110 nm and not within the

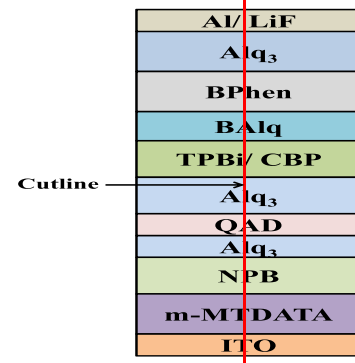


Fig. 6. Illustration of cutline drawn at the centre of triple HBL OLED.

QAD layer as shown in Fig. 7 (c). Therefore, the device shows poor luminescence.

Contrarily, Device B and C, triple HBL OLEDs (emission layer at 123.1 nm), illustrates a high 'h' and 'e' density in and near the emission layer as seen in Fig. 7 (a) and (b). Hence, both these devices depict a high rate of recombination within and near of the emission layer that is evident from Fig. 7 (c). Therefore, both these devices demonstrate a high value of luminescence. These results prove that the novel triple HBL architecture is influential in enhancing the overall device performance. Further, Device B depicts higher 'e' and 'h' concentration in comparison to device C. Thus, TPBi is more effective material in comparison to CBP.

Hence, the rate of recombination is highest for Device B as observed in Fig. 7(c).

V. ANALYTICAL ANALYSIS OF NOVEL TRIPLE HBL OLED

Analytical analysis also gives an insight into the device internal physics [25], [26] however, with a different perspective. Herein, the electric field and mobility behaviour within the OLED is examined. Poisson's equation is applied to determine the electric potential throughout the device [28], [29]. Whereas, Drift-Diffusion model examines 'e' and 'h' concentration and their respective current density [19]. Mobility of charge carriers is calculated utilizing Poole Frenkel model [30]. Poisson's equation obtains the inbuilt electric potential within the device. It is the result of variation in density distribution of charge carrier on the application of external electric potential. The equation for organic devices is expressed as:

$$E\left(z + \frac{\Delta z}{2}, t\right) = E\left(z - \frac{\Delta z}{2}, t\right) + \Delta z \frac{q}{\epsilon} \{p(z, t) - n(z, t) + N_D(z) - N_A(z)\} \quad (6)$$

The different parameters in (6) are: ' Δz ' represents the mesh width. These mesh points serve as nodes where different calculations are performed within the device. ' t ' symbolizes the time frame for the analysis to be performed. ' $E(z, t)$ ' is the electric potential obtained at various mesh points. Further, ' q ' and ' ϵ ' are electron charge and relative permittivity of the material, respectively. ' $p(z, t)$ ' and ' $n(z, t)$ ' are 'h' and

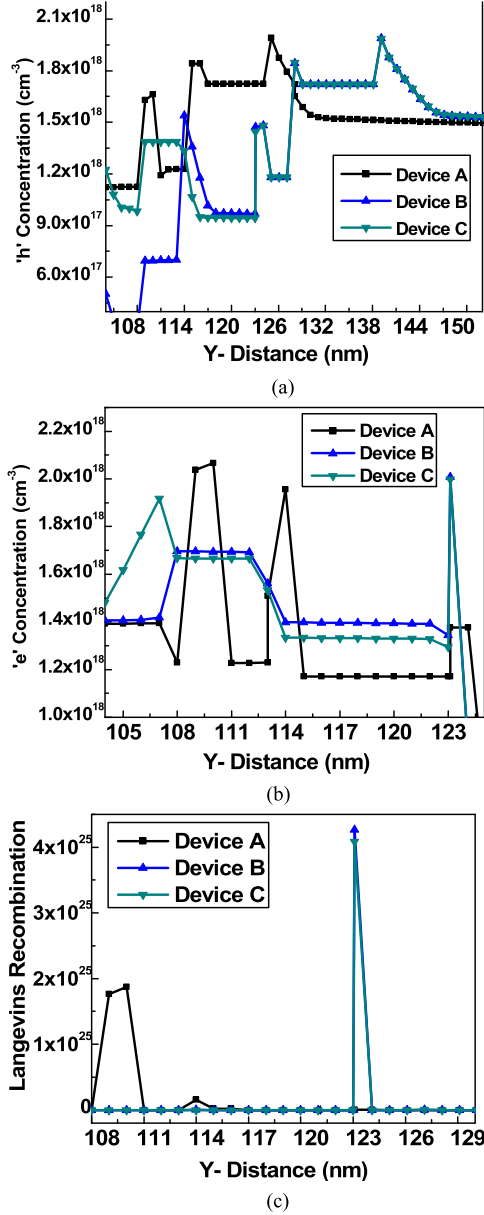


Fig. 7. Cumulative plots illustrating (a) 'h' concentration, (b) 'e' concentration and (c) Langevin's recombination rate for Devices A, B, and C.

'e' concentration correspondingly, as a function of z (device thickness) and t . At last, ' $N_A(z)$ ' and ' $N_D(z)$ ' represents the respective impurity concentration of acceptor and donor type, at point z . The calculation of inbuilt electric potential is dependent on determining the variation of charge carrier density. It is calculated with the help of Drift-Diffusion equation. The hole and electron density are expressed as:

$$p(z, t + \Delta t) = p(z, t) - \Delta t \left\{ \frac{1}{q} \frac{J_p \left(z + \frac{\Delta z}{2}, t \right) - J_p \left(z - \frac{\Delta z}{2}, t \right)}{\Delta z} + r(z, t) n(z, t) p(z, t) \right\} \quad (7)$$

$$n(z, t + \Delta t) = n(z, t) + \Delta t \left\{ \frac{1}{q} \frac{J_n \left(z + \frac{\Delta z}{2}, t \right) - J_n \left(z - \frac{\Delta z}{2}, t \right)}{\Delta z} - r(z, t) n(z, t) p(z, t) \right\} \quad (8)$$

' Δt ' is the time required for revising the values of these time dependent charge carrier densities. The recombination rate ' $r(z, t)$ ' is expressed in Langevin's form for the organic devices. Finally, ' J_p ' and ' J_n ': current density due to holes and electrons respectively, are obtained as:

$$J_p \left(z + \frac{\Delta z}{2}, t \right) = q \mu_p \left(z + \frac{\Delta z}{2}, t \right) \frac{p(z, t) + p \left(z + \frac{\Delta z}{2}, t \right)}{2} E \left(z + \frac{\Delta z}{2}, t \right) - K T \mu_p \left(z + \frac{\Delta z}{2}, t \right) \frac{p \left(z + \frac{\Delta z}{2}, t \right) - p(z, t)}{\Delta z} \quad (9)$$

$$J_n \left(z + \frac{\Delta z}{2}, t \right) = q \mu_n \left(z + \frac{\Delta z}{2}, t \right) \frac{n(z, t) + n \left(z + \frac{\Delta z}{2}, t \right)}{2} E \left(z + \frac{\Delta z}{2}, t \right) + K T \mu_n \left(z + \frac{\Delta z}{2}, t \right) \frac{n \left(z + \frac{\Delta z}{2}, t \right) - n(z, t)}{\Delta z} \quad (10)$$

' μ_p ' and ' μ_n ' in (9) and (10) represent hole and electron mobility values, correspondingly, whereas ' K ' and ' T ' have standard meanings. Since, mobility in organic devices follow Poole Frenkel mobility model, therefore, it is utilized for their determination. The model is expressed as:

$$\mu(E(z, t)) = \mu_0 \exp \left(\sqrt{\frac{E(z, t)}{E_0}} \right) \quad (11)$$

The unknown parameters in (11) are: ' μ_0 ' that denotes the mobility at null or zero electric field, whereas ' E_0 ' stands for the value of characteristic field (V/cm).

The complete model is validated for the multilayered OLED and the results are present in [30] and tabulated in Table V. Different parameters in the table illustrate a close trend between the analytically calculated results and the ones obtained through internal device analysis. Therefore, the model satisfactorily analyse the OLED and is utilized herein for the numerical analysis of the triple HBL device. The analysis is performed on Device B owing to its best performance. Thereafter, results are compared with internal analysis results. The analysis is aimed at justifying the reasons for enhanced performance of Device B.

Foremost, the electric field within the device is extracted numerically as well as through simulation. Fig. 8 depicts a comparative plot between these two analysis results for electric field. It is observed from the plot that both electric field curves are identical with a characteristic peak at the centre of

TABLE V
MULTILAYERED OLED: COMPARISON OF NUMERICAL
AND INTERNAL ANALYSIS RESULTS [30]

Name of Parameter	Maximum Magnitude	
	Analytical Analysis	Internal Analysis
Electric Field (V/cm)	3.6×10^6	4.27×10^6
'e' Concentration (cm^{-3})	2.4×10^{18}	1.97×10^{18}
'h' Concentration (cm^{-3})	4.2×10^{18}	1.99×10^{18}
'e' Mobility (cm^2/Vs)	0.027	0.82
'h' Mobility (cm^2/Vs)	0.033	0.68
'e' Current Density (A/cm^2)	0.401	0.443
'h' Current Density (A/cm^2)	0.401	0.446

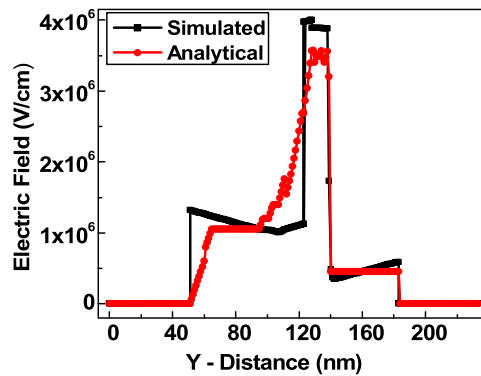


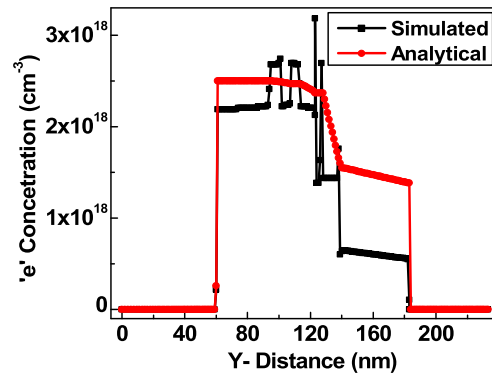
Fig. 8. Combined plots for simulated and analytical electric field.

OLED at 123 nm (i.e. the emission layer). Simulated values are little higher as compared to numerically extracted values. These are 4.03×10^6 and 3.59×10^6 respectively with an error rate of about 12%.

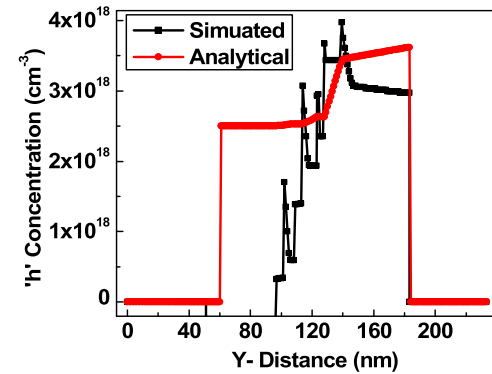
Charge carrier concentration is extracted utilizing drift diffusion model. The curves for 'e' and 'h' concentration are illustrated in Fig. 9 (a) and (b) respectively. Device B constitutes an array of layers and therefore, internal analysis (simulated) results depict multiple peaks for both 'h' and 'e' concentration as different layers are traversed. The analytical results follow a similar trend however, little less pronounced peaks are observed in charge carrier concentration. Both these results demonstrate a high 'e' and 'h' concentration in the vicinity of emission layer. Thereafter, charge carrier concentration of both types fall to a low value, suggesting recombination processing occurring in this region.

The mobility values for the charge carrier are extracted next as shown in Fig. 10. The curves for 'e' and 'h' mobility resemble closely to the electric field curves. Both these curves depict a distinctive peak around emission layer (123 nm). This suggests that the mobility depends on the electric field. Thereby, it demonstrates Poole Frenkel mobility behavior. Only analytical results are shown owing to the slight difference in the peak values of carrier mobility for analytical and internal analysis results, similar to multilayered OLED.

Fig. 11 show curves for 'e' and 'h' current density. The internal and analytical analysis values matches with minor error of 6%. However, it is observed that analytical 'e' and 'h' current values are observed in the entire region covered with organic semiconductors. Thereby, it suggests further scope to improve the architecture of the OLED. Table VI highlights the



(a)



(b)

Fig. 9. Cumulative plots for: (a) 'e' concentration and (b) 'h' Concentration.

TABLE VI
DEVICE B: COMPARISON OF ANALYTICAL
AND INTERNAL ANALYSIS RESULTS

Name of Parameter	Maximum Magnitude	
	Analytical Analysis	Internal Analysis
Electric Field (V/cm)	3.59×10^6	4.03×10^6
'e' Concentration (cm^{-3})	2.53×10^{18}	3.19×10^{18}
'h' Concentration (cm^{-3})	3.63×10^{18}	3.96×10^{18}
'e' Mobility (cm^2/Vs)	0.005	1.06
'h' Mobility (cm^2/Vs)	0.03	0.92
'e' Current Density (A/cm^2)	0.402	0.377
'h' Current Density (A/cm^2)	0.402	0.375

comparison of analytical and internal analysis results. Both these results show a similar trend. The analytical analysis results demonstrate performance improvement in the device as a result of higher 'h' and 'e' concentration. Novel OLED architecture neutralizes the impact of higher hole mobility in comparison to electron mobility. Consequently, a balanced recombination is achieved within the emission layer that enhances the device luminescence characteristic.

VI. OLED FOR DIAGNOSIS OF OVARIAN CANCER

The present section illustrates utilization of OLED as a light detector for the diagnosis of ovarian cancer. There are various other devices such as solar cells, photo diodes and photo detectors, etc., that can serve the same purpose. However, OLED is preferred for the detection purpose focusing basically

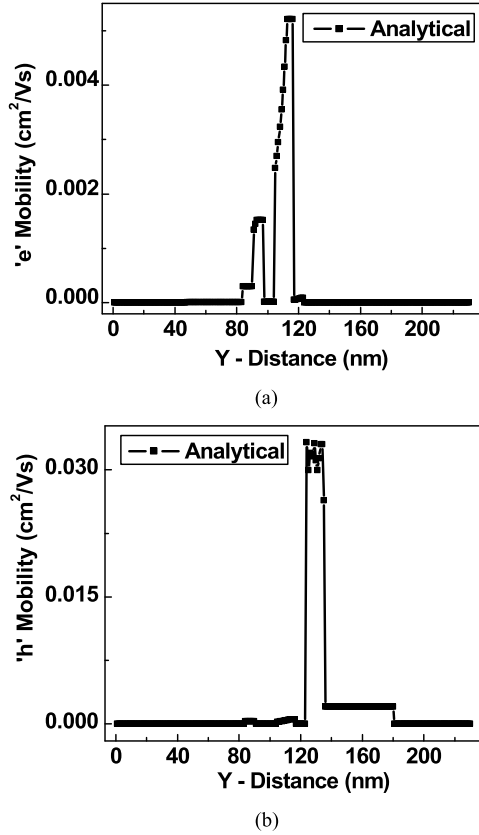


Fig. 10. Field dependent mobility plot: (a) 'e' mobility and (b) 'h' mobility.

on the ease of fabrication. It is well known and highlighted by many researchers [6], [31] that realization of two different devices on a single substrate is probable but even with similar materials, the fabrication process is quite complicated.

Therefore, if the same device can be utilized as a light source as well as for detection, the fabrication process becomes highly simplified and standardized. Moreover, OLED is also utilized herein because of its huge color gamut. This means that within a limited spectrum, the OLED can produce a larger variation of colors as compared to other devices. This depicts that the OLED can differentiate between wavelengths closer to each other much easily. Hence, the device is much more sensitive to light and thus can produce different current values for light with wavelength close to each other.

The method used for light based detection of ovarian cancer is suggested by Zavirik *et al.* [32]. Previous literature reviews [33], [34] are also available that highlights interaction of human urine and blood with light for the detection of various compounds present therein. Additionally, some researchers [35], [36] have also highlighted detection of different types of cancer by analysing fluorescence spectra of various compounds present in human urine. Their research depicted that upon excitation, fluorescence from the urine samples of cancer patients varies from that of a normal humans. The reason for this variation in emission spectrum is the result of lower pyridoxic acid concentration in an oncological patient [32].

Urine samples are excited at varying wavelength in the range: 250-530 nm and the emission spectrum are observed

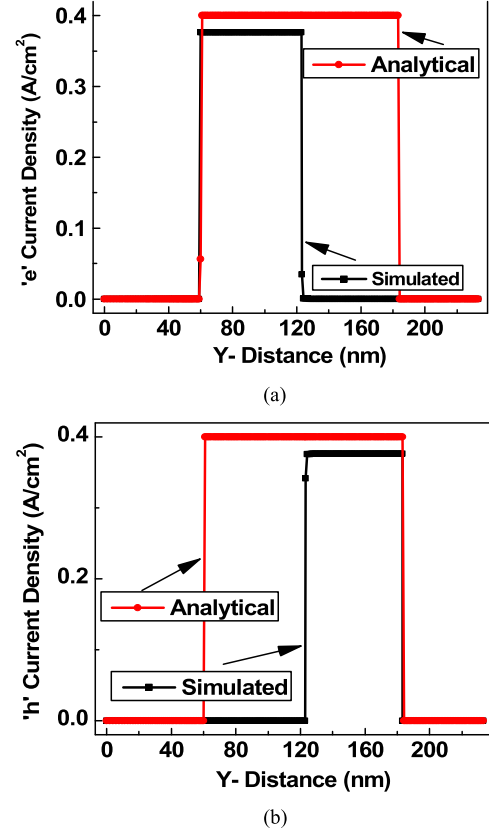


Fig. 11. Combined plots for (a) 'e' current density and (b) 'h' current density.

TABLE VII
FLUORESCENCE COMPARISON OF HEALTHY
HUMAN AND OVARIAN CANCER PATIENT

Excitation Wavelength	Fluorescence at 420 nm	Fluorescence at 440 nm	Outcome
330 nm	High ↑	Low ↓	Healthy Human
370 nm	Low ↓	High ↑	Ovarian Cancer Patient

in the interval 390-460 nm. Fluorescence emission for healthy human shows a peak at 420 nm when excited at 330 nm. However, the same for cancer patient observed at 440 nm for excitation of 379 nm [32]. Table VII tabulates the data in this regard. Light emitting diodes (conventional LEDs) depicts a property to produce current corresponding to any wavelength of light lower than its own emission wavelength. The same principle is applied herein for the OLED to detect light.

Methodology suggested by Zavirik *et al.* [32] utilized spectrophotometer for detection of fluorescence. Similarly, an effective methodology can be developed with the help of OLED for diagnosis of ovarian cancer as illustrated in Fig. 12. The novel OLED architecture utilized for the detection of ovarian cancer consists of different layers that results in a balanced electron and hole injection. There are hole injection and transport layer (m-MTDATA and NPB) similar to electron injection and transportation layer (Alq₃).

The role of these layers is to enhance charge carrier within the device. However, due to higher mobility of holes as

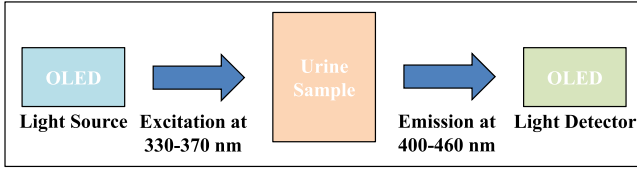


Fig. 12. Methodology to utilize OLED for diagnosis of ovarian cancer.

compared to electrons most of them get exhausted at the opposite electrode. Therefore, to prevent these holes from reaching the opposite electrode, hole block layers are used. Therefore, a higher recombination occurs in the emission layer and due to higher charge carrier injection, current within the device is also improved.

This OLED device is used as a light source and fluorescence sensor for the detection of ovarian cancer. As a result of these additional layers, an enhanced luminescence is achieved, that is helpful for better excitation of the urine sample. At the same time, the OLED based fluorescence sensor is able to produce a higher photo current due to the improved device architecture. Thus, the overall architecture is helpful in efficient detection of cancer. Therefore, the light from OLED interacts with human urine sample resulting in its excitation and fluorescence emission. OLED light sensor detects this emission. Therefore, the complete system can be used for the diagnosis of ovarian cancer. The first step towards utilization of OLED for the detection of light is to determine its emission wavelength as:

$$E_g = hc/\lambda \quad (12)$$

where 'E_g' represents the band gap of a particular light emitting material, 'h' stands for Plank's constant, 'c' is the speed of light and λ: the wavelength of light emitted.

The architecture of Device B consists of Alq₃ and QAD, both capable of emission. Therefore, the emission wavelength of the device will be somewhere in midst of the spectrum of these two materials. Based on the energy band gap for Alq₃ (2.8 eV) and QAD (2 eV) their emission wavelength is 443nm and 621nm, respectively. Hence, the proposed OLED is capable of producing a current for any wavelength below 443 nm. Analysis is carried out in ATLAS, wherein a light beam of intensity 1 W/cm² and wavelength varying from 0-720 nm is made incident on surface of the OLED. The cutline is drawn to analyse the performance of the device working as a fluorescence detector as illustrated in Fig. 7. The main focus is on the emission layer.

In the device architecture, the main emission layer is QAD, however Alq₃ is also capable of producing light. The device architecture is such that, when the device is used as a light source, maximum recombination occurs within the emission layer. However, while device is utilized as a fluorescence sensor, the light falls on both Alq₃ and QAD layer and these layers produce photo current. Since dimension of Alq₃ layer is much greater than QAD layer, therefore, the layer dominates the photo generation current. The photo-generation is observed in Device B, upon incidence of light and the result is depicted in Fig. 13.

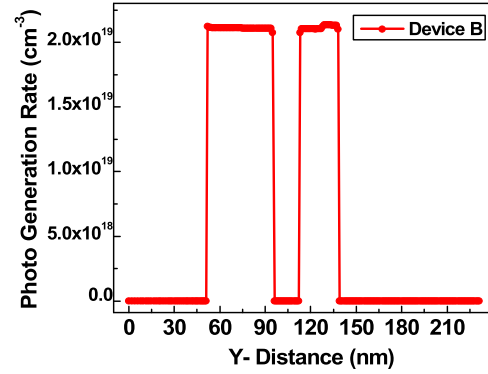
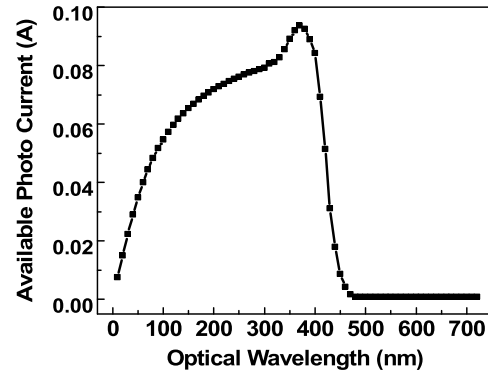
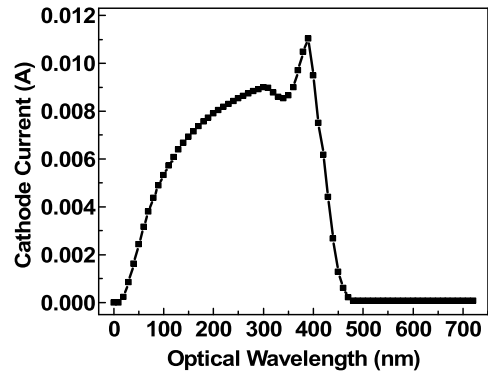


Fig. 13. Photo generation rate as observed in Device B.



(a)



(b)

Fig. 14. Current variation observed after light incidence on Device B (a) Total photo generation current available and (b) Cathode current.

Fig. 13 depicts that the photo-generation is observed only in QAD and Alq₃ layers of Device B. Alq₃ layer dominates the photo generation rate owing to its larger dimensions. If a small bias is applied on the device the generated electrons and holes can be easily separated to produce the corresponding photo current. Cutline analysis is performed to extract these photo-generated current values. Fig. 14 shows the available photo-generation current and cathode current for Device B. Their maximum values obtained are 93 and 11 mA, respectively.

Further, it is also observed that current values vary from 0-450 nm range only and thereafter, falls to zero. Its highest value is observed in range of 400-440 nm. These

TABLE VIII
PHOTO-CURRENT PARAMETERS OBSERVED FOR DEVICE B

Parameter	Device B
Total Photo-generation Current Available (mA)	93
Current at Cathode	11
Photo-generation Current Available at 420 nm	50
Photo-generation Current Available at 440 nm	7
Cathode Current Observed at 420 nm	5
Cathode Current Observed at 440 nm	1

values correspond to the emission wavelength of Alq₃ layer which dominates the photo generation in the OLED. Photo current generation is high in device B as is tabulated in Table VIII. Difference between cathode current at emission wavelength of 420 and 440 nm is 4 mA. Hence, Device B can easily differentiate between the emission spectra at 420 and 440 nm and therefore, between healthy person and oncological patient.

The complete system can be amalgamated to develop a low cost portable sensor based on organic LED that might replace expensive medical tests.

VII. CONCLUSION

The present research article proposes a novel triple HBL OLED architecture to enhance its performance. Thereafter, the article highlights its utilization as a light sensor for diagnosis of ovarian cancer. Triple HBL OLEDs, Device B and C, are analyzed and their performance is compared to multilayered OLED, Device A. Compared to Device A, both these devices depicted a luminescence improvement of 47.09% (25285 cd/m²) and 40.08% (24204 cd/m²), respectively. Improved luminescence results from enhanced recombination rate. Judicious selection of HBLs increase charge carrier concentration as it blocks the holes and enhances electron injection at the same time. This is also evident through analytical modeling and internal device analysis.

Both internal and analytical analysis are performed for the proposed OLED. Internal analysis is performed using Silvaco Atlas tool, whereas analytical analysis is performed with the help of model equations governing device working. Both these analysis results highlight different aspects related to internal physics of the device. The internal analysis is utilized herein to extract electron and hole concentration along with Langevin's recombination rate. On the other hand, analytical analysis is performed to observe the electrical properties and manner in which charge carrier concentration varies within the device.

The analytical modeling is based on Poisson's equation and drift diffusion equation. Parameters such as electric field, 'e' and 'h' concentration, etc. are extracted. The analytical results illustrate a high electron and hole concentration within the emission layer of triple HBL OLEDs as compared to Device A. Consequently, recombination takes place within the emission layer. The analytical analysis results also highlight presence of Poole Frenkel mobility behavior within the device.

Finally, article depicts the application of OLED for diagnosis of ovarian cancer. Fluorescence emission from urine samples of the healthy person and oncological patient shows

a peak emission at 420 and 440 nm wavelengths, respectively. Triple HBL OLED, Device B detects this fluorescence emission and produces a corresponding cathode current of 5 and 1 mA with respect to these two emission wavelengths. Hence, a healthy person is differentiated from oncological patient. Therefore, using the present methodology OLED based portable hand-held device can be developed for the diagnosis of ovarian cancer.

REFERENCES

- [1] J.-Y. Jeon, Y.-J. Jeon, Y.-S. Son, and G.-H. Cho, "A double zeros compensated direct fast feedback current driver for medium to large AMOLED displays," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 59, no. 10, pp. 2197–2209, Oct. 2012.
- [2] B. Kumar, B. K. Kaushik, and Y. S. Negi, "Organic thin film transistors: Structures, models, materials, fabrication, and applications: A review," *Polym. Rev.*, vol. 54, no. 1, pp. 33–111, Jan. 2014.
- [3] S. Negi, P. Mittal, and B. Kumar, "In-depth analysis of structures, materials, models, parameters, and applications of organic light-emitting diodes," *J. Electron. Mater.*, vol. 49, no. 8, pp. 4610–4636, Aug. 2020.
- [4] H. C. Chen *et al.*, "Polymer inverter fabricated by inkjet printing and realized by transistors arrays on flexible substrates," *J. Display Technol.*, vol. 5, no. 6, pp. 216–223, May 2009.
- [5] X. Liu *et al.*, "Iridium (III)-complexed polydendrimers for inkjet-printing OLEDs: The influence of solubilizing steric hindrance groups," *ACS Appl. Mater. Interface*, vol. 11, no. 29, pp. 26174–26184, 2019.
- [6] D. Threm, J. L. Gugat, A. Pradana, M. Radler, J. Mikat, and M. Gerken, "Self-aligned integration of spin-coated organic light-emitting diodes and photodetectors on a single substrate," *IEEE Photon. Technol. Lett.*, vol. 24, no. 11, pp. 912–914, Jun. 2012.
- [7] L. Zhou *et al.*, "Screen-printed poly (3, 4-ethylenedioxythiophene): Poly (styrenesulfonate) grids as ITO-free anodes for flexible organic light-emitting diodes," *Adv. Funct. Mater.*, vol. 28, no. 11, 2018, Art. no. 1705955.
- [8] A. Samore, M. Rusci, D. Lazzaro, P. Melpignano, L. Benini, and S. Morigi, "BrightNet: A deep CNN for OLED-based point of care immunofluorescent diagnostic systems," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 9, pp. 6766–6775, Sep. 2020.
- [9] K. F. H. Or, F. K., and J. M., "Exploiting the potential of OLED-based photo-organic sensors for biotechnological applications," *Chem. Sci. J.*, vol. 7, no. 3, pp. 1–10, 2016.
- [10] Y.-S. Son, Y.-J. Jeon, J.-Y. Jeon, and G.-H. Cho, "Transient charge feedforward driver for high-speed current-mode data driving in active-matrix OLED displays," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 3, pp. 539–547, Mar. 2010.
- [11] P. A. Haigh *et al.*, "Exploiting equalization techniques for improving data rates in organic optoelectronic devices for visible light communications," *J. Lightw. Technol.*, vol. 30, no. 19, pp. 3081–3088, Oct. 2012.
- [12] P. A. Haigh, Z. Ghassemlooy, and I. Papakonstantinou, "1.4-Mb/s white organic LED transmission system using discrete multitone modulation," *IEEE Photon. Technol. Lett.*, vol. 25, no. 6, pp. 615–618, Mar. 2013.
- [13] P. A. Haigh, Z. Ghassemlooy, I. Papakonstantinou, and H. Le Minh, "2.7 Mb/s with a 93-kHz white organic light emitting diode and real time ANN equalizer," *IEEE Photon. Technol. Lett.*, vol. 25, no. 17, pp. 1687–1690, Sep. 2013.
- [14] S. Negi, P. Mittal, and B. Kumar, "Impact of different layers on performance of OLED," *Microsyst. Technol.*, vol. 24, no. 12, pp. 4981–4989, Dec. 2018.
- [15] P. Mittal, Y. S. Negi, and R. K. Singh, "Mapping of performance limiting issues to analyze top and bottom contact organic thin film transistors," *J. Comput. Electron.*, vol. 14, no. 1, pp. 360–379, Mar. 2015.
- [16] H. Yang, Y. Zhao, J. Hou, and S. Liu, "Organic light-emitting devices with double-block layer," *Microelectron. J.*, vol. 37, no. 11, pp. 1271–1275, Nov. 2006.
- [17] X.-C. Li *et al.*, "Diindolotriazatruxene-based hole-transporting materials for high-efficiency planar perovskite solar cells," *ACS Appl. Mater. Interface*, vol. 11, no. 49, pp. 45717–45725, Dec. 2019.
- [18] Y. Chen, J. Chen, D. Ma, D. Yan, L. Wang, and F. Zhu, "High power efficiency tandem organic light-emitting diodes based on bulk heterojunction organic bipolar charge generation layer," *Appl. Phys. Lett.*, vol. 98, no. 24, p. 114, 2011.

- [19] J. Park, Y. Kawakami, and S.-H. Park, "Numerical analysis of multilayer organic light-emitting diodes," *J. Lightw. Technol.*, vol. 25, no. 9, pp. 2828–2836, Sep. 2007.
- [20] C.-H. Gao, X.-B. Shi, D.-Y. Zhou, L. Zhang, Z.-K. Wang, and L.-S. Liao, "Highly efficient white organic light-emitting diodes with controllable excitons behavior by a mixed interlayer between fluorescence blue and phosphorescence yellow-emitting layers," *Int. J. Photoenergy*, vol. 2013, pp. 1–7, Jan. 2013.
- [21] H.-Y. Li *et al.*, "Highly efficient green phosphorescent OLEDs based on a novel iridium complex," *J. Mater. Chem. C*, vol. 1, no. 3, pp. 560–565, 2013.
- [22] P. Chulkin, O. Vybornyi, M. Lapkowski, P. J. Skabara, and P. Data, "Impedance spectroscopy of OLEDs as a tool for estimating mobility and the concentration of charge carriers in transport layers," *J. Mater. Chem. C*, vol. 6, no. 5, pp. 1008–1014, 2018.
- [23] S. Reineke, F. Lindner, Q. Huang, G. Schwartz, K. Walzer, and K. Leo, "Measuring carrier mobility in conventional multilayer organic light emitting devices by delayed exciton generation," *Phys. Status Solidi (B)*, vol. 245, no. 5, pp. 804–809, May 2008.
- [24] S. Naka, H. Okada, H. Onnagawa, and T. Tsutsui, "High electron mobility in bathophenanthroline," *Appl. Phys. Lett.*, vol. 76, no. 2, pp. 197–199, Jan. 2000.
- [25] P. Mittal, Y. S. Negi, and R. K. Singh, "An analytical approach for parameter extraction in linear and saturation regions of top and bottom contact organic transistors," *J. Comput. Electron.*, vol. 14, no. 3, pp. 828–843, Sep. 2015.
- [26] B. Kumar, B. K. Kaushik, Y. S. Negi, S. Saxena, and G. D. Varma, "Analytical modeling and parameter extraction of top and bottom contact structures of organic thin film transistors," *Microelectron. J.*, vol. 44, no. 9, pp. 736–743, Sep. 2013.
- [27] P. Mittal, Y. S. Negi, and R. K. Singh, "A depth analysis for different structures of organic thin film transistors: Modeling of performance limiting issues," *Microelectron. Eng.*, vol. 150, pp. 7–18, Jan. 2016.
- [28] G. G. Malliaras and J. C. Scott, "Numerical simulations of the electrical characteristics and the efficiencies of single-layer organic light emitting diodes," *J. Appl. Phys.*, vol. 85, no. 10, pp. 7426–7432, May 1999.
- [29] J. Park, T. Kim, J. Lee, and D. Shin, "Energy loss mechanism in organic and inorganic light-emitting diodes," *IEEE Photon. Technol. Lett.*, vol. 20, no. 16, pp. 1408–1410, Aug. 2008.
- [30] S. Negi, P. Mittal, and B. Kumar, "Analytical modelling and parameters extraction of multilayered OLED," *IET Circuits, Devices Syst.*, vol. 13, no. 8, pp. 1255–1261, Nov. 2019.
- [31] E. Manna, T. Xiao, J. Shinar, and R. Shinar, "Organic photodetectors in analytical applications," *Electronics*, vol. 4, no. 3, pp. 688–722, Sep. 2015.
- [32] M. Zvarik, D. Martinicky, L. Hunakova, I. Lajdova, and L. Sikurova, "Fluorescence characteristics of human urine from normal individuals and ovarian cancer patients," *Neoplasma*, vol. 60, no. 5, pp. 533–537, 2013.
- [33] M. J. P. Leiner, M. R. Hubmann, and O. S. Wolfbeis, "The total fluorescence of human urine," *Analytica Chim. Acta*, vol. 198, pp. 13–23, Jan. 1987.
- [34] D. Yim, G. V. G. Baranoski, B. W. Kimmel, T. F. Chen, and E. Miranda, "A cell-based light interaction model for human blood," *Comput. Graph. Forum*, vol. 31, nos. 2–4, pp. 845–854, May 2012.
- [35] N. Bosschaart, G. J. Edelman, M. C. G. Aalders, T. G. van Leeuwen, and D. J. Faber, "A literature review and novel theoretical approach on the optical properties of whole blood," *Lasers Med. Sci.*, vol. 29, no. 2, pp. 453–479, Mar. 2014.
- [36] V. Masilamani, T. Vijmasi, M. Al Salhi, K. Govindaraj, A. P. Vijaya-Raghavan, and B. Antonisamy, "Cancer detection by native fluorescence of urine," *J. Biomed. Opt.*, vol. 15, no. 5, 2010, Art. no. 057003.



with the Department of ECE, Tula's Institute, Dehradun.

Shubham Negi (Member, IEEE) received the B.Tech. degree in electronic and communication engineering from Hemwati Nandan Bahuguna Garhwal University (Central University), Srinagar (Garhwal), India, in 2013, the Master of Technology degree in VLSI design and systems from Graphic Era (Deemed to be University), Dehradun, India, and the Ph.D. degree with a research focus on enhancing the performance of organic devices and their subsequent utilization in novel applications. He is currently working as an Assistant Professor



power VLSI circuits. She has published one patent on novel OTFT structure and text book titled *Organic Thin-Film-Transistor Applications: Materials to Circuits* (CRC Press) (U.K.: T&F, 2016). She is a Reviewer of many IEEE transactions and other international journals of IEEE, IET, Elsevier, Springer, IOP, Wiley, and T&F. She is the life member of many professional societies. She has received the research awards in 2012 and 2015 for dedicated research from Graphic Era University, Dehradun, India. She also received Commendable Research Award in 2019 and 2020 from DTU.

Poornima Mittal (Member, IEEE) received the B.Tech., M.Tech., and Ph.D. degrees. She has more than 15 years of academic and research experience. She is currently working as an Associate Professor with the Department of ECE, Delhi Technological University (DTU), Delhi, India. She has published more than 110 international journals/conference papers/book chapters. Her research interests include design/modeling of flexible electronic devices, material synthesis and characterization, thin film fabrication, OLED, solar cell, memory design, and low



(University of UP Government). He has more than 19 years of experience in the field of academic and research. He has received various awards and certificates of appreciations for his dedicated academic and research activities. He is currently working on novel structures of OTFT, organic solar cells, and OLED displays. His name has been listed in *Marquis Who's Who in the World, USA*. He is a Life Member of Indian Society Technical Education (ISTE) and International Association of Engineers (IAENG). He is a Reviewer of many international journals with reputed banners, including IEEE, IET, Elsevier, Springer, and Taylor & Francis publishers.

Brijesh Kumar (Member, IEEE) received the Ph.D. degree from the Indian Institute of Technology (IIT) Roorkee, India, in 2014. He has more than 120 reputed international, national journals and conference publications. His research interests include VLSI design and technology, organic material-based novel devices and circuits, and solid-state devices and circuits. He is currently working as a Professor with the Department of Electronics and Communication Engineering, Madan Mohan Malaviya University of Technology (MMMUT), Gorakhpur,

MODELING FOR THE ENERGY POTENTIAL OF BIOGAS POWER PLANTS IN NATIONAL CAPITAL TERRITORY

^a Rohit Agrawal, ^b S.K. Singh

^{a, b} Department of Environmental Engineering, Delhi Technological University, Delhi, India
Corresponding Author: ^a rohitagrawal_2k19ene04@dtu.ac.in, ^b sksinghdce@gmail.com

Abstract

Biogas is a renewable energy source which is being researched and widely developed as a future alternative energy source that is economical, sustainable, and environmentally friendly. Under the scheme for National Capital Territory, we used 150 cows for example where the dung from the cows is processed. Biogas production in every day which is 54 m³ or equivalent to 54,000 liters. Biogas can be used as a generating system capable of producing energy of 540 kWh each day with a power of 540 kW. The generator system in this study divided into 2 parts, namely first, a simple generator system (Digester-biogas-Genset 30000W-electricity biogas) which is assumed to operate for 24 hours a day with the energy output from this biogas power plant is 613.8 kWh per day. Generating system, the second is a generator system using HOMER (Thermal-Boiler-Generator Bio 2 kW-Converter of 10 kW-electric load) with an energy output of 613.8 kWh per day. In realizing an efficient biogas-based generating system, so in this study use HOMER software to optimize generator size and value economic power plant with coverage in the form of net present cost (NPC) of Rs.2,50,000.00 and the cost of Energy (COE) of Rs. 1.57 per unit

Keywords: Anaerobic, Biogas, Digester, Energy, Homer

1. Introduction

Energy has a very important role in the activities of human life; the increasing use of energy has become the world's talk, especially in India. Some of the energy which is used by the Indian people today comes from hydro, solar, and fossil fuels, namely petroleum, coal, and gas. Based on India's Energy outlook, the national energy demand continues to increase along with economic growth, populations, energy costs, and governmental policies. An average Gross Domestic Product (GDP) growth rate of 4.14% per year and the population growth rate of 0.78% per year during 2016-2020, the growth rate for the final energy demand is approx. 2.3% per year as per International Energy Agency. "The total primary energy consumption from coal (452.2 Mtoe; 55.88%), crude oil (239.1 Mtoe; 29.55%), natural gas (49.9 Mtoe; 6.17%), nuclear energy (8.8 Mtoe; 1.09%), hydroelectricity (31.6 Mtoe; 3.91%) and renewable power (27.5 Mtoe; 3.40%) is 809.2 Mtoe (excluding traditional biomass use) in the calendar year 2018." As time goes by, fossil fuels because they are classified as non-renewable energy will sooner or later be depleted or their availability crisis; Based on statistical data on New Renewable Energy and Energy Conservation by International Energy Agency, India has a new and renewable energy source, namely bio-energy with a potential of 18,000 MW with an installed on-grid system capacity of 220.8 MW, of which only a small amount is still being utilized. As a South Asian country, India has abundant bio-energy as energy potential that can be used as renewable energy source to replace fossil energy which is still widely used today, as well as to maintain national energy security [1-2].

Delhi is the National Capital Territory in India that has installed electrical energy capacity from several power plants in several locations generates approx. 2,000 MW of electricity. The power plant used to supply electrical energy by Gas only. Whereas, 54MW of energy is generated from Bio-Waste from 5250 ton of per day waste; The need for electrical energy in Delhi in 2021 is 2,160.30 MW consisting of the government sector of 63.56 MW, the household sector of 1,330.85 MW, the industrial sector of 130.05 MW, the business sector of 508.15 MW, the social sector is 91.31 MW, the public lighting sector is 35.36 MW. With such a large energy demand, Delhi is still experiencing a deficit of 47.32 MW of electrical energy. The impact of the deficit in electrical energy is the occurrence of rotating blackouts in the National Capital Region of Delhi. Utilizing new and renewable energy to overcome the deficit in electrical energy in the Delhi is the best way. In addition to fulfilling energy needs, the use of new and renewable energy can also reduce environmental pollution caused by the use of fossil energy. There is still a lot of potential from biomass in the National Capital Region of Delhi that has not been utilized to overcome the problem of energy in electricity, one of the abundant biomass potentials that are still underutilized is solid waste or garbage [3]. The National Capital Region of Delhi has fertile soils and wild plants that are easily available. With this geographic condition, it is easy to develop the livestock sector. Based on data from Animal Husbandry Statistics Division, livestock in Delhi has a population consisting of 6 types of animals, namely cows, buffaloes, dairy cows, goats, sheep, and pigs approx. 16,00,000 heads, and continued to increase in 2021 to 16,34,128 heads. Of the livestock population, the cattle population has the highest number compared to other animals [4-5].

Previous research has examined the potential for cow dung to be used as a source of power generation, but this study is still a hypothesis, so the test is less accurate in actual conditions. The biogas production process using the anaerobic digestion method made from cow dung, it goes through several stages, each of which has erratic changes that can affect the production of biogas produced [6-7]. In this study, it has also analyzed the economic value that gets positive values so that the design is

feasible to be realized. The process of forming biogas has several factors that can affect the production of biogas, namely the temperature in the digester, the growth of microorganisms, inhibiting agents, etc. Previous research that examined cow dung as a raw material for power plants still used potential based calculations, without examining the factors that influence the production of biogas. So that the results obtained are not accurate. A study to calculate the factors that influence the formation of biogas is very necessary because in the process of a biogas power- plant the factors that influence it are very much taken into account in order to obtain optimal results [8-9]. Overcoming the shortcomings in calculations to predict biogas production in previous research has been done by making a mathematical model of each stage of biogas production as a differential and algebraic equation that is simulated in MATLAB software. Modeling is done to make it easier to optimize each process and to control each biogas formation process. In the simulation process, the performance of all stages of biogas that is being in production can be seen so that estimating the biogas production that will be produced from the whole process is more accurate and can optimize the results of biogas production without disturbing the ongoing anaerobic digestion process activities (trial and error).[10]

2. Literature Review

2.1 *Related research*

Before conducting this research, it is necessary to conduct a literature study which aims to find references and research relevant to the research to be carried out. These references are obtained from journals, books or papers related to this research. Research on the biogas power plant with a balloon type digester, this study analyzes the potential of cow dung to generate electricity using a prototype. The balloon type digester is used for the reason that it is simple to install, easy to assemble and assemble and the price is relatively cheap. The results obtained from the prototype are a mixture of cow dung and water with a 1: 1 ratio of 624 liters of gas can be produced from a plastic drum capable of turning on electricity for 35 minutes with a power capacity of 700 watts [11-13].

Research on the modeling of biogas production in batch type reactors using the Hamming predictor-corrector method, in this study analyzes a model of the biogas production process with a batch type digester. The amount of biogas which is produced from biogas production process was predicted using a model that is commonly used in the anaerobic digestion process, namely Anaerobic Digestion Model No. 1 (ADM1) [14]. The ADM1 model is transformed into a system of differential equations and is solved using the Hamming predictor-corrector method. This method is a linear method from the previous points. The simulation of biogas production was carried out for 120 hours by defining the initial substrate concentration of 500 mgCOD / L. On the basis of simulation results, it is known that the maximum concentration of methane obtained at the end of the simulation is 417.48 mgCOD / L. In addition, the growth of microorganisms that digest glucose is faster than the growth of other microorganisms. The simulation results show II-3 that the initial concentration of glucose and microorganisms is very influential on the concentration of methane produced. [15] Research on the simulation of biogas from dairy cow dung, in this study analyzed a simulation model on Matlab for biogas production from cow dung on a dairy farm. Input in this study uses dairy cow dung which is diluted with 25% water after filtering with the output of high-quality fertilizer and biogas consisting of 70-73% methane that is produced from the diluted liquid fraction of dairy cow dung. The model in this study made several modifications based on the hill model to simulate the production of biogas methane in anaerobic digestion. The modified hill model is simulated in Matlab using the eulerian and ode solver methods to obtain changes in methane gas over time. And this study also uses the Matlab editor function block Simulink. The three simulators provide the same response curve with different simulation times [16,17].

Research on the modeling and anaerobic simulation of livestock manure into biogas, this study creates and analyzes a model of biogas production from livestock manure which aims to develop a method for testing the digestion of fertilizers and wastewater for biogas. The model made several additions from the basic anaerobic digestion model no 1 (ADM1), modeling was carried out using Matlab by implementing all the equations and parameters. guidelines to simulate biogas production in certain species [18]. Research on the modeling and simulation of biogas production based on anaerobic digestion of energy crops and manure, this study makes an anaerobic digestion model to improve the accuracy of predicting the dynamics of anaerobic digestion for plants and manure. The model is calibrated using an experimental dataset in a batch process which is mono-fermented corn amylases. Furthermore, the concept is being validated by experimental data in which corn silage has being digested and tested for twenty-eight days in a continuous pilot-scale biogas fermented at uninterrupted raw material loads. The resulting model accurately predicts the flow rate dynamics of CH₄ (methane) and the carboxylic acid concentration. After that, the II-4 calibration model was carried out using ADM1 (Anaerobic Digestion Model no 1) for silage grass and livestock manure. The calibrated model precisely predicts anaerobic digestion from subtract for biogas and methane flow rates, and volumetric concentration dynamics of biomass, carboxylic acid chains, inorganic carbon matter, organic matters, and the pH values. Process modeling in this research uses Matlab [19]. Based on several studies that have been carried out for the calculation of biogas production using mathematical equations based on its potential only and calculated manually; Several supporting studies have carried out the calculation of biogas production by adapting the actual conditions that are implemented in each biogas formation process into a differential equation that is solved using the Matlab simulator. However, this research still focuses on calculating biogas production. The author offers a modeling and simulation in producing biogas and the potential for electrical energy by utilizing cow dung waste. The simulation in this study not only examines the aspect of biogas production, but also involves the potential of electrical energy generated from biogas as well as analyzing the technical and economic aspects. By using modeling and simulation, the author can experiment in complex situations, save money, save time and focus on the important characteristics of the problem compared to the manual method

(trial-and-error) is less effective, and time consuming. Apart from that, modeling and simulation are also useful for analyzing system performance.

2.2 Cow Manure

2.2.1 Definition of Cow Manure

Cow manure is the result of digestion in the form of waste from cows which varies in color from green to black, depending on the food eaten by the cow. After exposure to air, the color of cow dung tends to darken. Cow manure is waste from the digestive process of cattle which is solid and in the process of its disposal it is often mixed with urine and gases, such as methane and ammonia. Nutrient content in cow dung varies depending on the state of the production level, type, amount of feed consumption, and individual livestock [9, 20, 21]; The composition of cow dung that has generally been studied can be seen in table 2.1.

Table 2.1: Composition of Cow Manure.

Compound	Percentage
Hemicelluloses	18.6%
Cellulose	25.20%
Lignin	20.20%
Protein	14.90%
Dust	13%

Specifications of cow dung produced from cows weighing 635 kg, the amount Total solids (TS) can generally also be estimated to be 10-15% of the initial impurity mass. Meanwhile, the number of volatile solids can be estimated at 8-10% of the mass of impurities early [22-24]. The specification of cow manure with a cow weight 636Kg can be seen in table 2.2.

Table 2.2: Cow Manure specifications with a total weight of 635 kg

Specifications	Cows with a weight of 635kg
Dirt	50.8kgs
Manure	51.1liters
Total solids (total solid, ts)	6.35kgs
Volatile solids (volatile solid, vs)	5.4kgs

2.2.2 Potential for Cow Manure

Waste Cattle farming in India has enormous potential which is spread over several regions. Cattle breeding business requires ideal geographical conditions for the survival of cows. Weather in Delhi is good for cattle farming. Cow manure is a potential raw material for making biogas because it contains starch and lignocelluloses. Usually, cow dung is used as fertilizer and the rest is used to produce methane gas using anaerobic processes. Cow manure is a biomass that contains carbohydrates, protein and fat. Biomass that contains high carbohydrates will produce low methane gas and high CO₂, when compared to biomass that contains high amounts of protein and fat. In theory, the methane production resulting from carbohydrates, protein, and fat is 0.37; 1.0; 0.58 m³ CH₄ per kg of organic dry matter. Cow manure contains the three elements of organic matter, so it is considered more effective to convert into methane gas. One way to determine the appropriate organic material to be used as an input for the biogas system is by knowing the ratio of carbon (C) and nitrogen (N) or what is called the C / N ratio. Several experiments that have been carried out by ISAT show that the activity of methanogenic bacteria will be best at a C / N ratio of around 8-20 [25-27]

2.2.3 Biogas

Biogas is a gas produced through anaerobic processes (without oxygen) where the molecules are complex carbon contained in organic matter degraded into molecules with simpler structures including CH₄ and CO₂. India mostly uses biogas for cooking or heating, whereas biogas which contains the main ingredient methane (CH₄) can be used as fuel in power plants because it has a fairly large heating value, which is 23,880 BTU / lbm [28]. Biogas is produced when microorganisms, especially bacteria, reduce levels of organic matter without air or anaerobic conditions. Compared

to air, biogas is about 20% lighter and has a flame temperature between 655° C to 750° C. Biogas is a gas that is odorless, has no color and burns with a blue embers color similar to liquid petroleum gas (LPG). Biogas burns with an efficiency of 60% in the conventional biogas furnace and a calorific value of 20 MJ / Nm³. The volume of biogas is usually expressed in normal units of meters per cubic (Nm³), namely the volume of gas at 0oC and atmospheric pressure. Biogas consists of 50% to 75% methane (CH₄), 20% to 44% carbon dioxide (CO₂) and small amounts of other substances. The biogas composition is as follows [29-33].

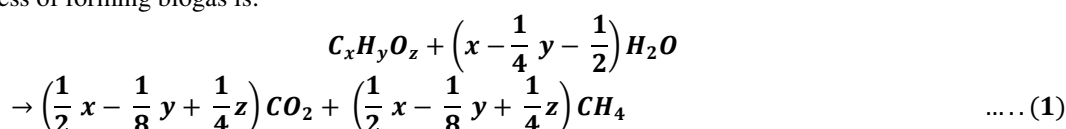
The composition of biogas can be seen in table 2.3.

Table 2.3: Compositional Biogas

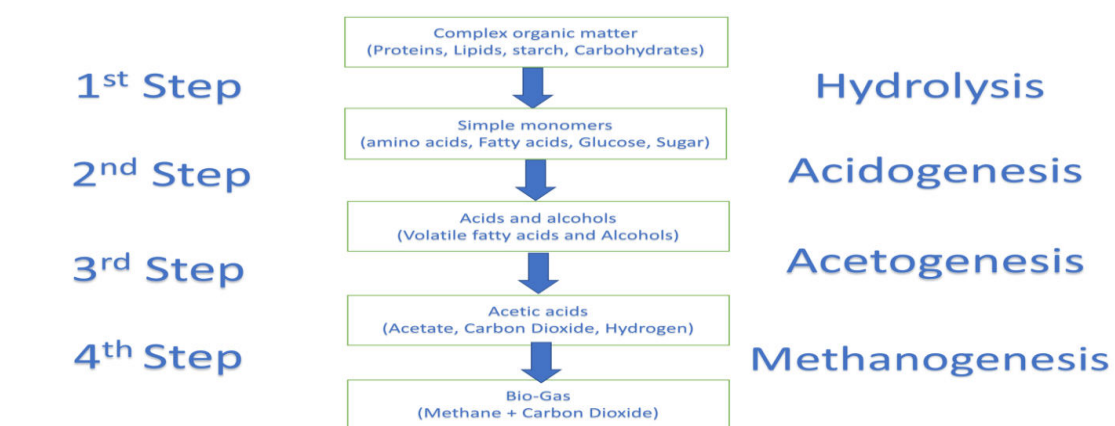
Concentration	Formulas	Elements (Volume %)
Methane	CH ₄	50-75
Carbon dioxide	CO ₂	25-45
Water Vapor	H ₂ O	2-7
Oxygen	O ₂	<2
Nitrogen	N ₂	<2
Hydrogen Fluid	H ₂ S	<2
Ammonia	Nh ₃	<1
Hydrogen	H ₂	<1

2.2.3 Biogas formation Process

The formation of biogas occurs based on chemical principles, namely the occurrence of fermentation of carbohydrates, fats and proteins by methane bacteria which are not mixed with air or what is called anaerobic digestion process. One gram of cellulosic material will produce 825 cm³ of gas at atmospheric pressure. One gram of fat produces 1.25 liters of biogas at atmospheric pressure. The process of forming methane gas by anaerobic digestion involves a complex interaction of several different bacteria, protozoa, and fungi. Some of the bacteria that play a role are Bacteroides, clostridium butyrum-coli and other intestinal bacteria. These two bacteria are the main bacteria producing methane and can live in anaerobic conditions. The fermentation process usually takes 7 to 10 days with an optimum temperature of 35°C and an optimum pH of 6.4-7.9. In general, the process of forming biogas is:

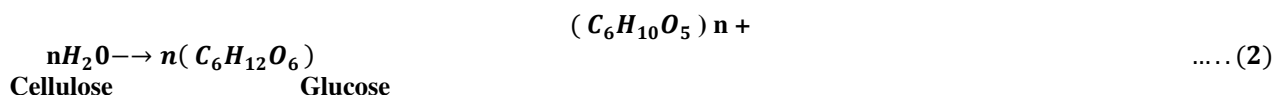


In the formation of biogas, the process consists of the acid hydrolysis step (acidification), and the methanogenesis stage [33-39]. The different stages of biogas formations can be seen in figure 2.1.



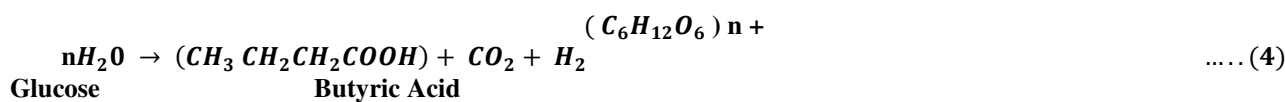
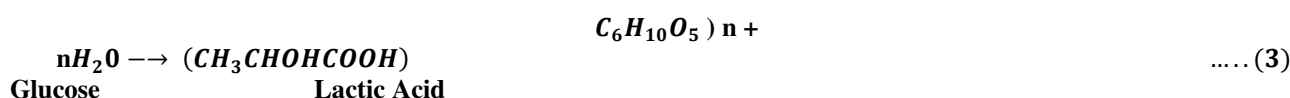
2.2.3 Hydrolysis

At this stage of hydrolysis, it is the breakdown of complex organic materials being simple, changes the structure of the polymer to the monomeric form an insoluble compound with a lighter molecular weight. Lipids turn into acids long chain fat one and glycerin, polysaccharides into sugars (mono and disaccharides), protein into amino acids and nucleic acids, into purines and pyrimidines. Lipid conversion occurs slowly below 20°C. The hydrolysis process requires exo-enzyme mediation excretion by fermentative bacteria. Hydrolysis of molecules is catalyzed by an extra enzyme's cells such as celluloses, lipases, proteases, etc. [33-39].



2.2.4 Acidification

At this stage of acidification, the bacteria will change the polymer simply as a result hydrolysis to acetic acid (CH₃COOH), hydrogen (H₂), and carbon dioxide (CO₂). To converting into acetic acid, bacteria need oxygen and carbon contained in solution. This stage is carried out by obligate anaerobic bacteria and some of them are bacteria facultative anaerobes. These bacteria are anaerobic bacteria that can grow in acidic conditions namely pH 5.5-6.5 which works optimally at a temperature of about 30°C. Acetic acid very much needed which will then be used by microorganisms for formation methane gas. In addition, mixing is necessary for an even metabolism with a water concentration of > 60% [33-39].



2.2.5 Acetogenesis

This acetogenesis stage is an advanced stage of the acidification stage, at this stage about 79% of COD is converted into acetic acid. The formation of acetate depends on the oxidation conditions of the organic matter which are usually accompanied by the formation of CO₂ and hydrogen. Ethanol, butyric acid and lactic acid are converted into acetic acid by acetogenic bacteria. The reaction is as follows [33-39]:



2.2.6 Methanogenesis

This stage of methanogenesis is the stage where methane and carbon are formed dioxide. Methane is produced from acetic acid or from the reduction of carbon dioxide by bacteria acetotropic and hydrogenotropic using hydrogen. Methane producing bacteria have appropriate atmospheric conditions due to the process of acid-producing bacteria. That acid the resulting acid-forming bacteria will be used for methane-producing bacteria. On at this stage low molecular weight compounds are decomposed by methanogenetic bacteria be a compound with a high molecular weight [33-39].

2.2.7 Biogas Formation Process Parameters

The factors that influence microorganisms are very important in determining speed of the biogas formation process, includes the temperature, pH, nutrition, concentration solid, volatile solid, substrate concentration, time of digestion, stirring of ingredients organic as well as pressure influences. The following is a discussion of these factors [33-39]:

1. Temperature

There are three conditions for anaerobic degasification based on the temperature of the digester, including:

- i. Psychrophilic conditions: In these conditions, the digester temperature is between 10-18°C, and liquid organic waste digested for 13-52 days.
- ii. Mesophilic conditions: In these conditions, the temperature of the digester is between 20-45°C, and liquid organic waste digested for 18-28 days. Compared to the digester in thermophilic condition, in mesophilic conditions, the operation is

- iii. Thermophilic conditions: In this condition, the temperature of the digester is between 50-70°C, and liquid organic waste is digested for 11-17 days. In thermophilic conditions it produces a lot of biogas, but the investment costs are high and the operation is complicated. The graph representative of anaerobic digestion temperature can be seen in figure 2.2.

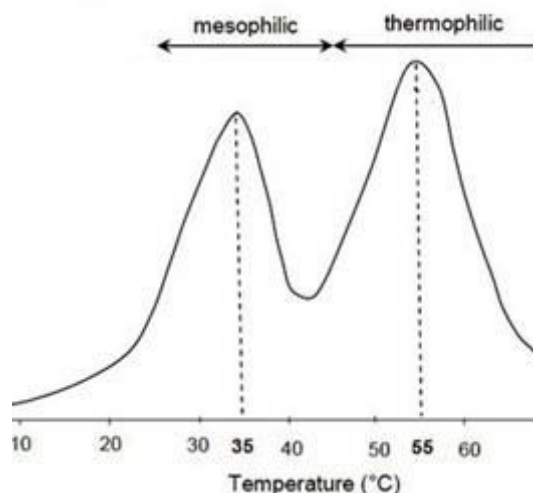


Figure 2.2: Graph Representative Anaerobic Digestion Temperature

- The optimal temperature for anaerobic digestion is temperature 30-35° C. These temperature range combines the best situations for bacterial growth and production of methane gas in the digester for a long-time short process. At 35° C, it will digest the mass of the material same will be digested twice as fast as a temperature of 15° C and produces nearly 15 times as much gas at the same processing time. As with the biological process, the methane gas production increases for each increase in temperature of 11-15° C. In other words, the number the total amount of gas which is produced in a fixed amount of material increases with each other with increasing temperature [33-39].
- iv. Degree of acidity (pH): In anaerobic decomposition, pH is a factor that affects microbes so that if the pH in the digester is not in accordance with the recommended pH range then the microbes cannot grow to the maximum. It can even cause microbial death which in turn will inhibit methane gas production. Anaerobic bacteria require an optimal pH between 6.2 - 7.6, but that is best is 6.6 - 7. At first the media has a pH of ± 6 then it rises up to 7.5. If the pH is smaller or greater, it will have toxic properties against methanogenic bacteria. When the anaerobic process is already on its way towards the formation of biogas, the pH ranges from 7-7.8. PH control is carried out naturally by the NH_4^+ and HCO_3^- ions. These ions will play a role in determining the pH value [33-39].
- v. Solids Concentration Factor (Total Solid Content / TS): Total solid content is the amount of solid material present in waste in organic material during the digester process occurs, which indicates the rate of destruction / decomposition of solid organic waste materials. Ideal concentration solids for producing biogas are 7-9% dry content, this condition can make the anaerobic digester process run well. It should be noted that TS concentrations should be kept at no more than 15% as it will inhibit metabolism. When introducing organic material into the biodigester must be added with a certain amount of water, the function of the water here is in addition to maintains TS <15%, also to simplify the mixing process, the process of flowing organic material into the biodigester and for facilitates that the gas stream formed at the bottom can flow to the passage over the biodigester [33-39].
- vi. Volatile Solids (VS): VS or volatile solids is part of the TS solids that change into the gas phase at the acidification and methanogenesis stages as in the process fermentation of organic waste. In laboratory scale testing, the current weight is part the solid organic material is burnt out in the gasification process at a temperature of 538 °C called volatile solid. The following is a table of volatile solids (VS) components. The volatile solid components can be seen in table 2.4

Table 2.4 Volatile Solid Components

Component	TS%
Cellulose	31
Hemicelluloses	12.2
Lignin	12
Kanji	12.4
Protein	12.6

Ether	2,6
Ammonia	0.5
Acid	0.1
Total	83.4

It can be seen from the table above that the components of volatile solids (VS) generally consist of cellulose, hemicellulose, lignin, starch, protein, ether, ammonia and acids. The size of VS is about 83.4% TS. Taking into account that the TS from animal feces is not far from 10%, it is necessary to add some animal food waste in the biodigester, apart from containing high C / N it also has the potential for high biogas production because it contains high TS [33-39].

vii. **Duration of the Digestion Process:** The duration of the digestion process (Hydraulic Retention Time) or HRT is the amount of time (in days) the digestion process in the anaerobic tank counts from the entry of organic matter to the initial process of forming biogas in the anaerobic digester. From the biogas generation as a whole HRT covers 70-80% of the total time. The total time of HRT depends on the type of organic material and the treatment of organic matter before the digestion / digester process is carried out. If too much volume of material is inserted (overload) it results in the filling time being too short, the raw material will be pushed out while gas is still produced in small quantities.

viii. **Carbon Nitrogen (C / N) Ratio** Anaerobic processes will be optimal if given food ingredients containing carbon and nitrogen simultaneously. Carbon is needed to supply energy while nitrogen is needed to form the structure of bacterial cells. The C / N ratio shows the ratio of the sum of the two elements. For materials that have a carbon amount of 15 times the amount of nitrogen will have a C / N ratio of 15 to 1. The C / N ratio with a value of 30 (C / N = 30/1 or carbon 30 times the amount of nitrogen) is a digestion process at an optimum level, if other conditions also support. The process will run slowly if there is too much carbon, because nitrogen will run out first. Conversely, if there is too much nitrogen (low C / N ratio; for example, 30/15) then the carbon will run out first and the fermentation process will stop. One study showed that the metabolic activities of methanogenic bacteria would be optimal at the C / N ratio of 8-20 [40]. The following is a table showing the C / N ratio of some organic materials in common use:

ix. **Volatile Solids (VS):** VS or volatile solids is part of the TS solids that change into the gas phase at the acidification and methanogenesis stages as in the process fermentation of organic waste. In laboratory scale testing, the current weight is part the solid organic material is burnt out in the gasification process at a temperature of 538 °C called volatile solid. The following is a table of volatile solids (VS) components. The C/N ratio of organic materials can be seen in table 2.5.

Table 2.5 C/N Ratio of Organic Materials

RAWH MATERIAL	C/NH RATIO
Human Decoration	8
Goat Dung	121
Sheep Dung	191
Corn Waste	601
Wheat Waste	901
Duck Waste	8
Chicken Poop	101
Pig Dung	181
Cow Dung	241
Dirt Gajah	43
Rice Waste	7
Saw Dust	2

x. **Stirring of Organic Materials:** Stirring is very beneficial for the ingredients in the anaerobic digester, which provides the opportunity for the material to remain mixed with bacteria and to maintain an even temperature throughout the digester. With stirring, it will minimize the potential for material which is settling at the bottom of digester and the concentration is firmly distributed, and the potential for all materials to undergo an anaerobic fermentation process is greater. In large digesters the mixing system is very important. The purpose of stirring is to keep the solid material away from settling on the bottom of the digester. In addition, stirring can facilitate the release of gas produced by bacteria to the biogas reservoir [33-39]. Effect of Pressure has an important role, the higher the pressure in the digester, the lower the biogas production in the digester, especially in the hydrolysis and acidification processes. The pressure is maintained between 1.15-1.2 bar in the digester [33-39].

- xi. Toxic and Inhibitor Compounds The anaerobic fermentation process of inhibiting compounds or inhibitors can be divided into 2 types, namely physical inhibitors and chemical inhibitors. Physical inhibitors are temperature and chemical inhibitors, also known as toxins, include heavy metals, antibiotics and volatile fatty acids (VFA) [33-39].

2.2.9 Equations for the Formation of Biogas

The following are some of the equations that determine the process of biogas formation from the fermentation of organic waste in anaerobic digester [34-36]. The theoretical decomposition time equation is the time the organic material is in the digester tank. When this process occurs, the growth of anaerobic bacteria decomposes, the process of decomposing organic matter, and stabilizes the formation of biogas to its optimum conditions. Overall, the hydraulic retention time or HRT covers 70% -80% of the total biogas formation time if the biogas formation cycle is idea, time the process of introducing organic matter directly obtains biogas as the final process without adding organic material again [33-39]. HRT can be formulated into the following equation:

$$HRT (days) = \frac{Volume\ Digester\ (m^3)}{Daily\ Organic\ Ingredient\ Addition\ Rate\ \frac{m^3}{day}} \dots (8)$$

If the dry solid material is DM (Dry Material) or it is also called Total Solid (TS) ranges from 4-12%, so the optimum breakdown time (Optimum Retention Time) ranging from 10-151 days. If the Dry material value is greater than 1 the percentage value of material solids dries above, it means that the organic matter has a denser concentration so it takes a long-time breakdown time becomes specific, so that1 the length of time equation applies the following specific retention time or SRT:

$$SRT = \frac{Organic\ Solids\ in\ Anaerobic\ Digester\ (kg)}{Daily\ Organic\ Ingredient\ Addition\ Rate\ \frac{kg}{day}} \dots (9)$$

For specific organic matter as above, the rate of addition of organic waste (Specific Loading Rate) or SLR can be seen as follows:

$$SLR = \frac{(kg\ ODM)}{m^3 - day} = \frac{Added\ Organic\ Ingredients(kg\ \frac{ODM}{day})}{Volume\ Digester\ (m^3)} \dots (10)$$

The depth of the digester1 tank greatly affects the SLR value and when the parameters otherwise it can be maintained in ideal conditions, the maximum SLR values obtained range from 3-6 kg ODM / m³-day.

- *Specific Biogas Production Equations*

Specific Biogas Production (SBP1) is a digester efficiency indicator value. Minimum conditions are 1.5 and the ideal target is 2.5.

$$SBP\ (day - 1) = \frac{Biogas\ Production\ (m^3\ /day)}{Volume\ Digester\ (m^3)} \dots (11)$$

- *Specific Methane Production Equations*

Methane Production Specific (Specific Methane Production) or SMP, relates to the total energy produced against that energy potential owned organic waste (feedstock). For organic waste from plants / seeds energy value between 0.3 - 0.4 (%) and for some types of animal waste can value up to 0.8%;

$$SMP\ (m^3\ CH_4/kg\ ODM(day - 1)) = \frac{Volume\ Gas\ (\frac{m^3}{day})}{Organic\ Material\ Addition\ Rate\ (kg\ ODM-day)} \dots (12)$$

2.2.10 Equations for the Forming energy to power

The conversion of biogas energy for electric power generation can be done using several technologies, namely, gas turbines, microturbines and the Otto Cycle Engine. The need for biogas, such as gas concentration of methane and biogas pressure, load requirements and availability of available funds, are significantly affected by the choice of this technology. [41-50]. In the book Renewable energy conversion, transaction and storage by Bent Sorensen, that 1 kg of methane gas is equivalent to 6.13 x 10⁷ J, while 1 kWh is equivalent with 3.6 x 10⁷ Joule. For a gas density of 0.656 kg / m³, so that is 1 m³ methane gas produces 11.17 kWh of electricity. Then it can be assumed that the conversion potential biogas into electrical energy as follows:

$$W = CH_4 * 11,17 \text{ kWh}$$

(13)

.....

Where:

W = Electrical energy that can be produced (kWh)

CH_4 = Methane (M^3)

The Energy conversion from methane gas to electrical energy can be seen in table 2.6.

Table 2.6 Energy Conversion from Methane Gas to Electrical Energy

Energy Type	Energy Equivalent
1 kg of Methane Gas	$6.13 * 10^7 J$
1 kWh	$3.6 * 10^6 J$
1 m ³ of Density Methane Gas Methane gas is 0.656 Kg / m ³	$4.0213 * 10^7 J$
1 m ³ of Methane Gas	11.17KWh

2.2.11 Equations for the Formation of Biogas

A biogas generator set or known as a biogas generator is a tool which has two main components, namely the engine and the generator that uses it biogas to produce electricity. This driving machine can be moved because of the combustion of gas fuel in it which is then used for generate mechanical energy. In the presence of mechanical energy which is coupled with generator, then there is a conversion of mechanical energy into electrical energy [51-57]. The following are the main components of a Biogas Generator:

1. Compressor is a mechanical power generator that functions to generate heat energy comes from the atmospheric air to meet the needs of the process combustion gas in the gas turbine combustion chamber. In the process of operation, the compressor is supported by tools, namely the intake air filter and the inlet gate fan.
2. Combustor is the combustion chamber which is the generator of heat energy from the process gas fuel. In the operation process, the combustor is supported by assistive tools namely gas station, control system, fuel nozzle, and igniter system (ignition system).
3. Gas Turbine is a mechanical energy generator from the heat energy conversion process into kinetic energy and further into capable mechanical energy drive the turbine shaft with mass gas burning fuel. In the operating process, the gas turbine is supported by tools, namely a lubricating oil system, control oil system, turning motor, pony motor, starting motor, cooling water system, exhaust duck system, and turbine supervisory instrument.
3. Generator is a tool to convert mechanical energy from the turbine shaft into electrical energy.

The Biogas generator image can be seen on figure 2.3:



Figure 2.3: Biogas Generator

2.2.12 Economic Model

○ Total Net Present Cost (NPC)

NPC is the value of all costing that are incurred during the lifetime, less the present value of all income earned over the lifetime. This Costs include capital costs, replacement costs, O&M costs, fuel costs, emission fines, and power purchase costs from the grid. Economic models for the HOMER simulation use the Net Present Cost (NPC) which is the total cost of installing and operating the system during the project lifetime [58-65]. Net Present Cost (NPC) itself can be calculated using the following equation:

$$NPC = \text{Capital Cost} + \text{Replacement Cost} + \text{O\&M Cost} + \text{Fuel Cost} + \text{Salvage}$$

Where:

Capital Cost = Cost of capital components (Rs)

Replacement Cost = Cost of component replacement (Rs)

O&M Cost = Operational and maintenance costs (Rs)

Fuel Cost = Fuel cost (IDR)

Salvage = Cost remaining on components (Rs)

○ Cost of Energy (COE)

The Cost of Energy (COE) is the costing that is being required to produce each 1 kWh of electrical energy, that is, the result of dividing the annual costs and energy production annual. The COE value of each scenario uses the following equation:

$$COE = \frac{TAC}{E_{tot,served}} \quad \dots\dots (14)$$

Where:

$E_{tot,served}$ = Total annual energy used to serve the load(kWh)

TAC = Total Annualize Cost or total annual costs incurred for generating reserves. (Rs.)

3. Proposed Methodology

This research is a simulation research that aims to find a picture through a simple or small-scale system (modeling), in the model will be made changes to variable or control to see the effect. Research simulation aims to provide an overview of the application of a technique through process modeling so that the technique will not suffer unexpected losses prior to its application. This research was conducted using a mathematical model has been made in previous studies and continues the study of research previously, namely utilizing biogas into electrical energy. The conversion process to be done through a simulation to get a model that fits the case or the problem is there. In this research will also analyze technical aspects and economy, namely NPC (Net Present Cost) and COE (Cost of Energy). In this study, it was started from a literature study related to previous research to support the research that will be carried out. Furthermore, potential data collection raw materials for cow and buffalo dung in National Capital Region, then doing modeling and simulation of biogas production from cow dung using the anaerobic method digestion. Furthermore, modeling and simulating the conversion of electrical energy from biogas production that has been obtained from previous simulations. After that he will do it analysis of electrical and economic energy resulting from the modeling process that has been done. The stages to be carried out in this research are in accordance with figure in 3.1.1 and 3.1.2:

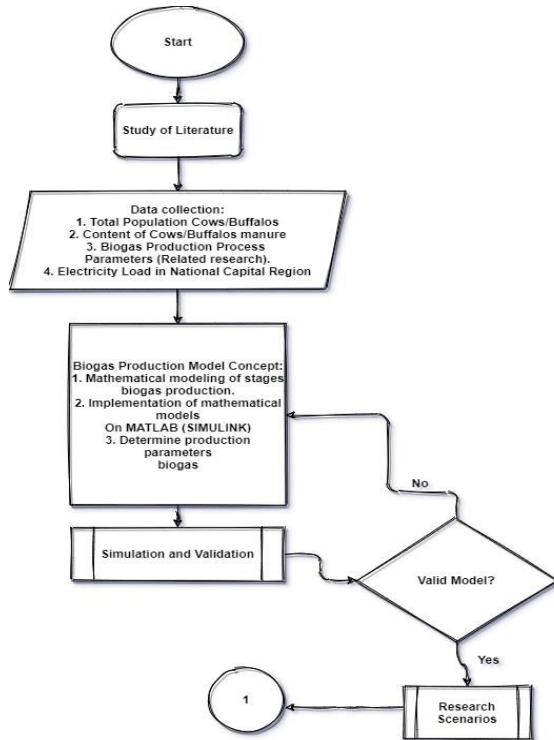


Figure 3.1.1: Research Flowchart

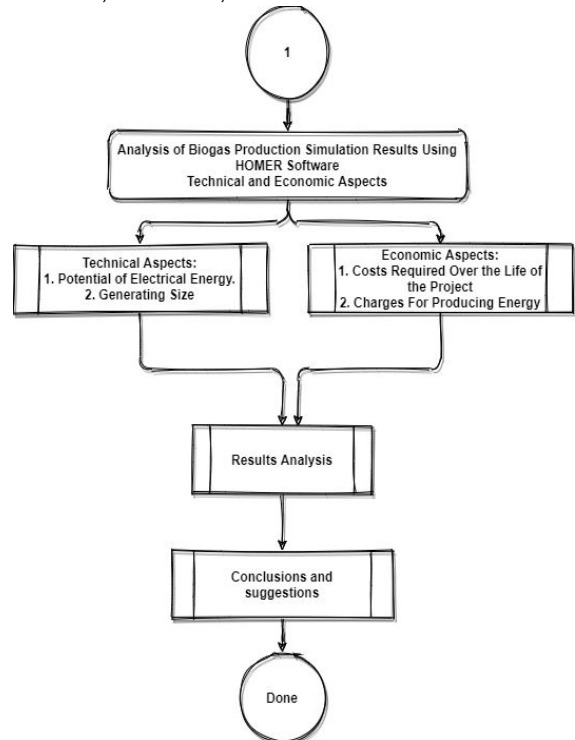


Figure 3.1.2: Research Flowchart continued

○ Literature Study

This stage collects some of the research needed to be used references in this research, namely Final Project, Thesis, books and journals. On each Related research will be analyzed in the theory used, research methods and results. In the book you will find a theory that supports this research in better results.

○ Data Collection

This research uses modeling and simulation methods with an approach literature using MATLAB software. Therefore, data is needed to support it this research. The data required consists of:

1. The originating statistics for the population of cattle cow and buffalo in National Capital Territory. This data is used to obtain potential cow & buffalo dung that will be produced.
2. Data on the content of cow & buffalo dung will be used as input parameters simulation of biogas production.
3. Anaerobic digestion process parameter data from related research (Experimental results and related journals)
4. National Capital Territory electricity load data which will be used to analyze technical aspects.

* Mathematical Modeling of Biogas Production Stages

This simulation research needs to determine the stages of the intermediate biogas production process other hydrolysis, acidogenesis, acetogenesis, and methanogenesis to be carried out on Matlab worksheet.

○ Hydrolysis Process Modeling

The first stage of the anaerobic digestion process is the process of hydrolysis. The process can be interpreted as the rate of change during the anaerobic digestion process in the biodegradable concentration of volatile solids (BVS) in the reactor. This mechanism depends on the feed material type, the flow rate of the feed, the effective reactor volume and the reactor temperature. The following equation is the hydrolysis process:

Defines the portion of raw waste that can serve as a substrate:

$$= B_o \cdot Svs_{in} \quad Sb_{in} \quad \dots (15)$$

Defines the portion of the biodegradable raw material which is originally in acidic form:

$$= A_f \cdot Sb_{in} \quad Sv_{in} \quad \dots (16)$$

Mass balance of biodegradable volatile solids:

$$\frac{d(Sb)}{dt} = (Sb_{in} - Sb) \left(\frac{F_{feed}}{V} \right) + \frac{\mu_m \cdot K_1 \cdot X_{acid} \cdot X_{meth}}{\frac{K_s}{S_b} + 1} \quad \dots (17)$$

Where:

B_o = Biodegradability Constant $\left(\frac{kg \text{ BVS} / m^3}{kg \text{ VS} / m^3} \right)$

A_f = Acidity Constant $\left(\frac{kg \text{ VFA} / m^3}{kg \text{ BVS} / m^3} \right)$

Sb = biodegradable volatile solid concentrations in (kg / m^3)

Sb_{in} = The concentration of biodegradable volatile solids in the reactor feed (kg / m^3)

F_{feed} = Feed flow rate (m^3 / day)

K_1 = Yield factor

K_s = Half Monod constant velocity for acidogens (kg / m^3)

X_{acid} = Concentration of acidogens (kg / m^3)

X_{meth} = Methanogens concentration (kg / m^3)

μ_m = Max. growth rate of acidogens (d^{-1})

V = Reactor volume (m^3)

The maximum growth rate for methanogens that can be expressed as a function of the temperature dependence of the reaction rate using the following empiric:

$$\mu_m T_{react} = \mu_{mc} \cdot (T_{react}) = 0.013 \cdot T_{react} - 0.129 \quad \dots (18)$$

Where:

μ_{mc} : Maximum growth rate for methanogens (d^{-1})

T_{react} : Reactor temperature $(^{\circ}C)$.

○ Acidogenesis Process Modeling

The acidogenesis stage is the rate of change in the concentration of volatile fatty acids during the fermentation process. This process depends on the total concentration of VFA in the reactor (type of feed material), feed flow rate, volume effective reactor, and reactor temperature. The following equation is the acidogenesis process:

$$\frac{d(Sv)}{dt} = (Sv_{in} - Sv) \left(\frac{F_{feed}}{V} \right) + \frac{\mu_m \cdot K_2 \cdot X_{acid}}{\frac{K_s}{S_b} + 1} + \frac{\mu_{mc} \cdot K_3 \cdot X_{meth}}{\frac{K_{sc}}{S_v} + 1} \quad \dots (19)$$

Where:

Sv = Total volatile fatty acid concentration in the reactor (kg / m^3)

Sv_{in} = The total volatile fatty acid concentration in the reactor feed (kg / m^3)

K_2 = Yield factor

K_3 = Yield factor related to methane gas growth rate

K_{sc} = Half Monod constant velocity for methanogenesis (kg / m^3) .

○ Acetogenesis Process Modeling

The 3rd stage of the AD process is the acetogenesis process. This process depends on the acidogens concentration, the type of feed ingredient, the feed flow rate, the effective reactor volume and the reactor temperature. The following equation is the process of acetogenesis:

$$\frac{d(X_{acid})}{dt} = \left[\frac{\mu_m}{\frac{K_s}{S_b} + 1} - K_d - \frac{F_{feed}/V}{b} \right] \cdot X_{acid} \quad \dots (20)$$

Where:

b = Retention time factor estimated
 K_d = Acidogens specific mortality rate (d^{-1}).

3.8 MODELING OF THE METHANOGENESIS PROCESS

The methanogenesis stage determines the concentration of methanogens which is used to produce methane. This process depends on retention time, the feed flow rate, the effective reactor volume and reactor temperature:

$$\frac{d(X_{meth})}{dt} = \left[\frac{\mu_m}{\frac{K_{sc}}{S_v} + 1} - K_{dc} - \frac{F_{feed}/b}{V} \right] \cdot X_{meth} \quad \dots (21)$$

Information:

K_{dc} = Specific mortality rate from methanogens (d^{-1})
 X_{meth} = concentration of methanogens presents in (kg/m^3)

The equation for the amount of methane output is as follows:

$$F_{meth} = V \cdot \frac{\mu_{mc}}{\frac{K_{sc}}{S_v} + 1} - K_4 \cdot X_{meth}$$

Information:

K_4 = Yield factor related to methane gas flow.

- Implementation of a Mathematical Model of Biogas Production Using Matlab (Simulink)

After all the data is obtained, anaerobic digestion system modeling simulation is performed at SIMULINK. This modeling is done based on existing mathematical modeling. Implementation of mathematical models in Matlab can use text code or SIMULINK. SIMULINK uses blocks contained in the SIMULINK library, these blocks function as mathematical functions, each block has a different function such as add, subtract, multiply, divide, integral etc. Biogas production parameters will be entered into each block according to the respective parameter values.

• Reactor Modeling

SIMULINK block shape shown in Figure 3.2 is a form of model based on combining all mathematical models of the biogas production process into a system.

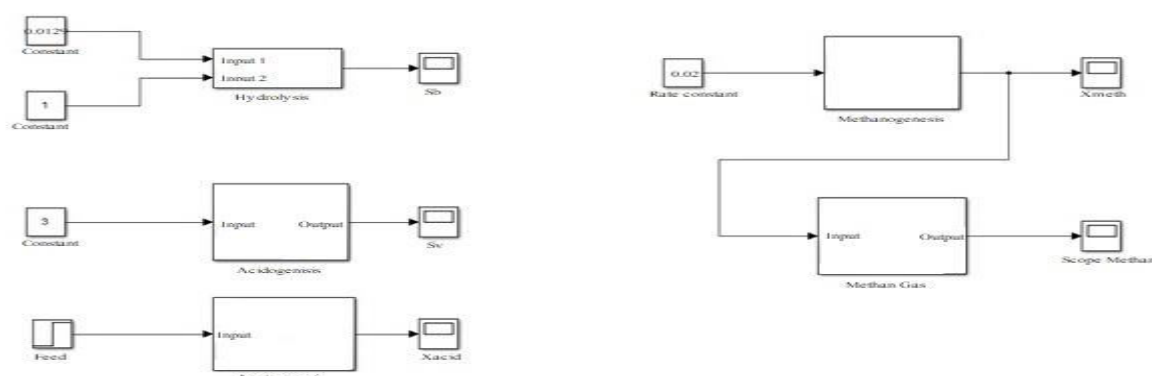


Figure 3.2: SIMULINK Reactor Model

• DETERMINING BIOGAS PRODUCTION PARAMETERS

The rate of the biogas formation process which is largely depend on the factors that affect microorganisms, including temperature, pH, nutrients, solids concentration, volatile solids, substrate concentration, digestion time, stirring of organic matter and the influence of pressure. The following are the parameters for each biogas production process:

• Simulation and Validation

After modeling, simulation testing of the system that has been modeled is carried out. Simulations were carried out using SIMULINK MATLAB R2014a. To run the simulation on Simulink by selecting the run button to run the created model. If, when running there is an error, a diagnostic viewer will appear which will explain the errors that occurred in the system. Then there will be a re-evaluation of the system model until there are no more errors. To see the results of the model, by selecting the Scope Fmeth block, a graph will appear which is the result of the model created. After obtaining the results from the system model that is made, verification is carried out. Whereas, validation is an action to prove a research is correct with a benchmark so as to achieve the desired results. The purpose of validation is to ensure the results of a study are close to real or maintain the credibility of proposed scheme.

In this study, validation was carried out using a mathematical equation from the study entitled "Simulation of a Biogas Reactor for Dairy Manure" then the initial step of modeling will use the same variables and parameters from the research above, then see the results of the model made whether it is in accordance with the results of the reference research or not, if appropriate, the research can be continued to the next stage. Verification aims to ascertain whether the model is suitable or not. The suitability refers to the standard theory of conversion of potential cow dung into biogas, namely 1 kg of cow dung (cattle dung) produces **0.023-0.04 m³** of biogas. If, the results of the system model made are in accordance with the reference standard, then change the input parameters according to the case will be examined, the parameters that are changed are F_{feed} (input feed), T_{react} (temperature), and V (digester volume). If the results of the model made deviate too far from the reference standard, then re-evaluate the model. After the results are in accordance with the reference standard, you can proceed to the next stage.

4. Result and Simulation

• Calculation of the Potential of Biogas in a Dairy Cattle

In calculating a biogas power plant from cow dung, a farm location is required to obtain the basic model of the generator. This calculation uses a model in the dairy cows/buffalo's dung as an example of a potential case as proposed in section below.

• Potential hypothesis for the Biogas Plant

The fixed dome type biogas reactor is designed for 10 cows (with 25 kg / day / head of cow dung and 45 days of retention time) with a reactor capacity of 18 m³. Based on the results of the Matlab and Homer test of these activities and literature references as shown in the following table 4.1: -

Table 4.1 Sample performance of the biogas plant

Description	Referenced	Test Results and Analysis
1. Material Conditions (Cow Dung)		
• Total Solid, Kg/ Head / Day	4.8	4.2
• Volatile Solid, Kg / Head / Day	3.9	3.8
• Water Content, %	6.9 – 8.9	13.59
• C/N Ratio	1:20 1:30	1:17
• Cod, Mg/Ld	-	19 800
• -BOD/COD	-	0.06
2. Conditions in The Reactor (Process)		
- Temperature, °C	35	25–27
- pH	7.0 - 8.01	7-8.61
3. Chemical Content of Biogas		
- CH ₄ , %	501-1601	76.131
- CO ₂ , %	301-1401	21.881
- H ₂ S, µg / m ³	<1%	1543.461
- NH ₃ , µg / m ³	-	40.121

4. Discharge sludge Condition from the Reactors (Effluent) - Cods - Bod/Cods - Nutrient content (Main), - % • Nitrogen • Phosphorus • Potassium	5001-12500	1 960
	0.51	0.37
	1.451	1.82
	1.101	0.73
	1.101	0.41
5. Performance • Lighting, m ³ / Hour • Gas Stove, m ³ / Hour	10.11 - 0.151 (Illumination Equivalent To 60 Watts of Bulbt≅T100 Candles power≅ 620lumens). Pressure: 170-851 Mmmh ₂ o1 0.2 - 0.45 0.3 m ³ / Person/Day Pressure: 75-901 Mmmh ₂ o	0.15-0.3 Pressure=30-60 mmh ₂ o 0.2-0.4 Pressure=60-85 mmh ₂ o

From the existing data, we try to calculate the biogas capacity generated from the existing potential, the percentage of Total Solid and Volatile Solid obtained is the sample of 20 kg / day cow dung as: -

%Total Solid = 14.2 kg / head / day: 120 kg / head / day = 121 %

%Total Volatile Solid = 13.8 kg / head / day: 120 kg / head / day = 119 %

So, for the example Delhi Capital Region, which produces (8-20 kg/head/day), taking 15 kg / head / day and considering data for 150 cattle as: -

Total Solid = 121% * 115 kg / head / day * 1501 = 472.51kg / day

Volatile Solid = 119%*1 15 kg / head / day * 1501 = 427.5 kg / day

Based on the potential mentioned above for biogas for cow dung as: -

Potential Biogas Volume = 0.04 m³/kg *2,250 kg/ day = 90 m³/ day

K = Volume of biogas production as: -

V_S = 190 m³ / day:427.5 kg / day

K = 121% m³/ kg ≈ 121%

• Methane Production Calculation

Energy production using biogas is proportional to the amount of methane gas production. With a known biogas production value (V_{BG}) of 90 m³ / day and by using table 4.3 (depicted under); then it can be seen that the production of methane gas (V_{MG}) is,

$$V_{MG} = 65.7\% * V_{BG} = 65.7\% \times 90 \text{ m}^3 / \text{day} = 59.13 \text{ m}^3 / \text{day}.$$

• Electrical Energy Productions calculations:

With the known volume of methane gas produced, namely 59.13 m³ / day, and Conversion Factor (F_K) (m³ of methane gas is equivalent to 11.17 kWh), so that the potential for electrical energy produced is, E = V_{MG} x F_K = 59.13 * 11.17 = 660.428 kWh / day. The power generated by the Biogas Power Plant is the energy generated per day divided by 24 hours, namely:

P = (E / 24) *time = (660.428 / 24) *24 = 660.428kW ≈ 0.660 MW So from the calculation of the available potential data, the following results are obtained: and presented in table 4.2: -

Table 4.2: Result of biogas capacity calculations

No.	Type of the Calculation involved in the Process	Calculated Results
1.	Potential of Cow Dung (Q)	15 Kg / Day

2.	Calculation of the sum Of Total Solids (Ts)	4.725 Kg / Day
3.	Calculation of The Amount of Volatile Solid (Vs)	4.275 Kg / Day
4.	Calculation of The Volume of Biogas Production (V_{bs})	90 m ³ / Day
5.	Calculation of The Volume of Methane Gas (V_{gm})	59.13 / Day
6.	Calculation of Potential Electrical Energy (E)	660.428 kWh / Day
7.	Power Generating from the Power Plant (Biogas Power)	660.428 Kw in a Day

○ **Digester Design**

Digester Type and Dimension Design

From the available potential it is possible to design a digester for produce biogas. As explained in chapter II a design digester there are several considerations that must be considered. design digester with consideration of several aspects as follows: -

- **Temperature**

For countries such as India, an unheated digester is used for soil temperature conditions of 20 - 30 ° C (Mesophilic - 20 – 40° C).

- **Degree of Acidity (pH)**

Bacteria thrive in moderately acidic conditions (pH between 6.6 - 7.0) and pH should not be below 6.2. Therefore, the main key in the operational success of the biodigester is to keep the temperature constant (fixed) and the material input accordingly. Filling material C / N ratio - The ideal requirement for the digestion process is C / N = 20 - 30. Therefore, to obtain high biogas production, it is necessary to extract carbon (C) materials such as straw, or N (for example: urea) to achieve a C / N ratio = 20 - 30. Based on the data obtained, Cow dung has C / N = 24 so that it is sufficient for the process to obtain the required pH.

- **Digester Design**

As initial data, the potential for dairy cow dung is 2,250 kg / day. In simple terms, the sequence of biodigester facility design begins with the calculation of the volume of the biodigester which includes the potential of the raw materials present in producing methane gas, determining the biodigester model, designing storage tanks and ending with determining the location. The digester used in this plan uses the fixed dome type or fixed dump digester type. This model is the most popular model in Indonesia, where all digester installations are made in the ground with a permanent construction. Besides being able to save land space, making a digester in the soil is also useful for maintaining a stable digester temperature and supporting the growth of methanogenic bacteria. Digester of this type has advantages Low construction costs due to simple construction and long life. The digester uses a flow type, where the raw material flow is entered and the residue is removed at certain intervals. The length of time the raw material is in the digester reactor is called the retention time (RT). The construction parts in this type of digester include:

- a. Gas storage room (gas collecting chamber).
- b. Gas storage chamber.
- c. Fermentation Chamber Volume.
- d. Hydraulic Chamber Volume (hydraulic chamber).
- e. Volume of the sludge layer.

Furthermore, the size and type of digester used will be designed based on the potency and available literature data.

Digester size planning is seen from the daily amount of cow dung, the ratio of the composition of the mixture of water and cow dung, digestion time and the volume of biogas produced. The daily amount of manure produced at Capital is approx. 2.25 tons or 2,250 kg while the composition of the mixture of water and organic waste is to obtain 8% solids, solids refer to the amount of Kg t_s (total solid). Based on the calculations above, the total solid produced is 472.5 Kg. To obtain water that is added to make biogas raw material, fresh cow dung is mixed with water in a ratio of 1: 1. So that the amount of water added = the potential amount of cow dung = 2,250 kg / day then; Q_t = 4,500 kg / day Based on available data storage time (HRT) of cow dung in the digester. Storage time depends on the ambient temperature and biodigester temperature. With tropical conditions like India, at a temperature of 25- 35 ° C, the digestion time is approximately 25-35 days, a short digestion time can reduce the volume of the digester and vice versa the digestion time length can increase the volume of the digester. By determining the digestification time is 30 days, then with equations given below, can be determined the working volume of the digester, where the working volume of the digester is the sum of the digestification room volume (V_{DR}) and the storage volume ($V_{Storage}$),

namely: -

The working volume of the digester = $V_{DR} + V_{Storage}$, where $V_{Storage} + V_{DR} = Q_t \times \text{HRT}$ (digestion time), then:

$$\begin{aligned} V_{DR} + V_{Storage} &= Q_t \times \text{HRT} \\ &= 4,500 \text{ Kg / day} \times 30 \text{ days} \\ &= 1,35,000 \text{ Kg} \end{aligned}$$

Because approximately 80% of the total Q (raw material) is water, we assume the density Q (raw material) \approx density of water (1000 kg / m³)

$$\text{Volume} = m / \rho$$

$$V_{DR} + V_{Storage} = 1,35,000 \text{ Kg} / 1000 \text{ kg / m}^3 = 135 \text{ m}^3 = V_1$$

Based on table 4.3 the assumption of geometrical equations for the size of the digester tank is obtained:

$$V_{Storage} + V_{DR} = 80\% V_1 \text{ or } V_1 = (V_{Storage} + V_{DR}) / 0.8$$

$$V_1 = 135 / 0.8$$

$$V_1 = 168.75 \text{ m}^3$$

If building a digester size of 168.75 m³ apart from being impractical in maintenance it is also less possible due to limited land, so look for a much smaller size digester with more than 1 digester, making it possible for maintenance and if there is damage to one of the digester others are still able to produce biogas as fuel for their electricity generation. It is determined that the digester to be built is 50 m³ in size so that the number of digester sizes that must be built is:

$$\text{Number of digester} = 168.75 \text{ m}^3 / 50 \text{ m}^3 = 3.375 \approx 4 \text{ digester pieces.}$$

For the digester size (V) 50 m³, by reviewing the geometric equation assumptions in Table 4.3 it is obtained:

$$V_{Storage} + V_{DR} = 80\% V$$

$$= 80\% * 50$$

$$= 40 \text{ m}^3$$

$$\text{Volume of gas storage room (Vc)} = 5\% * V = 5\% * 50 = 2.5 \text{ m}^3$$

$$\text{The volume of the sludge storage layer (V}_{\text{Sludge storage}}) = 15\% * V$$

$$= 15\% * 50 = 7.5 \text{ m}^3$$

$$\text{Storage volume (V}_{\text{Storage}}) = 0.5 (V_{\text{Storage}} + V_{DR} + V_s) C.$$

C is the rate of gas production per m³ per day, based on the table the C value 3.4, for cow dung is 0.21, then:

$$V_{\text{storage}} = 0.5 (V_{\text{storage}} + V_{DR} + V_{\text{sludge storage}}) K.$$

$$= 0.5 * (40 + 7.5) * 0.21 = 4.99 \text{ m}^3$$

From the value of $V_{\text{storage}} = 4.99 \text{ m}^3$ so that it can be seen the value of V_{DR} , namely: $V_{\text{storage}} + V_{DR} = 40$, $V_{DR} = 40 - 4.99 = 35.01 \text{ m}^3$

From the geometric assumption it is also known that $V_{\text{Storage}} = V_H = 4.99 \text{ m}^3$, meaning that the biogas will occupy the entire gas storage space (fixed drump digester type) according to the volume of gas produced. So that the volume of each part of the digester is known, namely:

$$V - \text{Total Digester volume} = 50 \text{ m}^3$$

$$V_c - \text{Volume of the gas collecting chamber} = 2.5 \text{ m}^3$$

$$V_{\text{storage}} - \text{Volume of gas storage chamber} = 4.99 \text{ m}^3$$

$$V_{DR} - \text{Volume of the fermentation chamber} = 35.01 \text{ m}^3$$

$$V_H - \text{Volume of the (Hydraulic chamber)} = 4.99 \text{ m}^3$$

$$V_{\text{sludge storage}} - \text{Volume of the sludge layer} = 7.5 \text{ m}^3$$

- **Process Hydrolysis Acidogenesis, Acetogenesis and Methanogenesis.**

The anaerobic decomposition in biopolymers organic complexes into methane gas are carried out by combined activities microbes. In general, this decomposition can be classified into four reactions, namely: hydrolysis, acidogenesis, acetogenesis and methanogenesis.

- **Process Hydrolysis Modeling**

The hydrolysis stages, the decomposition of the polymeric organic material into soluble monomers, such as carbohydrates (polysarides) are broken down into glucose: -



Therefore, the hydrolysis stride was incorporated into the Anaerobic Digestion Model in which the degradable particulate organic substrate, X_c are partially disintegrated into carbohydrates (X_{CH}), proteins (X_{PR}) and lipids (X_{LI}) and is described by

Equation shown below. The hydrolysis of X_{CH} , X_{PR} and X_{LI} are defined as under: -

$$\frac{dXc}{dt} = -K_{dis}Xc + D_{in}(Xc_{in} - Xc) + k_{dec,x1}X1 + k_{dec,x2}X2 \quad \dots (23)$$

$$\frac{dXch}{dt} = -K_{hyd,ch}Xch + D_{in}(Xc_{in} - Xch) + f_{ch,xc} + k_{dis}Xc \quad \dots (24)$$

$$\frac{dXpr}{dt} = -K_{hyd,pr}Xpr + D_{in}(Xpr_{in} - Xpr) + f_{pr,xc} + k_{dis}Xc \quad \dots (25)$$

$$\frac{dXli}{dt} = -K_{hyd,li}Xli + D_{in}(Xli_{in} - Xli) + f_{li,xc} + k_{dis}Xc \quad \dots (26)$$

In where k_{dis} is the constraint for degeneration process; the macrobiotic substrate application, $S1$, incorporate the conditions interrelated to hydrolysis, as shown:

$$\begin{aligned} \frac{dS1}{dt} = & D_{in}(S1_{in} - S1) - (k_1\mu_1X1) + k_7(k_{dis}Xc - k_{dec,x1}X1 - k_{dec,x2}X2) + k_8\{(k_{hyd,ch}Xch - f_{ch,xc}k_{dis}Xc) \\ & + (k_{hyd,pr}Xpr - f_{pr,xc}k_{dis}Xc) \\ & + (k_{hyd,li}Xli - f_{li,xc}k_{dis}Xc) \} \end{aligned} \quad \dots (27)$$

In where k_7 is the yield-coefficient of-substrate disintegration and k_8 the yield-coefficient of hydrolysis of-carbohydrates, proteins and lipids however the simulation depicting the scenario using Matlab is as under that can be seen in figure 4.1, SIMULINK block shape shown in, is a form of model based on mathematical modeling in formula 17: -

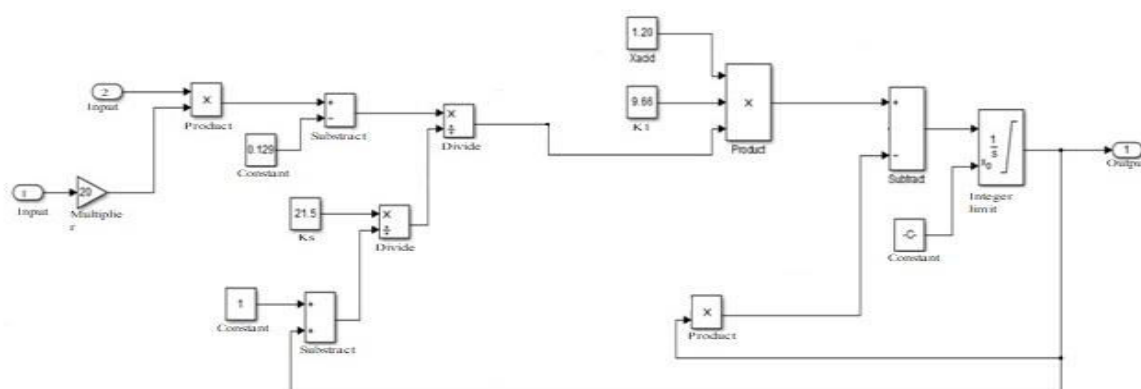
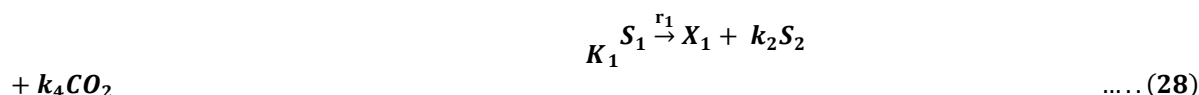


Figure 4.1: Process Hydrolysis Modeling

• Process Acidogenesis

In the acidogenesis stage, the bacteria generate acid and transform short-chain compounds into acetic acid, hydrogen and carbon dioxide formed in the hydrolysis stage. These bacteria are anaerobic bacteria that can grow and thrive in acidic conditions. To produce acetic acid these bacteria, require oxygen and carbon obtained from dissolved oxygen in solution, the formation of acid under anaerobic conditions is very important to form methane gas by microorganisms in the next process. In addition, these bacteria also turn alcohol, organic acids, amino acids, carbon dioxide, H_2S and a little methane gas into low molecular compounds. Acetic acid, propionic acid, butyric acid, H_2 and CO_2 are the most relevant compounds in the stage of acidogenesis. Furthermore, it contains small quantities of formic acid, lactic acid, valeric acid, methanol, ethanol, butadienol and acetone. Acid-forming bacteria can usually survive more abrupt conditions than methane-producing bacteria. These bacteria, if in anaerobic conditions, are able to produce staple food for producing methane gas and the resulting enzyme activity on proteins and amino acids will free amino salts which are the only source of nitrogen that can be accepted by methane-producing bacteria, therefore the simulation of below formulation is as under, SIMULINK block shape shown in Figure 4.2 is a form of model based on mathematical modeling in formula 19 and can be seen in figure 4.2.



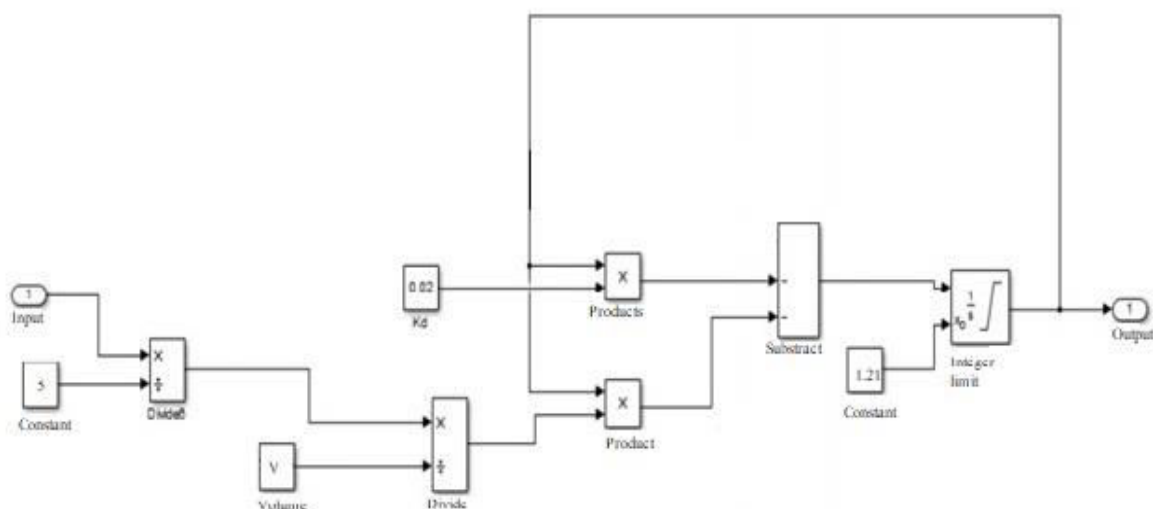
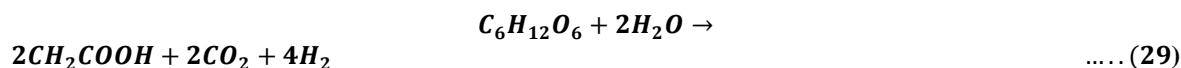


Figure 4.2: Process Acidogenesis Modeling

• Process Acetogenesis

Other results such as alcohol and Volatile Fatty Acid (VFA) from the process acidogenesis will be oxidized by acetogenic bacteria. On phase in this acetic acid and H₂ gas are produced which are used to form the gas methane. In the acidogenesis phase, about 20% acetic acid and 4% have been produced H₂ gas. In anaerobic environments, facultative microbes are able to break down acids long carbon chain fats such as propionic and butyric acids become acids acetate and H₂ gas. Therefore, at this stage, the acetogenic bacteria are producing hydrogen converting fatty acids and ethanol / alcohol to acetate, carbon dioxide and hydrogen. This advanced conversion is very important for success in biogas production, because methanogens cannot use fatty acid compounds and ethanol directly as :-



The Acetogenesis modeling can be seen in figure 4.3, SIMULINK block shape shown in Figure 4.3 is a form of model based on mathematical modeling in formula 20.

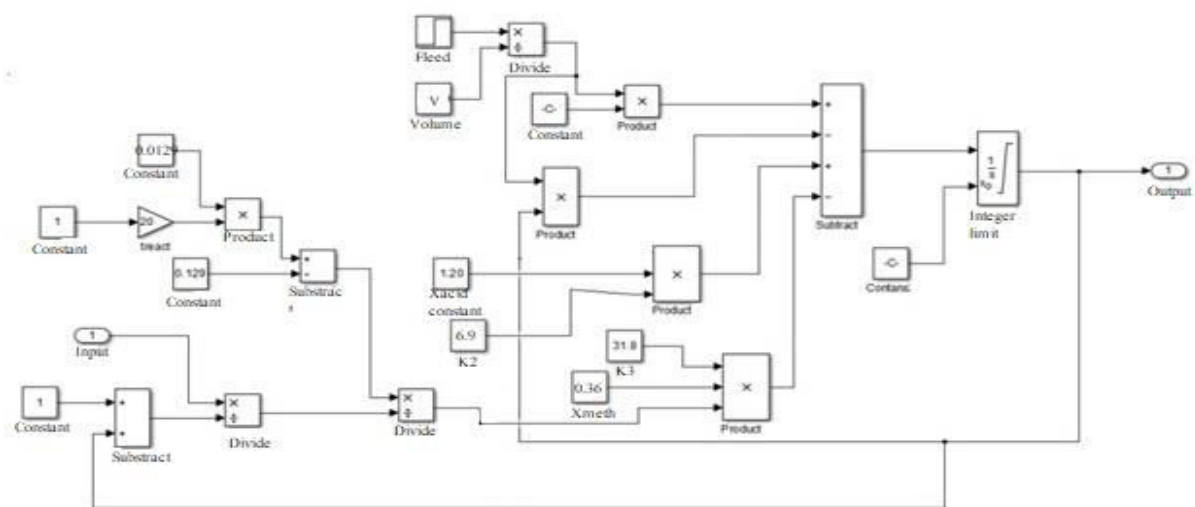


Figure 4.3: Process Acetogenesis Modeling

4.3.4 PROCESS METHANOGENESIS

Methanogenesis is the final stage of all conversion stages anaerobic organic matter into methane and carbon dioxide. On In the

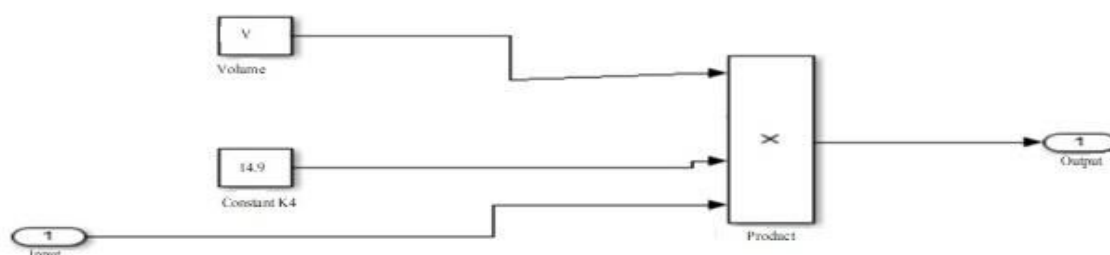
a. Acetolactic methane bacteria break down acetic acid into:


$$2H_2 + CO_2 \rightarrow CH_4 + 2H_2O$$

.....(31)

The block diagram illustrates a control system for a tank's water level. It starts with a 'Feed' input and a reference 'R'. These are combined in a 'Divide' block. The output of this block is then multiplied by a 'Volume' input in another 'Divide' block. The result is fed into a 'Subtract' block, which also receives an 'Input' signal. The output of the subtract block is multiplied by a 'Constant' (represented by a box with an infinity symbol) and then integrated by an 'Integrate limit' block. The output of the integrator is fed back to the 'Input' of the subtract block, completing the feedback loop. A '0.36' constant is also shown, which is connected to the 'Integrate limit' block.

Therefore, using the above simulation modeling techniques, the amount of methane gas output is depicted as under and can be seen in figure 4.5: -



- **Gas Rector**

650

in reactor pressure. Movement of parts the reactor also marks the start of internal gas production biogas reactor. When viewed from the flow of raw materials (waste), a biogas reactor can also divide into two, namely the batch type (tub) and continuous (flow). In tub type, Reactor raw materials are placed in the container (specified space) from the start until the completion of the contamination process. This is only commonly used at stages experiments to determine the gas potential of a type of organic waste; whereas, in the flow type, there is a flow of raw materials in and out of residues at any given time. The length (time) the raw material is inside a biogas reactor is called the hydraulic retention time. The contact between raw materials and acid bacteria / methane, are two important factors that play a role in the biogas reactor. Schematic of fixed dome biogas reactor is inculcated in the solution the simulation depicted as under and can be seen in figure 4.6:

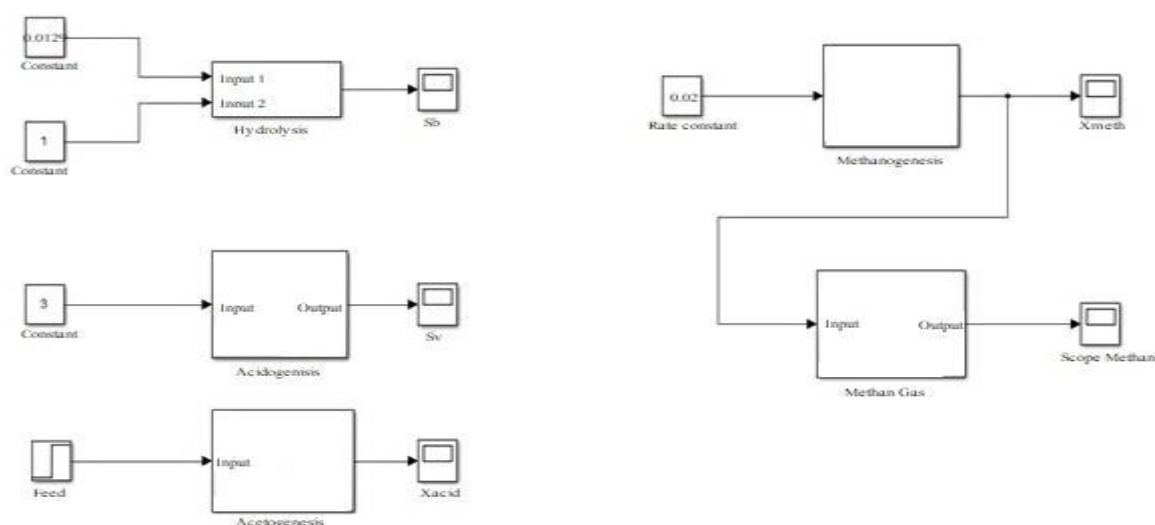


Figure 4.6: Reactor Model

• Simulation Results

The pressure in the reactor in the study, this is measured directly using Matlab so it can be seen that the process the reaction in the reactor for generate how much pressure that is generated as well as to know the increase that happens until the gas content is inside the material runs out. Based on the test preliminary, each repetition is carried out during the simulation to the point i.e., increase in methane gas production based on various temperatures. (Below simulated graphs shows X-Axis - time in days, while Y-axis consists of pressure in SI unit i.e., Pascal or N/m^2) and the result of simulations at temperature of 20°C, 30°C, 40°C, can be seen in figure 4.7, 4.8, 4.9 respectively.

• Simulation results with scenario 20° C

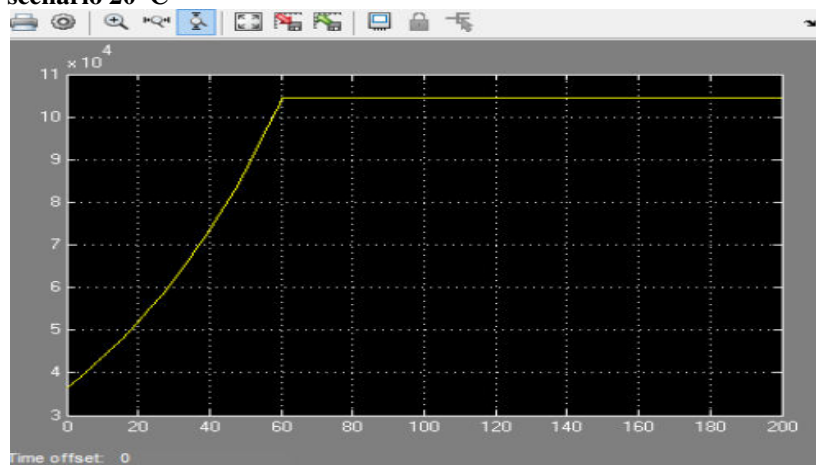


Figure: 4.7: Simulation Result of Converting Biogas Energy Scenario 20°C

- **Simulation results with scenario 30° C**

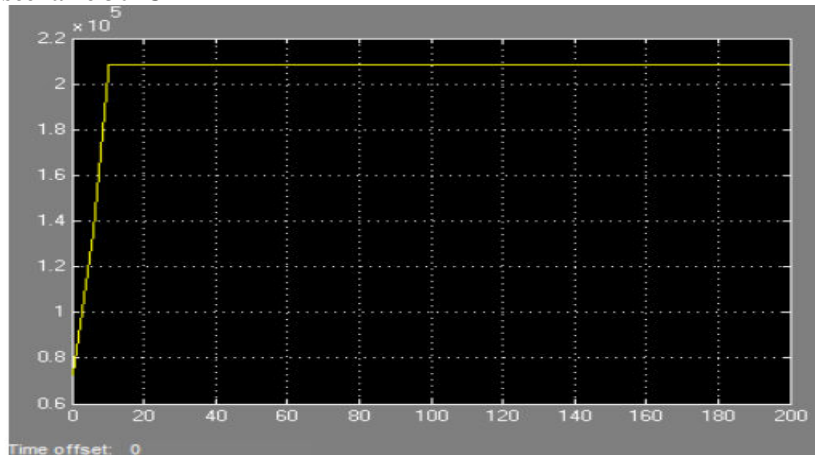


Figure: 4.8: Simulation Result of Converting Biogas Energy Scenario 30°C

- **Simulation results with scenario 40° C**

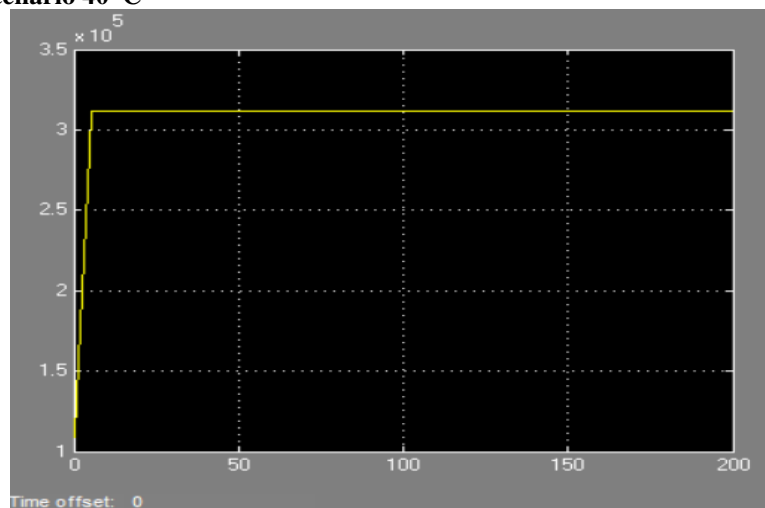


Figure: 4.9: Simulation Result of Converting Biogas Energy Scenario 40°C

- **Biogas Production**

The biogas production process begins by diluting the manure with water with 1:1 ratio. Excessive water in the system can block the biogas channel, lowers the heat level of the fire, and makes the fire red. Stirring is possible at any given time to prevent solids from settling on the tank bottom. The manure that has been mixed with water is forwarded into the digester until it closes the channel enter and output, then wait for approximately 10-40 days. Then filling the digester can be done twice a day, namely morning and evening. The first gas produced must be discarded because it is dominated by CO₂ gas. Next, biogas production can be carried out normally so that CH₄ gas production will increase and CO₂ gas will decrease by a percentage of 54%: 27%. Furthermore, biogas can be obtained connected to a stove or electric generator. The gas produced is very good for combustion because it is able to generate a high enough heat, the fire is blue, no smelly, not smoky. The following is a schematic of the biogas production process.

- **Break Event Point**

Break Even Point is a situation in which a company in a position does not experience gains or losses. This can happen if fixed costs and sales volume are used by the operation only to offset fixed costs and variable costs. The company suffers a loss if profits are only adequate to cover operating expenses and a portion of the fixed costs. In the meantime, the corporation will experience a profit if profits outweigh sales expenses and fixed costs. The components of the cost calculation at Break Even Point are as follows:

- **Fixed Costs**

Fixed costs are the costings whose value tends to be stable without being influenced by the units produced. So, this one component is constant or appears during production or when production is not carried out by the company. Examples of fixed costs at companies are labor costs, machine depreciation costs, water costs, and so on.

- **Variable Cost**

Variable costs are the costings whose value depends on the quantity of units or goods produced. So, this one component is its cost per unit does not remain or change according to the ongoing production action. So, for example, if production stops, variable costs decrease or don't exist and if production increases, variable costs will also increase. Some examples of variable costs are raw material costs, electricity costs, and so on.

- **Selling Price**

The selling price is the selling price of electricity produced by the company. This selling price needs to be known because it is included in the Break Event Point calculation formula. The following is a formula for finding the Break Even Point value based on sales.

$$BEP = \frac{FC}{VC/P}$$

Where;

FC is a Fixed Cost

P is the Price per Unit

VC is Variable Cost

- **Biogas Potential**

As mentioned earlier cow can produce 8 kg to 20 kg of manure, of which one cows can produce biogas 0.36 m³ / day, so if calculated, biogas produced from the cattle farm (having 150 cows) is 54 m³ / day. It is known that 1 m³ biogas can generate electric power of 11.17 kWh so that for 54 m³ biogas can generate energy for:

The amount of energy = volume of biogas * energy generated per m³

=2542m³*11.17 kWh

= 603.18 kWh

So, the theoretical amount of biogas energy in per cattle farm in National Capital Territory is 603.18 kWh with a power output of 603.18 KWh per day

- **Biogas Generator**

The specifications of the biogas generator that will be used depicted in table 4.3 below: -

Table 4.3 Biogas Generator Set Specifications

Features	Double Cylinder
	4-stroke
	OHV
	Air-cooled
	Three phase AC
AC voltage-	220/230V
AC Outputs-	Running Power: 30KW
	Peak Power
Frequency	50/60Hz
Starting system	Recoil or Electric start
Fuel	Biogas
Weight	150Kg
Other	Min. Fuel Consumptions: 2m ³ /hour

Assuming the biogas generator will be operated 18 hours a day (in two shifts: morning and evening), and then the energy output from this biogas-based power plant is:

Energy = Power * Time

= 30 KW * 18 hours

= 540 KWh

The capacity of the digester with a power of 603.18 kwh and if you know the generator engine with a capacity of 540 kwh with the capability 603.18 kwh digester, then the digester can hold gas for 1.117 days. As for the biogas needed to turn on the generator for 18 hours based on minimum consumption of biogas stated in the generator specifications (2m³ / hour) is = 18

hours * 2m³ / hour = 36 m³ / day. Meanwhile, the gas produced by 150 cows is 54 m³ in 18 hours. So, the process of forming biogas to drive a generator is at least 36 m³. Therefore, m³ for 18 hours of use, it may take as long $\frac{36m^3}{54m^3} = 0.667 \text{ days}$ the length of time the generator set operates for a biogas volume of 54 m³ can be determined by the following calculation:

$$\frac{\text{The volume of biogas production}}{\text{Biogas for generator}} = \frac{54}{2} = 27 \text{ hours}$$

• Homer Analysis

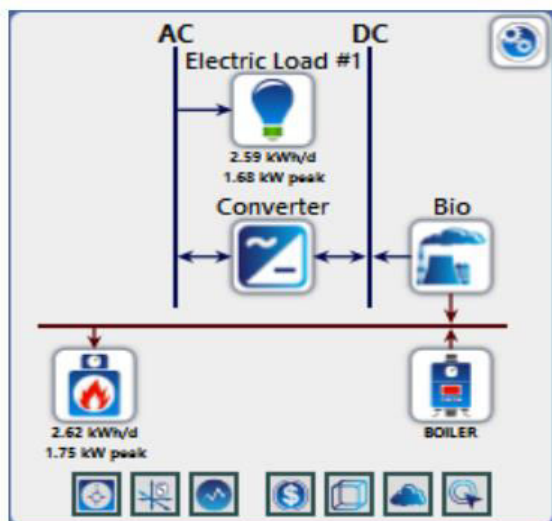


Figure 4.10 Schematic of the biogas power plant design

In Figure 4.10 there are several components including thermal loads, boilers, generators biogas, converters, and AC loads. The function of thermal load is as energy for heating water in the boiler. The function of the boiler in the circuit above is as a container where the process occurs heating water to produce hot steam. The function of the generator is as energy generator electricity. The converter function is to convert the DC current from the generator into the acceptable AC current used by AC loads. The following is a systematic process in Figure 4.13.

- 1- The thermal load heats the water in the boiler.
- 2- Water that has been heated in the boiler will produce hot steam.
- 3- The hot steam is directed to rotate the motor in the generator which then will generate electrical energy that is DC.
- 4- DC current from the generator flows to the converter to be converted into AC current.
- 5- AC current is applied to the load.

• Economic Analysis

The economic value of the biogas power plant is shown in Table- 4.4

Table 4.4 The Economical Value of the Biogas Power Generation System

Assessment Criteria	Score
Total energy production (KWh) Per Year	1,97,100
Net present cost Of Materials (Including Generator, Digester, Tank, Pipeline and Other)	Rs-2,50,000
Cost of energy (RS/KWH) Per Year	1.268

The total energy production using the biogas power generation system generates power amounting to Rs. 1.268 per kWh / year. The result of the total energy produced can be seen in Table 4.5 and will be calculated using formulation as Energy = Volume of Biogas * Energy generated per m³.

Table 4.5: Data on Total Energy Production per Year

Component	Output Production (KWh) In A Year
Generator (Operating 18 Hrs. A Day) R(540kwh*365days) in a year	1,97,100

$$E_{tot.prod} = 1,97,100 \text{ kWh (197.10MWh) in a complete year}$$

The results of net present cost calculation using Homer software can be seen in table 4.6 below:

Table 4.6: Results of Net Present Cost Calculation for HOMER Software

Component	Capital	Replacement	O&M	Fuel	Salvage	Total
Generic biogas generator	200000	0	10	14,608.01	-152.581	2,14,455.43
Homer Load Following	10	10	10	10	0	10
System Converter	399.99	10	39.97	10	01	439.97
Other	50,000	10	45,249.7	10	0	95,249.71
System	2,50,999.98	0	45,289.68	14,608.01	-152.581	3,10,745.09

$$NPC = 2,50,999.98 + 0 + 45,289.69 + 14,608.01 - 152.58$$

$$NPC = 3,10,745.09$$

Cost of Energy system Rs.3,10,745.09, total energy yield data serve loads and costs total per year used to calculate the Cost of Energy can be seen in Table 4.7 and calculated using equation $NPC = Capital\ Costs + Replacement\ costs + O\&M\ costs + Fuel\ Costs - Salvage$.

Table 4.7: Data on Total Energy Serving Expenses and Costs per Year

Assessment Criteria	Score
Total Energy Serving Load (kWh / year)	1,97,100 (Average Consumption)
Total Annual Fee (Rs)	3,10,745.09

$$COE = \frac{Rs.3,10,745.09}{1,97,100kWh} = 1.57Rs. / kWh \text{ in a year}$$

5. Conclusion and Suggestions

○ Conclusion

From the research conducted it can be concluded that:

- National Capital Territory with a population of 150 heads (example) of cattle has the potential to produce 0.36 m³ per cattle of biogas with the potential for electrical energy generated 603.18 kWh (**603.18 KW power**).
- Selection of a biogas generator set with a capacity of 30 kW suitable for use as an engine for converting biogas energy into electrical energy in a simple generator system with the HOMER generating system using a frequency of 50Hz.
- A simple generator system (Digester-biogas-Genset Biogas 30000 W-electricity) which is assumed to operate for 18 hours a day can generate energy of 540.00 kWh while the HOMER generator system for 24 hours can produce the 603.18 kWh electric energy).
- The remaining 63.18 m³ can be utilized for cooking and boiling as fuel usages in ordinary course of business, livelihood and to counter any losses.
- Investment in the construction of a simple generator system (Digester-biogas-Genset Biogas 30 kW of power resulting 540.00 kWh) will be more easily realized as compared to the construction of a HOMER generating system (Thermal-Boiler-Generator Bio 30 kW-Converter 603.10 kWh electricity load) with consideration of a fairly the economical price of Rs.1.57 of per unit in a year. Consequently, which is much cheaper and more economical to existing electricity.
- By the use of cow dung which is Municipal solid waste (basically biodegradable organic waste), we can reduce the municipal solid waste generation and air pollution to some extent in Delhi.

○ Suggestions

The suggestions in this study are:

- For the construction of a biogas-based power plant system in National Capital Region, it will be more efficient and economical with a simple generator system (Digester-biogas-Biogas Generator 30000 W-electricity) compared to the HOMER generating system (Thermal-Boiler-Generator Bio 30 kW -Converter 10 kW-electric load).
- It is necessary to hold further studies of course with different parameters regarding the biogas-based power plant system in National Capital Region so that the construction of the generator can be realized so that it can be distributed to the community.

- The bio-waste that is being generated during the production of biogas, can be used as manure for agriculture and farming which is much more effective and better than the harmful chemical compounds that is being used for agriculture and farming.

○ Declaration of Competing Interest

The authors declare no conflicts of interest concerning the research, authorship and publication of this article.

○ Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- [1] https://niti.gov.in/sites/default/files/2020-01/IEA-India%202020-In-depth-EnergyPolicy_0.pdf
- [2] <https://www.world-nuclear.org/information-library/energy-and-the-environment/renewable-energy-and-electricity.aspx>
- [3] http://delhiplanning.nic.in/sites/default/files/11%29%20Energy_0.pdf
- [4] <https://dahd.nic.in/about-us/divisions/cattle-and-dairy-development>
- [5] <https://static.pib.gov.in/WriteReadData/userfiles/key%20results.pdf>
- [6] Khan, M E., Martin, A R. (2016) Review of biogas digester technology in rural Bangladesh. *Renewable & sustainable energy reviews*, 62: 247-259 <http://dx.doi.org/10.1016/j.rser.2016.04.044>
- [7] <https://www.eesi.org/papers/view/fact-sheet-biogasconverting-waste-to-energy>
- [8] Artanti, D & Saputro, Roy & Budiyo, B. (2012). Biogas Production from Cow Manure. *International Journal of Renewable Energy Development (IJRED)*. 1. 61-64. 10.14710/ijred.1.2.61-64.
- [9] Tallou, Anas & Haouas, Ayoub & Jamali, Mohammed Yasser & Atif, Khadija & Amir, Soumia & Aziz, Faissal. (2020). Review on Cow Manure as Renewable Energy. 10.1007/978-3-030-37794-6_17.
- [10] Baredar, Prashant & Khare, Vikas & Nema, Savita. (2020). Biogas digester plant. 10.1016/B978-0-12-822718-3.00003-4.
- [11] Caposciutti, Gianluca & Baccioli, Andrea & Ferrari, Lorenzo & Desideri, Umberto. (2020). Biogas from Anaerobic Digestion: Power Generation or Biomethane Production? *Energies*. 13. 743. 10.3390/en13030743.
- [12] Jankowska, Ewelina & Zieliński, Marcin & Dębowski, Marcin & Oleskiewicz-Popiel, Piotr. (2019). Anaerobic digestion of microalgae for biomethane production. 10.1016/B978-0-12-815162-4.00015-X.
- [13] Li, Hailong & Mehmood, Daheem & Thorin, Eva & Yu, Zhixin. (2016). Biomethane production via anaerobic digestion and biomass gasification.
- [14] Jeong, Hyeon-Seok & Suh, Changwon & Lim, Jae-Lim & Lee, Sang-Hyung & Shin, Hang-Sik. (2005). Analysis and application of ADM1 for anaerobic methane production. *Bioprocess and biosystems engineering*. 27. 81-9. 10.1007/s00449-004-0370-4.
- [15] Assegaf, Ali & Umbara, Rian & Kurniawan, Isman. (2019). Pemodelan Produksi Biogas pada Reaktor Tipe Batch Menggunakan Metode Hamming Predictor-Corrector. *Indonesian Journal on Computing (Indo-JC)*. 4. 1. 10.21108/INDOJC.2019.4.1.138.
- [16] Saeed, Mohammed & Fawzy, Samaa & El-Saadawi, Magdi. (2018). Modeling and simulation of biogas-fueled power system. *International Journal of Green Energy*. 16. 1-27. 10.1080/15435075.2018.1549997.
- [17] Pathmasiri, Kalpani & Haugen, Finn Aakre & Gunawardena, Sanja. (2013). Simulation of a Biogas Reactor for Dairy Manure. *Annual Transactions of IESL. The Institution of Engineers, Sri Lanka*. 394-398.
- [18] Danielsson, Oskar. 2014. Modeling and Simulation of Anaerobic Manure Digestion into Biogas. Gothenburg, Sweden. Master's Thesis. Chalmers University of Technology.
- [19] Antonio, Juan. 2018. Modeling and Simulation of Biogas Production Based on Anaerobic Digestion of Energy Crops and Manure. Chemical Engineering Dissertation. Technischen Universität Berlin.
- [20] Raj, Dr & Jhariya, Manoj & Toppo, Pratap. (2014). COW DUNG FOR ECOFRIENDLY AND SUSTAINABLE PRODUCTIVE FARMING. *INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH*. 3. 201-202.
- [21] Artanti, D & Saputro, Roy & Budiyo, B. (2012). Biogas Production from Cow Manure. *International Journal of Renewable Energy Development (IJRED)*. 1. 61-64. 10.14710/ijred.1.2.61-64.
- [22] Matos, Camila F., Paes, Juliana L., Pinheiro, Érika F. M., & Campos, David V. B. De. (2017). BIOGAS PRODUCTION FROM DAIRY CATTLE MANURE, UNDER ORGANIC AND CONVENTIONAL PRODUCTION SYSTEMS. *Engenharia Agrícola*, 37(6), 1081-1090. <https://doi.org/10.1590/1809-4430-eng.agric.v37n6p1081-1090/2017>
- [23] Amanda D Cuéllar and Michael E Webber, Cow power: the energy and emissions benefits of converting manure to biogas, 2008 *Environ. Res. Lett.* 3 034002 <https://doi.org/10.1088/1748-9326/3/3/034002>
- [24] N O Abdullah and E S Pandebesie, The Influences of Stirring and Cow Manure Added on Biogas Production from Vegetable Waste Using Anaerobic Digester, 2018 *IOP Conf. Ser.: Earth Environ. Sci.* 135 012005 <https://doi.org/10.1088/1755-1315/135/1/012005>
- [25] Abdallah, Mohamed & Shanableh, Am & Adghim, Mohamad & Ghenai, Chaouki & Saad, S.A... (2018). Biogas Production from Different Types of Cow Manure. 10.1109/ICASET.2018.8376791.
- [26] Cuéllar, Amanda & Webber, Michael. (2008). Cow Power: The Energy and Emissions Benefits of Converting Manure to Biogas. *Environ. Res. Lett.* 3. 10.1088/1748-9326/3/3/034002.
- [27] Drapcho. et al. 2008. *Biofuel Engineering Process Technology*. USA: The McGrawHill Companies, Inc

JOURNAL of CRITICAL REVIEWS

ISSN- 2394-5125

VOL 8, ISSUE 01, 2021

- [28] Effendy, Sahrul. 2018. Biogas from Converting Cow Manure Waste as Material Generator Set To Generate Electrical Energy Capacity 0.3 kWA. National Seminar on Innovation and Technology Application in Industry. Polytechnic Sriwijaya. ISSN 2085-4218
- [29] Sloby, C. (2020). biogas. 10.1002/9783527809080.cataz01953.
- [30] Velivela, Asmitha & Barham, Husam & Bauer, John & Roschke, Jon & Daim, Tugrul & Meissner, Dirk. (2020). Biogas: Converting Waste to Energy. 10.1007/978-3-030-58301-9_18.
- [31] Adisasmito, S.. (2020). Biogas Generation from Jakarta Municipal Waste. 10.1201/9781003078883-22.
- [32] Khalil, Mohamed & Mansour, Hanaa. (2020). Biogas production. 10.13140/RG.2.2.30410.31680.
- [33] Lamb, Jacob. (2020). Biogas and the Energy Sector. 10.48216/9788269203325CH1.
- [34] http://www.fluid-biogas.com/?page_id=125&lang=en
- [35] <https://bigadan.com/p/biogas-technology/how-to-make-biogas>
- [36] <https://vikaspedia.in/energy/energy-production/bio-energy/biogas>
- [37] Velmurugan, Sivasubramanian & Deepanraj, Balakrishnan & Jayaraj, Simon. (2014). Biogas Generation through Anaerobic Digestion Process – An Overview. RESEARCH JOURNAL OF CHEMISTRY AND ENVIRONMENT. 18. 80-94.
- [38] <https://www.epa.gov/agstar/how-does-anaerobic-digestion-work>
- [39] Meisam Tabatabaei, Hossein Ghanavati, Biogas Fundamentals, Process, and Operation Series: Biofuel and Biorefinery Technologies <https://www.springer.com/gp/book/9783319773346>
- [40] Danielsson, Oskar. 2014. Modeling and Simulation of Anaerobic Manure Digestion into Biogas. Gothenburg, Sweden. Master's Thesis. Chalmers University of Technology.
- [41] https://energypedia.info/wiki/Electricity_Generation_from_Biogas
- [42] <https://www.britannica.com/technology/steam-engine>
- [43] <https://www.sciencedirect.com/science/article/pii/S1876610217337840/pdf?md5=0cb79069294f4190dcccdd47f7f449403&pid=1-s2.0-S1876610217337840-main.pdf>
- [44] <http://www.bioenergyfarm.eu/de/klein-biogasanlagen-gulle-kleinanlagen/biogasnutzung/>
- [45] Taizhou Bison Machinery., Ltd
- [46] <https://eepafrica.org/about-us/success-stories/bio2watt/>
- [47] <https://sistema.bio/success-stories/>
- [48] <https://www.bmwgroup.com/en/responsibility/sustainable-value-report/popup/bio.html>
- [49] <https://www.gree-energy.com/ground-breaking-of-hamparan-project/>
- [50] <http://task37.ieabioenergy.com/files/daten-redaktion/download/Technical%20Brochures/Smart Grids Final web.pdf>
- [51] <https://blog.anaerobic-digestion.com/biogas-generator/>
- [52] https://www.researchgate.net/post/How_is_the_power_capacity_kW_or_MW_of_a_biogas_installation_determined
- [53] <https://www.biogasworld.com/>
- [54] <https://www.biogas-info.co.uk/about/biogas/>
- [55] <https://www.energy.gov/eere/bioenergy/biopower-basics>
- [56] Ogur, Eric. (2013). Design of a Biogas Generator. Concurrent Engineering Research and Applications. 3. 2248-9622.
- [57] Price, Elizabeth & Cheremisinoff, Paul. (1985). METHODS OF BIOGAS GENERATION AND COMPARISONS.
- [58] Motjoadi, Vinny & Adetunji, Kayode & Joseph, Prof. (2020). Planning of a sustainable microgrid system using HOMER software. 1-5. 10.1109/ICTAS47918.2020.233986.
- [59] Djalilova, Nigora. (2021). Feasibility study of hybrid wind-solar stand-alone energy systems using HOMER software. 10.4324/9781003110071-6.
- [60] Raveena, Battula & Rao, Bathina & Nadakuditi, Dr Gouthamkumar. (2018). Optimization of Hybrid off Grid Power System Using HOMER software. 696-700. 10.1109/RTEICT42901.2018.9012219.
- [61] Kansara, B.U. & Parekh, B.R. (2011). Modelling and simulation of distributed generation system using HOMER software. 328-332. 10.1109/ICONRAEECE.2011.6129804.
- [62] Aly, Abdelmaged & Kassem, Ahmed & Sayed, Khairy & Aboelhasan, Ismail. (2019). Design of Microgrid with Flywheel Energy Storage System Using HOMER Software for Case Study. 485-491. 10.1109/ITCE.2019.8646441.
- [63] Shahinzadeh, Hossein & Moazzami, Majid & Fathi, Shabnam & B. Gharehpetian, Gevork. (2016). Optimal sizing and energy management of a grid-connected microgrid using HOMER software. 1-6. 10.1109/SGC.2016.7882945.
- [64] Noor, Md. Fahel & Mallick, Bijoy & Habib, Ahsan & Noor, Nahiyah. (2019). COMPARISON OF PERFORMANCE AND COST OF WIND AND SOLAR HYBRID SYSTEM FOR SAINT-MARTIN ISLAND USING HOMER SOFTWARE.
- [65] S.K. Singh, M.kumar and S.Singh.(2020).PLASMA TECHNOLOGYAS WASTE TO ENERGY: A REVIEW.

List of figures

Figure 2.1: Biogas Formation Stages/Phase
Figure 2.2: Graph Representative Anaerobic Digestion Temperature
Figure 2.3: Biogas Generator
Figure 3.1.1: Research Flowchart
Figure 3.1.2: Research Flowchart continued
Figure 3.2: SIMULINK Reactor Model
Figure 4.1: Process Hydrolysis Modeling
Figure 4.2: Process Acidogenesis Modeling
Figure 4.3: Process Acetogenesis Modeling
Figure 4.4: Process Methanogenesis Modeling
Figure 4.5: Modeling of the amount of methane gas output
Figure 4.6: Reactor Model
Figure: 4.7: Simulation Result of Converting Biogas Energy Scenario 20°C
Figure: 4.8: Simulation Result of Converting Biogas Energy Scenario 30°C
Figure: 4.9: Simulation Result of Converting Biogas Energy Scenario 40°C
Figure 4.10 Schematic of the biogas power plant design

List of Tables

Table 2.1: Composition of Cow Manure
Table 2.2: Cow Manure specifications with a total weight of 635 kg
Table 2.3: Compositional Biogas
Table 2.4 Volatile Solid Components
Table 2.5 C/N Ratio of Organic Materials
Table 2.6 Energy Conversion from Methane Gas to Electrical Energy
Table 4.1 Sample performance of the biogas plant
Table 4.2: Result of biogas capacity calculations
Table 4.3 Biogas Generator Set Specifications
Table 4.4 The Economical Value of the Biogas Power Generation System
Table 4.5: Data on Total Energy Production per Year
Table 4.6: Results of Net Present Cost Calculation for HOMER Software
Table 4.7: Data on Total Energy Serving Expenses and Costs per Year

Multi Domain Fake News Analysis using Transfer Learning

Pratyush Goel

Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
pratyushgoel99@gmail.com

Samarth Singhal

Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
samarthsinghal1402@gmail.com

Snehil Aggarwal

Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
snehil160399@gmail.com

Minni Jain

Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
minnijain@dtu.ac.in

Abstract—Fake news detection is a significant problem where information is available from multiple sources across the internet. Most of the research on fake news has only targeted politics-related articles, but such models would not be robust enough to tackle fake news in the real world. To solve this problem, this research work incorporated transfer learning using attention-based transformers (BERT, RoBERTa, XLNet, DeBERTa, GPT2) and trained them on multi-domain datasets FakeNews AMT and Celebrity across different domains i.e. Politics, Entertainment, Sports, Business, Education and Technology. The proposed model has obtained state-of-the-art results while doing multi-domain and cross-domain testing, having beaten previous papers conformably. Also, the model has achieved a 99.3% accuracy on FakeNewsAMT and 84% accuracy on celebrity dataset. We believe the synergy of transfer learning in a multi-domain setting will make a robust model, which would be relevant in the real world. This idea originated from the fact that multi-domain research's critical challenge is that data distribution is varying, and the key benefit of transfer learning is that it can perform well even when it is trained and tested on different data distributions.

Keywords—Fake News; Multi-domain; Transfer learning; Transformers; Cross-domain; Natural language processing; Attention

I. INTRODUCTION

In the era of digitalization, there has been a drastic increase in internet consumption. Nowadays, people rely more on the internet and social media websites for gathering news and information. 90% of the world's data has been generated in the last two years itself. Every day humans create 2,500,000 Terabytes of data, and not all of it is correct.

To check the Truthfulness of content is of utmost priority for the media aggregator to prevent spreading fake news on their platform and maintain their credibility. Fake news impacts people's decision-making and may induce them to make harmful decisions. Various social networking sites have

deployed different ways to tackle fake news, such as moderators using fact-checking websites (ex. Snopes.com, PolitiFact). These websites play an essential role. However, they contain regional and domain-specific news and require constant updating as a news article can be factually correct in the present but fake in the future. All this requires a lot of time and effort.

Fake news detection until recently has only been concentrated in one domain which is politics. The datasets for fake news analysis have either been from the websites like Buzzfeed or Twitter that are used to track the news content, or from the genre of satire news like "The Onion" dataset or from websites like PolitiFact, which are the fact-checking websites.

However, these datasets have their own set of issues, challenges, and drawbacks. For example, in satire news, the news is made to mimic the real news but with humor or sometimes irony, and some news is factually correct but has turned and twisted them to their own convenience to make fake news. But the real challenge remains to solve the fake news in a multi-domain setting.

Models trained on a single domain dataset, although highly accurate, fail to perform on news from a different domain. In real world application, news can originate from any domain be it politics, sports, environment, technology, business, economy, etc. Classical machine learning on multi-domain is hard as the data distribution varies in train and test set. These factors have motivated to solve this problem, mainly because of its challenging nature and the real-world relevance it holds. If a good, well-built, robust, multi-domain fake news detector is presented to the world, it will be possible to filter any news without worrying about its domain.

Due to the work of Perez-Rosas et al [1], we are able to get a dataset that can help solve the multi-domain fake news classification challenge. This work offered the world two new datasets, namely FakeNewsAMT and Celebrity, that focus not only on politics but also on other domains, other genres. With

the help of these datasets, one can build a robust model that can be used to classify fake news from a range of domains.

Transfer learning is the concept of applying knowledge on a dataset gained from some other large dataset. There are various pretrained sequence to sequence classifiers that can be optimized for further use. A major advantage of transfer learning models is that data distribution changes do not impact them much. This is because these transformers have 100s of millions of parameters and are trained on billions of sequences. This advantage will help them perform much better in both multi-domain and cross-domain analysis. This shows the synergy between multi-domain analysis and transfer learning.

In this paper, we fine-tune various transformers (BERT [2], XLNet [3], GPT2 [4], RoBERTa [5], DeBERTa [6]) on both FakeNewsAMT and Celebrity dataset individually to perform multi-domain analysis, as FakeNewsAMT consists data spread across different domains, and evaluate the results. Later we perform cross-domain analysis i.e. training on Celebrity, testing on FakeNewsAMT and vice-versa. In the end we perform multi-domain training and domain-wise testing i.e. tested on each domain one by one and trained on remaining domains for FakeNewsAMT dataset. With all this analysis we show the possibility of solving the problem of fake news using transfer learning once and for all.

II. RELATED WORK

Fake news detection has become a crucial research area with an increase in internet usage and data consumption in this ever-evolving world. Every research aims to improve upon the existing methods of classifying the news.

The work of Perez-Rosas et al. [1] introduced two novel datasets FakeNewsAMT and Celebrity. The former being crowd-sourced, and the latter collected from the internet. FakeNewsAMT consists of news spread across six domains that will help us to do multi-domain analysis on our classifiers.

A group of studies has explored the idea of training classical machine learning-based models. Paper [1] also proposed a linear SVM classifier and conducted several exploratory analyses to identify linguistic properties, provided a different set of words as input features to the model, and achieved an accuracy of 78%. Paper [7] proposed a Random forest model and used paraphrasing tools and Grammarly to extract linguistic features from the news articles to train the model. A major drawback with these classical machine learning-based studies in the context of fake news detection is that these models fail to provide higher accuracy in real-world applications because of their inability to work in multi-domain and cross-domain settings.

Another group of studies trains various deep learning models to improvise over classical machine learning models. Previous work [8] proposed a hybrid model using LSTM, speaker profiles, and attention models on the Liar dataset [9]. It draws a comparison between CNN, LSTM and also shows how adding a speaker profile could improve the results. Previous work [10] proposes a bidirectional GRU approach on FakeNewsAMT and celebrity datasets. With a score of 0.68, it shows the scope of improvement in the cross-domain analysis.

Reference [11] leverages the information provided by three characteristics: text of an article, the response it receives, and the source users and proposed a hybrid model on the Twitter dataset. Paper [12] explores a unique propagation-based geometric deep learning approach that uses a four-layer graph CNN with two convolution layers on a custom Twitter dataset that contains entries from 2013 to 2018. This approach could achieve an approximate ROC AUC of 92.70. Reference [13] introduced a dataset FNAD which was collected by scraping data from satirical news sites (The Onion, Faking news, The hard times) and proposed a neural network architecture using both bi-directional LSTM layer and bi-directional GRU layer and concatenates the output of both the layers to achieve an accuracy of 80.2%. The author also analyzed their model on FakeNewsAMT, Celebrity, and PolitiFact with accuracy 81%, 82.61%, and 75.44%, respectively.

One excellent research work in the line of fake news is done by [14]. The author proposed a semantic matching model to do claim verification. This work was based on the FEVER dataset [15] that contained a list of claims made by humans taking Wikipedia as the reference. These approaches require a large dataset to train a robust model but provide great results on their respective dataset only. However, in the real world, we need a classifier that can classify news of any genre, which these models fail to do.

Few groups of researchers have explored the idea of using attention-based transformers for single domain settings. Reference [16] draws a comparison of BERT [2], Roberta [5], XLNet [2] on the FNC-1 dataset in the single domain settings. Paper [17] used Bert classifier and performed single domain analysis on the Liar dataset [9]. Reference [18] introduces the new dataset "Fake news Filipino" and explores the idea of transfer learning in the Filipino language using transformers such as BERT, GPT-2. Previous work [19] proposes a multimodal framework for fake news detection. It classifies using both textual and visual features of an article. It uses XLNet with a dense layer and achieves an accuracy of 0.856 on the dataset GossipCop.

None of the existing research shows how transfer learning-based models improve the model's performance on multi-domain and cross-domain settings. This creates a unique opportunity for us. We feel that the nature of transfer learning and multi-domain analysis complement each other. This thought originated from the fact that multi-domain research's critical challenge is that data distribution is changing, and the key benefit of transfer learning is that it can perform well even when it is trained and tested on different data distributions. Our research tries to exploit this nature of the two.

III. METHODS USED

A. Transformers

Transformers are described as sequence-to-sequence architecture. This type of architecture converts a sentence or a sequence of words into some other sequence. These models are good at translation and convert text written in one language to a sequence in another language made of different words.

Transformers are often described as Long-Short-Term-Memory (LSTM)-based models. LSTM models help filter out more important words from unimportant words or, in other words remembering only essential words.

Transformers is made of two parts which are Encoder and Decoder. The encoder converts a sentence or sequence into n-dimensional vectors. The Decoder, on the other hand, takes this abstract vector and converts it into a sequence. Thus, Encoder and Decoder can be worked in tandem to transform one sequence to another sequence, which may be in different language, symbols, or a same as input. Training the model is very important for encoder and decoder to work efficiently.

Another part of Transformers is the concept of Attention. This attention-mechanism helps the transformer in deciding which elements of the input sequence are essential at each step. Encoder reads input sequences and marks important words attributing different weights to those inputs. The Decoder then takes encoded sentences and weights provided by the attention.

The self-attention is an integral part of the transformer structure. For each word using self-attention the model is able to predict better encoding for a word. It does so by taking in account the position of other words in the sequence. Thus by looking at other words model can better understand the context of the word used further in the sequence. This concept is similar to the ones used in RNNs, like RNNs maintain a hidden state where it stores all the words that it has already come across while processing the current batch of words.

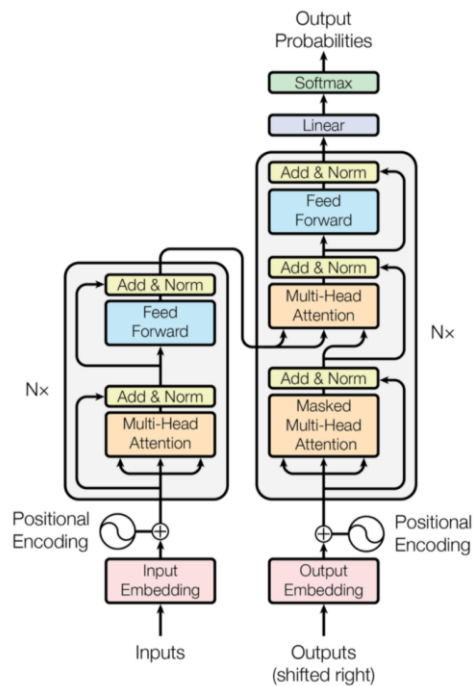


Figure 1: The Transformer – model architecture [20]

Using the figure 1 we can see that both encoder and decoder made of Feed Forward Layers and Multi-Head Attention. The modules of the transformer can be piled

multiple times on top of each other. Figure 1 uses Nx to denote this

Positional encoding is a crucial part of the module because transformers lack any recurrent networks that help in remembering how sequence is fed into the model. The embedded representation of each word consist of this relative positioning.

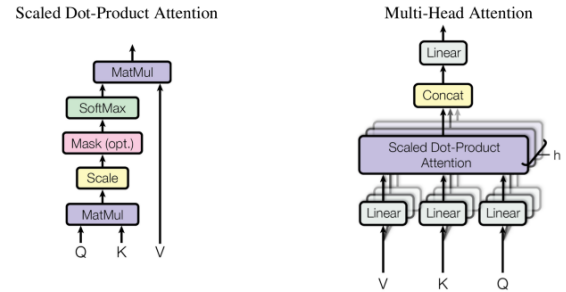


Figure 2 (Left) Scalar dot-product attention. (Right) Multihead attention consist of several attention layers running in parallel. [20]

$$Attention(Q, K, V) = softmax \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (1)$$

Q = Matrix that contains the query

K = all Key values which are vector representation of all words in sequence

V = all Values which are also vector representation of all words in sequence

This equation signifies the overall result of left side of Figure 2. These values of K, Q and V are calculated using the embedding that are trained during training process. The scalar dot product is used as it both time and space efficient.

$$a = softmax \left(\frac{QK^T}{\sqrt{d_k}} \right) \quad (2)$$

This a signifies the attention weights. Thus, each word in V is multiplied and summed with a value of a . a is given by a value of each Q word of the sequence, defined by how much is this word governed by supplementary words in the sequence given by K. SoftMax function ensures that this value of weight remains between 0 and 1. a value for words in sequence is calculated separately.

Also, as in Figure 2, a attention mechanism consist of scaled dot product of a given values of Q, K and V. These values pass through matrix multiplication and SoftMax filters as shown on left side of Figure 2. This forms a part of multi-head Attention component. As seen on the right side of Figure 2 using different linear projectors, many linear projections of Q, K and V are generated which are fed to the scaled dot-product layer this helps in system train better from varying

representation of values. Then all values from product are concatenated to form final attention component.

B. Transfer Learning

The main problem encountered while training any machine learning model is the unavailability of data or poor model generalization. Due to this, whenever we train a new model, accumulating the data is a daunting task. Also, if we train any new model on a dataset and apply it on another dataset, the performance decreases. This happens because the model fails to extrapolate the data and deduce the pattern. Thus, to sum this up, the main issue is to tackle ever-changing data like news and avoid retraining models frequently.

Thus, transfer learning helps us in solving this problem by enabling us to take an existing pre-trained model and use it on a novel dataset. Transfer Learning has the edge over traditional learning because it needs less generalized data and can adapt faster to new data.

Domain Transfer is one of the most prominent form of concept used for transfer learning. In this type the model is trained on one of the source dataset and then used to predict labels on target dataset.

C. Models fine-tuned

BERT [2] uses a transformer as an attention mechanism which helps in selecting important words and learns contextual relationships between them. BERT encoder reads the entire sequence of words together instead of directional mode, which reads the sequence of words sequentially. Thus, model can better learn the context of each word taking advantage of its surroundings.

XLNet [2] consolidates bidirectional context while at the same time dodging the [MASK] tokens and independent predictions. It uses permutation language modeling which predicts all tokens in some random order. Transformer XL is used primary architecture for XLNet.

RoBERTa [5] is a better-trained version of BERT [2]. It is trained with bigger batches. RoBERTa does not predict next sequence. It trains on longer sequences and mask pattern is changed dynamically.

DeBERTa [6] improves on BERT [2] and RoBERTa [5] models using two techniques. The first one is a disentangled attention mechanism where every word is represented using two vectors that display its content and position. Second, the use of an enhanced mask decoder helps to incorporate absolute position in the decoding layer and predicts masked tokens in model pre-training.

GPT2 [4] consists of solely stacked decoder blocks from transformer architecture. In GPT-2, the context vector is zero-initialized for the first word embedding. Also, GPT2 uses masked self-attention and is a close copy of basic transformer architecture.

IV. EXPERIMENTS

In this research, we performed three kinds of experiments that helped us verify and validate the idea of using transfer

learning via transformer-based machine learning algorithms to detect fake news in both multi-domain and cross-domain setting. In this section, we would be describing the data we used and how we preprocessed our data which is followed by the Evaluation Metrics and the Model Parameters. Further we explain all the experiments and their results in detail. The target is to have better accuracy with greater precision and recall. Qualitative formulation with its objective is elaborated in the evaluation metrics section.

A. Dataset

Most of the datasets in the field of fake news focus on the politics genre. However, fake news is way larger than just to be studied with politics as it is present in every life domain. This widespread impact leads to the need for multi-domain fake news analysis. We are using the datasets released by [1] for these specific research needs. They, in this work, released two novel datasets named FakeNews AMT and Celebrity. These datasets are a collection of news articles that can be fake or legit. These news articles have a title, news body, and a label (fake/legitimate) associated with them. Table 1 that follows shows us the statistics related to the dataset.

TABLE 1.
DATASET STATISTICS

Dataset	Number of examples	Average words	Average sentences	Label
FakeNewsAMT	240	132	5	Fake
	240	139	5	Legit
Celebrity	240	399	17	Fake
	240	700	33	Legit

The First dataset, FakeNews AMT, contains a corpus of 480 news articles. The distribution of fake and legit news articles in this dataset is even, with half of the articles being legit and the other half fake. The authors collected the legit articles from various mainstream American news sites. They then used crowdsourcing via Amazon Mechanical Turks to manually get these articles annotated, genre verified, and the corresponding fake news articles generated based on these legit articles. This dataset's 480 news articles belong to six news genres: technology, education, business, sports, politics, and entertainment. Having news from such diverse genres makes the dataset a multi domain one in the real sense. Here also, the distribution is hundred percent even with eighty articles of each of the six genres. Out of these eighty articles, forty are fake, and forty are legit. The second dataset, Celebrity, focuses on celebrity news. This dataset was crawled from the web to get news related to celebrity gossip. The distribution in this dataset, just like the first one, is even, with two hundred and forty articles being legit and the other two hundred and forty being fake.

The datasets we are using are in the form of a tab-delimited text file. Each data point is an annotated news article along with its title and genre. To get the dataset ready for use, first, the text files were converted to CSV files and then they were

encoded in UTF-8 so that they could be processed by the tokenizer.

B. Data-preprocessing

The First step was to decide what would be the input for the transformer models. We decided that genre should not be given in the input so that the model is more robust in the real world where information about the news genre might not always be available. Next, the title and article body were concatenated into one string which would become our input. The reason behind this concatenation is the fact that transformers take inputs in the form of one sentence.

Finally, the input was then tokenized so that it could be fed into the transformers. In tokenization, the words are converted into tokens, and additional special tokens called [SEP] and [CLS] are also appended. We used the tokenizers available in the Hugging Face [21] library for this purpose. The following table 2 shows the tokenization style for different NLP transformers. For each transformer we have used, we have given the specific format in which tokenization is to be done, so that the model can take the input properly.

TABLE 2.

TOKENISATION STYLES FOR DIFFERENT TRANSFORMERS

Transformers	Tokenization
BERT	[CLS] + tokens + [SEP] + padding
DeBERTa	[CLS] + tokens + [SEP] + padding
RoBERTa	[CLS] + prefix space + tokens + [SEP] + padding
GPT-2	Padding + tokens + [SEP] + [CLS]
XLNet	Padding + tokens + [SEP] + [CLS]

C. Evaluation metrics

We have used various metrics to analyze our results and determine which transformer model is the most capable in predicting make news in the multi domain settings. The most important metric in our study has been accuracy. The random baseline would be fifty percent, as half of the articles are legit. We also compared the results we found with the previous research results on this topic. The other three metrics we used were precision, recall, and F1-score. To calculate these metrics, we were also required to find the true positive, true negative, false positive, and false negative values for each of our experiments. The goal here was to have as high an F1-score as possible, which required us to maintain the right balance of accuracy and precision.

One of the primary reasons we chose to focus on F1-score instead of either precision or recall is that when we are trying to weed out fake news, we need to ensure that every real news is marked genuine but also need to take care that no fake news slips our hands at the same time. Ultimately, we need to aim to neither have false positives nor false negatives. Both are equally harmful.

D. Model parameters

The smallest possible models for the transformers were used to minimize the computation time and resources. We want

an ideal fake news detector to work very fast in an efficient way. All the transformer models we used were trained on lower case English alphabets, had 12 layers of transformer blocks with 12 attention heads and parameters in the range of 100 million. After a series of experiments and fine-tuning, the hyperparameters were also fixed for all the transformers. While providing a list of all arguments will become very tedious, we have given the main arguments in the following table 3. For the hyperparameters not mentioned in the table 3, we have used the standard given by the transformer model's author. One can find the list of default parameters in the huggingface [21] documentation.

The batch size is one of the most important hyperparameters that is needed to be tuned in a deep learning model. There needs to be a perfect balance, the batch size should not be so large that we lose generalization and not too small to lose speed. We experimentally found 8 to be perfect for our experiments. We have decreased the Adam optimizer's learning rate from 5.0×10^{-5} to 2.0×10^{-5} in order to increase the number of updates taken to reach the optimal weight. As good generalization is very important in multi domain settings, we have used a weight decay of 0.01. Warm up steps are 0 due to the small size of the dataset it is not possible to use some steps for warm up. Rather the learning rate has been kept low.

TABLE 3.

HYPERPARAMETER VALUES

Hyperparameter	Value
Batch size	8
Learning rate	2.0×10^{-5}
Weight decay	0.01
Adam epsilon	1.0×10^{-8}
Warmup steps	0
Epoch count	5

E. Experiments and results

We performed three types of experiments in the research in order to test any new algorithm thoroughly. We have done two kinds of comparison: First was an internal comparison between all the transformers we tried using transfer learning. The second comparison was with the prior research done in the field of multi-domain fake news analysis by [10], [7] and also with the baseline set by the dataset authors. The three experiments are as follows:

1) *Multi-domain analysis*: For multi-domain analysis, experiments were performed on both datasets individually. Training and test sets were divided in a seventy - thirty ratio. This experiment is an actual test of verifying that the algorithm can work on data that can have a different vocabulary and data distribution. While testing the transfer learning approach on top transformer models, we observed that RoBERTa [5] not only gave the best results on the FakeNews AMT dataset, but also gave the best results on Celebrity Dataset. These transformers were also able to beat any other previous model ever applied on these datasets as seen in Table

4. We can see the detailed results of our experiments in the following tables 4 and 5. Our results clearly indicate that transfer learning on transformers emerges as the best method to do multi-domain fake news analysis.

Table 4.
RESULTS ON FAKE NEWS AMT

Model	Accuracy (%)	F1-Score	Precision	Recall
BERT	85.40	0.86	0.82	0.88
DeBERTa	86.10	0.86	0.83	0.88
GPT-2	97.70	0.97	0.98	0.97
RoBERTa	99.30	0.99	1	0.98
XLNet	91.60	0.91	0.90	0.92

Table 5.
RESULTS ON CELEBRITY

Model	Accuracy (%)	F1-Score	Precision	Recall
BERT	74.00	0.65	0.82	0.54
DeBERTa	80.60	0.78	0.76	0.84
GPT-2	74.00	0.68	0.77	0.60
RoBERTa	84.00	0.82	0.89	0.75
XLNet	74.60	0.72	0.71	0.74

Table 6.
ACCURACY COMPARISON WITH PREVIOUS RESEARCH (%)

Model	FakeNewsAMT Dataset	Celebrity Dataset
RoBERTa	99.30	84.00
SVM [1]	74.00	76.00
SGG [7]	95.00	78.00
Model 1 [10]	77.08	76.53
Model 2 [10]	83.30	79.00

2) *Cross-domain analysis*: The second set of the experiment performed was the cross-domain analysis. Here, the transformers were trained on one dataset, and the prediction was performed on the other. This experiment verifies the transfer learning aspect of this research. As the datasets for the training and testing are different, both vocabulary and data distribution are also very different. In the true sense, it is a classic example of an algorithm that learns from one place and applies it to another. RoBERTa [5] gave us state-of-the-art results in this experiment as seen in table 7. It has also improved upon the results from past papers mentioned in table 8. In the two tables, 7 and 8, we have specified the training and testing dataset as it can be seen in the first two columns of the tables. Our results clearly indicate that transfer learning on transformers is very robust and very apt for multi-domain fake news analysis.

TABLE 7.
RESULTS OF CROSS-DOMAIN ANALYSIS

TRAINING	TESTING	MODEL	ACCURACY (%)
FakeNewsAMT	Celebrity	BERT	56.20
		DeBERTa	55.40
		GPT-2	54.60
		RoBERTa	59.40
		XLNet	55.20
Celebrity	FakeNewsAMT	BERT	56.00
		DeBERTa	55.00
		GPT-2	53.50
		RoBERTa	70.20

	XLNet	61.67
--	-------	-------

TABLE 8.
COMPARISON WITH PREVIOUS RESEARCH

TRAINING	TESTING	MODEL	ACCURACY (%)
FakeNewsAMT	Celebrity	RoBERTa	59.40
		SVM [1]	52
		SGG [7]	56
		Model 2 [10]	54.3
Celebrity	FakeNewsAMT	RoBERTa	70.24
		SVM [1]	65.00
		SGG [7]	70.00
		Model 2 [10]	68.55

3) *Multi-domain training and domain-wise testing*: We performed the third and final set of experiments on the FakeNews AMT dataset. The dataset had news articles from six genres namely – Technology, Education, Business, Sports, Politics and Entertainment. The design of the experiment was such that five out of the six genres were made part of the training set, and the sixth one was the test set and like this every genre was made a test set once. This experiment was to test how the algorithms performed on heterogeneous data. From table 9 we can see that this experiment's results were also outstanding. Table 10 confirms that we performed better than the previous attempts at this. DeBERTa [6], GPT2 [4] and RoBERTa [5] performed really well in these tasks. These experiments again reinitiated the essence of multi-domain analysis and transfer learning.

TABLE 9.
RESULTS OF EXPERIMENT 3

TEST DOMAIN	MODEL	ACCURACY (%)
Technology	BERT	88.75
	DeBERTa	93.75
	GPT-2	95.00
	RoBERTa	98.70
	XLNet	88.75
Education	BERT	95.00
	DeBERTa	97.50
	GPT-2	100.00
	RoBERTa	100.00
	XLNet	98.70
Business	BERT	72.50
	DeBERTa	98.70
	GPT-2	97.50
	RoBERTa	98.70
	XLNet	91.20
Sports	BERT	62.50
	DeBERTa	100.00
	GPT-2	98.75
	RoBERTa	100.00
	XLNet	92.50
Politics	BERT	96.25
	DeBERTa	97.50
	GPT-2	96.25
	RoBERTa	100.00
	XLNet	98.75
Entertainment	BERT	62.50
	DeBERTa	100.00
	GPT-2	100.00
	RoBERTa	100.00
	XLNet	88.70

TABLE 10.
COMPARISON WITH PREVIOUS RESEARCH

TEST DOMAIN	MODEL	ACCURACY (%)
Technology	RoBERTa	98.70
	SVM [1]	90.00
	SGG [7]	98.70
	Model 2 [10]	88.75
	Model 1 [1]	76.22
Education	RoBERTa	100.00
	SVM [1]	84.00
	SGG [7]	96.20
	Model 2 [10]	91.25
	Model 1 [1]	77.25
Business	RoBERTa	98.70
	SVM [1]	53.00
	SGG [7]	93.70
	Model 2 [10]	78.75
	Model 1 [1]	74.75
Sports	RoBERTa	100.00
	SVM [1]	51.00
	SGG [7]	96.20
	Model 2 [10]	73.75
	Model 1 [1]	70.75
Politics	RoBERTa	100.00
	SVM [1]	91.00
	SGG [7]	100.00
	Model 2 [10]	88.75
	Model 1 [1]	73.75
Entertainment	RoBERTa	100.00
	SVM [1]	61.00
	SGG [7]	96.20
	Model 2 [10]	76.25
	Model 1 [1]	68.25

V. CONCLUSION

In this paper, the problem we tackled was fake news detection. In comparison with the usual fake news research that is based on political news, we took a step further. We focused on news articles from multiple domains to ensure that the model we train is both robust and usable in the real world. Machine learning on multi-domain and cross-domain is hard as data distribution varies in the training and the test sets. We used transfer learning models because data distribution changes do not impact them much. This synergy of multi-domain and transfer learning bore fruits. We achieved 84% accuracy on Celebrity Dataset and 99% accuracy on the FakeNews AMT dataset, a significant 6 percent and 4 percent improvement from previous work.

Talking about the limitations, we would say that this research gave us the proof of concept. Transfer learning and transformers are huge fields in themselves. Improvements could be done while training the transformer. We used general vocabulary, future researchers can use more specific words to train. The other limitation is the cross-domain. Even though the

results are state of the art, there is still a huge scope of improvement there.

VI. FUTURE WORK

If we analyze our work closely, it is pointing towards two ideas. The first is fake news detection in multi-domain settings. This space has been tried before but on a tiny scale. In the future, there should be more research work focusing on the multi-domain setting, which is more relevant to the real world. We also attempted the cross-domain settings where our results are better than any previous attempts, but still, there is a lot of scope of improvement in that aspect. The second idea is the use of transfer learning to achieve our goal. This idea gave us outstanding results. In the future, more transformers can be fine-tuned specifically to the fake news detection task, and other transfer learning techniques could also be tried. Ultimately, the aim of any future work should be a robust and relevant fake news detection algorithm.

REFERENCES

- [1] Veronica P´erez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In Proceedings of the 27th International Conference on Computational Linguistics, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
- [3] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237, 2019.
- [4] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.
- [5] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907.11692.
- [6] He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. ArXiv, abs/2006.03654.
- [7] Gautam, A., & Jerripothula, K. R. (2020). SGG: Spinbot, Grammarly and GloVe based Fake News Detection. 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), 174–182.
- [8] Wenpeng Yin and Dan Roth. 2018. TwoWingOS: A two-wing optimization strategy for evidential claim verification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 105–114, Brussels, Belgium. Association for Computational Linguistics.
- [9] V. Slovikovskaya, “Transfer Learning from Transformers to Fake News Challenge Stance Detection (FNC-1) Task”. arXiv preprint arXiv:1910.14353, 2019.
- [10] Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake News Detection Through MultiPerspective Speaker Profiles. Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2:252–256.
- [11] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. 2019. Fake News Detection on Social Media using Geometric Deep Learning. arXiv preprint arXiv:1902.06673 (2019).
- [12] William Yang Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), volume 2, pages 422–426, 2017.

- [13] Jan Christian Blaise Cruz, Julianne Agatha Tan, and Charibeth Cheng. 2019. Localization of Fake News Detection via Multitask Transfer Learning. arXiv preprint arXiv:1907.00409.
- [14] Singhal, Shivangi & Kabra, Anubha & Sharma, Mohit & Shah, Rajiv Ratn & Chakraborty, Tanmoy & Kumaraguru, Ponnurangam. (2020). SpotFake+: A Multimodal Framework for Fake News Detection via Transfer Learning
- [15] Rodriguez, Alvaro and Lara Iglesias. (2019) "Fake News Detection Using Deep Learning." <https://arxiv.org/pdf/1910.03496.pdf>
- [16] Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In The ThirtyThird AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 6859–6866.
- [17] Barua, R., Maity, R., Minj, D., Barua, T., & Layek, A. K. (2019). F-NAD: An Application for Fake News Article Detection using Machine Learning Techniques. 2019 IEEE Bombay Section Signature Conference (IBSSC). doi:10.1109/ibssc47189.2019.8973059.
- [18] Saikh, T., De, A., Ekbal, A., & Bhattacharyya, P. (2020). A Deep Learning Approach for Automatic Detection of Fake News. ArXiv, abs/2005.04938.
- [19] Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A Hybrid Deep Model for Fake News Detection. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.
- [20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. ArXiv, abs/1706.03762.
- [21] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

Multimedia Data Summarization Using Joint Integer Linear Programming

Sidhant Allawadi

Department of Computer Engineering
Delhi Technological University
New Delhi, India
sidhantallawadi_2k17co341@dtu.ac.in

Ritika

Department of Computer Engineering
Delhi Technological University
New Delhi, India
ritika_2k17co272@dtu.ac.in

Vivek Rana

Department of Computer Engineering
Delhi Technological University
New Delhi, India
vivekrana_2k17co381@dtu.ac.in

Minni Jain

Department of Computer Engineering
Delhi Technological University
New Delhi, India
minnijain@dtu.ac.in

Abstract—In recent years, there has been a massive increase in multimedia data due to the increasing use of social media and communication technology. So, extracting useful information from a large set of data has become difficult and very time consuming. Multimedia Data (Text, Image, Audio, Video) summarization is a very useful technology that overcomes this challenge by eliminating information that is redundant or useless, and extracting only the relevant key details of the events in summaries. A lot of work has been done in this field to generate summaries in the form of text and images, but very limited research has been done to produce a multimodal summary especially on Asynchronous Data. This research work proposes an ILP based model, which takes a multimodal dataset (text, images, videos) as input and generates textual and image video summary as output. The results were obtained by comparing the two basic baseline's ROUGE values with the proposed model. Results for different modalities have confirmed that the proposed model performs better than the other baseline approaches.

Keywords — Asynchronous data, multimodal summarization, integer linear programming, ROUGE, Hybrid Gaussian-Laplacian Mixture Model (HGLMM), VGG-19 model

I. INTRODUCTION

Multimedia data (counting text, image, audio and video) has greatly increased since late, making it difficult for consumers to access substantial data. In this fast and busy world, people always seek to manage their time and try to gain more knowledge when compared with the required time. Thus, more efforts are directed towards generating the summaries of the multimedia data. Multi-modal summaries can provide consumers with written documents, images & videos which gives a full overview of the topic. This technique not only saves the users' time, but also saves the effort of understanding the complicated reports or going through the entire records. A multi-modal summary has several advantages over a unimodal summary. It generates a variety of perspectives on the same matter and has various modes to represent the generated summary so that the viewer can have the freedom to choose his

way to gain that particular information [12]. A large number of viewers can benefit with multimodal summaries.

There are few things which we observed that were not considered in past researches:

- 1) Most of the work has been done on synchronous data, therefore brings consideration towards asynchronous data.
- 2) Most of the work focuses on unimodal summarization, and only a few researches have explored multimodal summaries of multimedia data.

Thus we felt the need to generate a multimodal summarization method for multimedia data.

We have taken few aspects into consideration:

- 1) High relevance of the content in the generated summary.
- 2) The summary should be readable.
- 3) The redundancy should be as low as possible.
- 4) All the essential parts of the content should be covered in the generated summary.

To sum up, the major contributions of our work are as follows:

The asynchronous multimedia (text, video, image) data are considered as input and propose a multimodal summarization method which generates summaries with diverse representation.

- (1) A joint ILP framework is formulated to perform the proposed multimodal summarization method.
- (2) Different baselines are considered to compare the proposed method with the past research work.

Flowchart Description:

- (1) Given dataset consists of documents, images, audio and videos.
- (2) Documents and audio (converted using IBM Watson Speech-to-Text Service) forms the document set.
- (3) Images and videos (key frames extracted from videos) form the image set.
- (4) From the document set each document is taken and each sentence of that document is encoded (using HGLMM) resulting in encoding of the whole document.

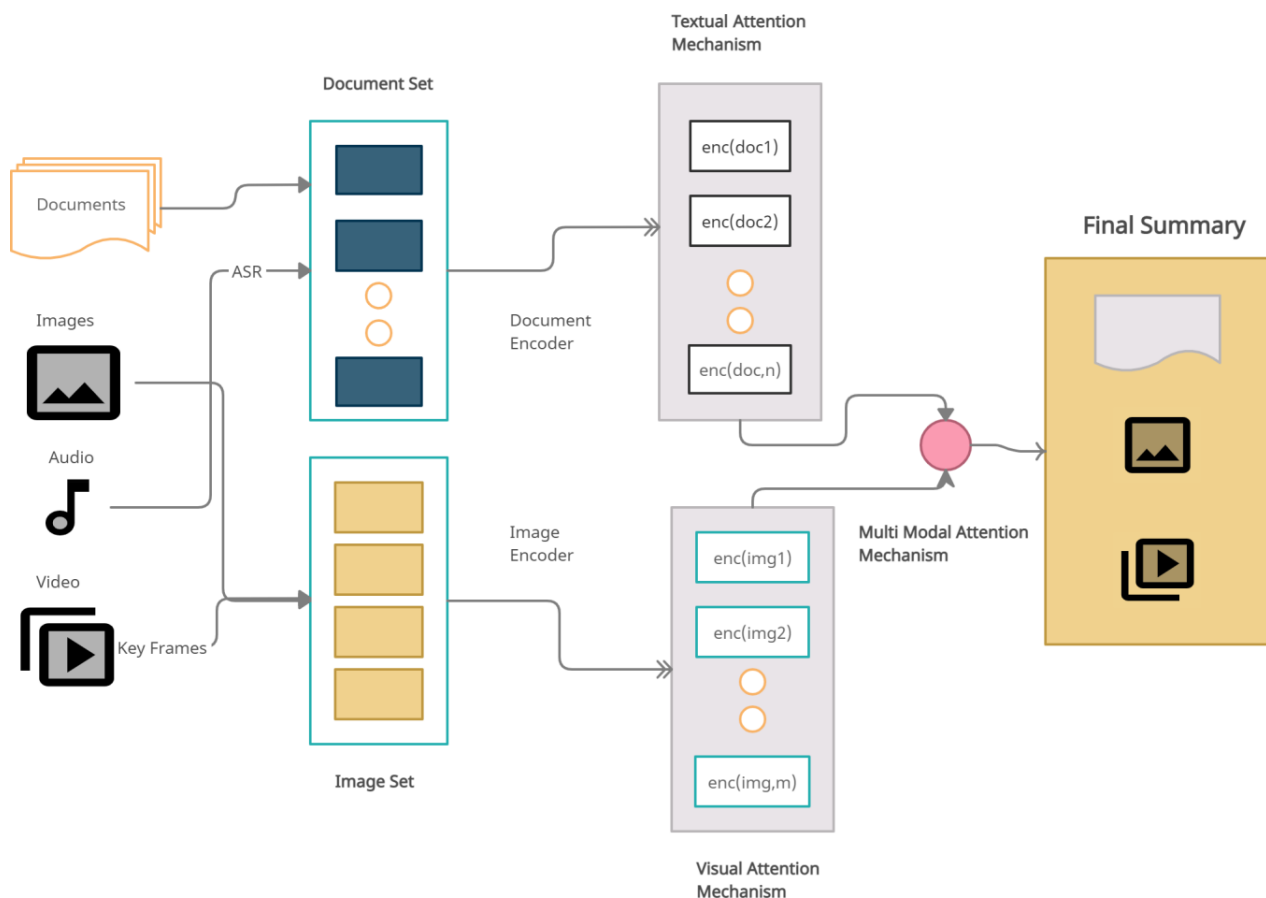


Fig. 1. Working of the proposed models

- (5) All the documents are encoded in this way to form a textual attention mechanism.
- (6) Similarly, all the images are encoded (using VGG-19) forming a visual attention mechanism.
- (7) These are fed into a multi multi-modal attention mechanism to get the final summary.

II. RELATED WORK

Past research has shown great insights in text-image summarization with a few researches to produce multimodal summaries [2], [5], [14]. Much work has recently been undertaken to sum up events, sport videos, films, pictorial stories and social media. Genetic Algorithms have been used for text summarizations [9]. Erol et al. (2003) aimed at the development of main audio, text and visual behavior analysis segments of the meeting recording [4]. Tjondronegoro et al. (2011) suggested a way of analyzing textual information from various databases and recognizing relevant content from a sports video in order to summarize a sporting event [11]. In order to identify salient events in a film, Evangelopoulos and

others (2013) used an attention mechanism. For the generation of a pictorial history and time-line description, Wang et al. (2012) and Wang et al. (2016b) used the image-text pairs. Li et al. (2016) created a multimedia news summary method for Internet search results that introduces the hLDA model to discover the newspaper subject structure [13]. A news article and a picture will then be selected to represent each subject. A lot of work has been done to summarize recordings of conferences [4], sporting videos [11], films, pictorial history and social contact. Duan et al. introduced a framework using joint-ILP to produce text summaries of text documents separated temporally [3]. Automated text report is a major NLP application which aims at summarizing the content of a given matter in a shorter model. The rapid growth in the transfer of data across the network includes multimodal reporting (MMS) from asynchronous text-image-audio-video combinations. Analysis reflects a similar MMS system that uses information science and speech process technology to analyze and reinforce aspects of transmission news reporting on complex data in multi-modal statistics. The key structure is

to link the linguistic gaps between multimodal material. The audio signals are converted to the content format for audio themes. Text is extracted using the OCR photo victimization technique for visual subject matter. The template is created afterwards for the required visual details by matching the material or by multimodal theme modeling. Finally, all multimodal considerations evaluate the thinking by optimizing the value, non-reinforcement, efficiency and reach of the allocated accumulation of submodular choices. The work is structured to discuss the theme of the visual subject. The experimental result demonstrates that different competitive techniques output the multi-modal report system.

III. PROPOSED METHODOLOGY

A. Pre-processing

We have several images, text documents and videos as inputs for a specific subject augmented from [6]. We take key-frames extracted from the videos [15] and merge them with the given pictures as inputs to make the image collection in order to acquire essential features from the raw/initial staged data. The audio is converted to text using IBM Watson's Speech-to-Text Service. The resulting transcriptions together with text then forms the text set. The following models are used for the next steps of our preprocessing.

- Hybrid Gaussian-Laplacian mixture model (HGLMM) which is proposed in [1], based on Laplacian and Gaussian distribution's weighted geometric mean, is used for encoding text in text-set. With the use of HGLMM we are getting benefits of both the distributions (Laplacian and Gaussian) as each dimension from each component is going to be modeled with the appropriate distribution.
- VGG-19 model [10] is used for encoding images in image-set. We first studied CNN (Convolutional neural network) then VGG-16 model and then VGG-19, VGG-19 is the most advanced model to give more accurate results.

These model specific encodings are next loaded to a two branch neural-network [8] which finally gives 512-dimensional sentence and image vectors.

B. Main Model

We are proposing a model which uses ILP framework (a framework which is used to maximise/minimise an objective function in accordance with some constraints) which generates summaries with diverse representation of the given dataset. To the best of our knowledge, this model gives better results than the previous works done for multimodal summarization.

C. Decision Variables:

The decision variables in an integer linear program are a sequence of quantities which are required for solving the problem, i.e., completing the tasks when the best values of

the variables are found.

The decision variables used in our model are:

$$M_{txt} = [m_{ij}^{txt}] \quad n \times n \text{ binary square matrix} \quad (1)$$

$$M_{img} = [m_{ij}^{img}] \quad p \times p \text{ binary square matrix} \quad (2)$$

$$M_c = [m_{ij}^c] \quad n \times n \text{ binary square matrix} \quad (3)$$

$M_{i,i}^x$: (sentence s_i /image I_i) is an exemplar or not;

x : txt, img

$M_{i,j;(i \neq j)}^x$: whether x_i votes for x_j as its exemplar;

x : txt, img

$M_{i,j}^c$: if there is a threshold level of correlation between i -th sentence and j -th image,
 c : cross-model

D. Objective Function:

$$f(x) =$$

$$\begin{aligned} & \text{Argmax} \{ \lambda_1 * m * k_{txt}^2 * ([\sum_{i=1}^n M_{txt_{i,i}} * SIM_{cosine}(s_i, O_{txt})]^{(1)} \\ & + [\sum_{i=1}^n M_{img_{i,i}} * SIM_{cosine}(s_i, O_{img})]^{(2)} + \\ & \lambda_2 * (k_{txt} + k_{img}) * k_{txt}^2 * ([\sum_{i=1}^n \sum_{j=1}^p M_{i,j}^c * SIM_{cosine}(s_i, I_j)]^{(3)} \\ & - \lambda_3 * (k_{txt} + k_{img}) * m * ([\sum_{i=1}^n \sum_{j=1}^n M_{txt_{i,i}} * M_{txt_{j,j}} * SIM_{cosine}(s_i, I_j)]^{(4)}) \} \end{aligned} \quad (4)$$

Parameters:

$$\lambda_1 \lambda_2 \lambda_3 : \text{Variables to define weight of equations.} \quad (5)$$

$$O_x : \text{Central vector of the cluster } x : \text{txt, img} \quad (6)$$

$$m : \text{Used to balance out the weights of } \lambda_1 \lambda_2 \lambda_3 \quad (7)$$

$$SIM_{cosine} : \text{Cosine similarity} \quad (8)$$

Number of (sentence, images) in the final summary, there must be exactly k_{txt} and k_{img} clusters in their respective uni-modal space.
 x : txt, img

$$k_x : \quad (9)$$

$$S_i : \text{ith sentence vector} \quad (10)$$

$$I_j : \text{jth image vector} \quad (11)$$

The components of equation are:

- (1) gives the salience score of the text-set.
- (2) gives the salience score of the image-set.
- (3) gives the cross-modal correlation score
- (4) gives the redundant part of the summary.

In our function, we are adding salience score of the text-set and image-set, cross-modal correlation score and subtracting the redundant part. Our aim is to maximise this objective function using ILP while following constraints given in the next section.

E. Constraints

- $m_{ij}^x \in \{0, 1\}; x \in \{\text{txt}, \text{img}, \text{c}\}$
#All the variables are binary
- $\sum_{i=1}^p m_{ij}^{\text{img}} = k_{\text{img}}$
#number of image clusters
- $\sum_{i=1}^n m_{ij}^{\text{txt}} = k_{\text{img}}$
#number of text clusters
- $\sum_{j=1}^n m_{ij}^{\text{txt}} = 1; i \in \{1, \dots, n\}$
#for a sentence s_i , it could either be an exemplar or a part of another cluster
- $\sum_{j=1}^p m_{ij}^{\text{img}} = 1; i \in \{1, \dots, p\}$
#for an image I_i , it could either be an exemplar or a part of another cluster
- $m_{ij}^x - m_{ij}^y \geq 0; X \in \{\text{txt}, \text{img}\}$
For a sentence or image to be a part of final (txt-img) summary, each of them must be an exemplar in their respective categories

F. Post Processing

The output of the ILP framework used in the model is a text summary and top n (n is variable, we take $n=10$ for convenience) images. This is used to gear up the final video and image summary.

The methods used for their extraction are:

1) Extracting images::

- Sort all those images which are not key-frames from the top 10 images.
- Since the redundancy was removed in the objective function itself, to get a high annotation score for the images and support users to get a better understanding, the range of similarity was set.
- In the model, range for similarity was set between 0.3 and 0.7, i.e, images with minimum of 30% similarity and maximum of 70% of similarity were added to the final image summary.

2) Extracting Videos::

- We take key frames and speech transcriptions of each video.
- Then we calculate cosine similarity of key frames with the generated image summary and named the visual score.
- Similar process is carried out for the audio transcriptions and then named its verbal score.
- The final video summary was the one with the the highest visual and verbal score.

IV. RESULTS

TABLE I
PERFORMANCE FOR GENERATING TEXTUAL SUMMARY
USING ROUGE VALUES

Systems	ROUGE L	ROUGE 2	ROUGE 1
Baseline 1	0.221	0.061	0.257
Baseline 2	0.223	0.072	0.261
Main Model	0.230	0.076	0.265

The performance of the model is compared with the 2 baselines. These are:

1) Baseline-1: :

- Input: sentence vectors
- Central vector: average of all the sentence vectors

2) Baseline-2: :

- Input: sentence and image vectors.
- Central vector: weighted average of sentence and image vectors.

The preprocessed data which has been driven from [14] was then modeled with the above baselines. Table 1 and Table 2 compares the performance of our model with these baselines.

ROUGE, (Recall-Oriented Understudy for Gisting Evaluation) values are used to analyse the textual summary [7]. ROUGE values for ROUGE-L, ROUGE-2, ROUGE-1 systems have clearly indicated that our proposed model is performing better than the baseline approaches as the values are higher for our model. It is clear from table 1, our model performs better than these baselines taken into account for comparison.

In table 2, Variance Recall, Variance Precision, Average Recall and Average Precision was calculated to analyse the performance of the proposed model. Our model performed better for low threshold values. The accuracy of the model for extraction of the most suitable video for summary was 47%. The random selection of images from the dataset for 20 iterations gave 18% accuracy.

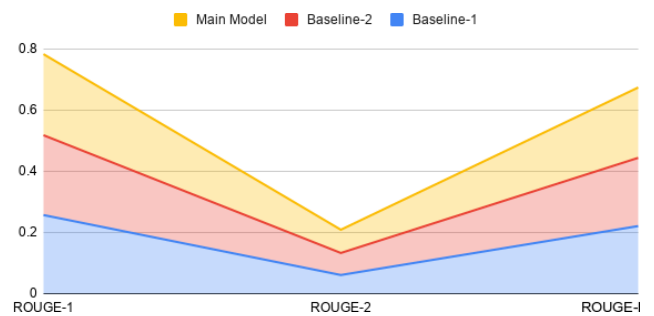


Fig. 2. Performance for generating textual summary using ROUGE values

TABLE II
PRECISION AND RECALL OF IMAGE SUMMARY.
AAS (AVERAGE ANNOTATION SCORES) IS THE THRESHOLD VALUE OF GENERATED IMAGE SUMMARY

AAS	Variance Recall	Variance Precision	Average Recall	Average Precision
5	0.046	0.004	0.064	0.018
4	0.155	0.110	0.320	0.260
3	0.80	0.140	0.385	0.605

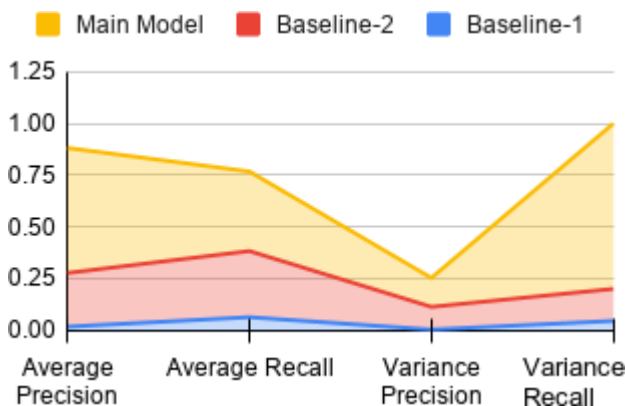


Fig. 3. Precision and Recall of Image

V. CONCLUSION AND FUTURE SCOPE

In this paper we proposed an Integer Linear Programming based model for extractive multimodal summarization. Our results show that the proposed model surpasses the existing baseline approaches. The proposed approach has shown remarkable achievements but still there are some areas which we would like to consider in further research. The model is an extractive summarization technique which doesn't work on the relationship between the concept and images/sentences. We would be working on that to move forward in the direction of an abstractive summarization technique. With the abstractive summarization technique, we will also be working on sentence compression in our generated summaries.

REFERENCES

[1] Gil Sadeh Benjamin Klein, Guy Lev and Lior Wolf.: *Fisher vectors derived from hybrid gaussian-laplacian mixture models for image an-*
[7] C.Y.: Lin. *ROUGE: A package for automatic evaluation of summaries*. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain, Jul 2004.

notation. In: arXiv preprint arXiv:1411.7399, 2014.
[2] Zhuge H.: Chen, J. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4046–4056, 2018. Abstractive text-image summarization using multi-modal attentional hierarchical rnn.
[3] Jatowt A.: Duan, Y. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. pp. 735–743. ACM, 2019. Across-time comparative summarization of news articles.
[4] Lee D.S. Hull J.: Erol, B. In: 2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698). vol. 3, pp. III–25. IEEE, 2003. Multimodal summarization of meeting recordings.
[5] Zhang J. Zhu J. Liu T. Zong C. et al.: Li, H. In: *Multi-modal sentence summarization with modality attention and image filtering*. 2018.
[6] Zhang J. Zong Zhu J. Ma C. C. et al.: Li, H. *Multi-modal summarization for asynchronous collection of text, image, audio and video*. 2017.
[8] Yin Li Liwei Wang and Svetlana Lazebnik.: In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. Learning deep structure preserving image-text embeddings.
[9] Anubhav Jangra Naveen Saini, Sriparna Saha and Pushpak Bhat-tacharyya.: In: Knowledge-Based Systems 164 (2019), 45s' 67. Extractive single document summarization using multi-objective optimization: Exploring self-organized differential evolution, grey wolf optimizer and water cycle algorithm.
[10] Karen Simonyan and Andrew Zisserman.: In: International Conference on Learning Representations, 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition.
[11] Tao X. Sasongko J. Lau C.H.: Tjondronegoro, D. In: Applications of Computer Vision (WACV), 2011 IEEE Workshop on. pp. 471–478. 2011. Multi-modal summarization of key events and top players in sports tournament videos. IEEE, 2011.
[12] Bigham J.P. Allen J.F.: UzZaman, N. *Proceedings of the 16th international conference on Intelligent user interfaces*. pp. 43–52. ACM. 2011.
[13] X. Wang J. Liu Z. Li, J. Tang and H. Lu.: In: ACM Transactions on Intelligent Systems and Technology, vol. 7, no. 3, p. 33, 2016. Multimedia news summarization in search.
[14] Li H. Liu T. Zhou Y. Zhang J. Zong C.: Zhu, J. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing pp. 4154–4164, 2018. Multimodal summarization of complex sentences
[15] Rui Y. Huang T.S. Mehrotra S.: Zhuang, Y. *Adaptive keyframe extraction using unsupervised clustering*. In: Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269). vol. 1, pp. 866–870. IEEE, 1998.

Multi-modal biometric recognition system based on FLSL fusion method and MDLNN classifier

¹Ajai Kumar Gautam, ²Rajiv Kapoor

¹²Department of Electronics & Communication Engineering, Delhi Technological University, Delhi , India

¹ajai.gautam@gmail.com, ²rajivkapoor@dce.ac.in

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: A Multi-Modal Biometrics (MMB) system incorporates information as of more than ‘1’ biometric modality for enhancing each biometric system’s performance. Numerous prevailing research methodologies focused on MMB recognition. However, the recognition system encompasses robustness, accuracy, along with recognition rate issues. This paper proposed the MMB recognition system centred on the FLSL fusion method and Modified Deep Learning Neural Network (MDLNN) classifier in order to enhance the performance. The face, ear, retina, fingerprint, and front hand image traits are considered by the proposed method. It comprised image enhancement, segmentation, Feature Extractions (FE), Feature Reduction, feature fusion, rule generation, and identification phases. The Improved Plateau Histogram Equalization (IPHE) algorithm enhances all the inputted traits. After that, Viola-Jones Algorithm (VJA) segmented the facial parts, and the Penalty and Pearson correlation-based Watershed Segmentation (PPWS) algorithm eliminates the unwanted information in the ear and finger traits and also segmented the blood vessel of the retina image. Region of Interests (ROI) calculation separates the palm region. Next, the features are extracted as of images, and then, the Kernelized Linear Discriminants Analysis (KLDA) algorithm reduces the features’ dimensionalities. Next, the Features Level and Scores Level (FLSL) fusion method fuse the features. Therefore, the features’ fused output is inputted to the MDLNN to classify the person as genuine or imposter. The investigational evaluation of the proposed MDLNN with the prevailing classifiers is analyzed. The proposed MDLNN centred MMB recognition system trounces the top-notch methods.

Keywords: Improved Plateau Histogram Equalization (IPHE), Viola-Jones Algorithm (VJA), Penalty and Pearson correlation-based Watershed Segmentation (PPWS), Kernelized Linear Discriminant Analysis (KLDA), Feature Level and Score Level (FLSL), and Modified Deep Learning Neural Network (MDLNN) algorithm.

1. INTRODUCTION

Biometrics could be found anywhere from unlocking of mobiles to airport border control recently. A modern system that utilizes the information of biometrics as of one person aimed at authentication along with verification is a biometric system [1]. For recognizing individuals, the technology of utilizing humans’ physical along with behavioral characters is a biometrics system [2]. BR systems are utilized in multiple areas for recognition [3]. Besides different other factors namely cost, convenience, security level, memory requirements, etc, the BR system is also centered upon its verification or identification accuracy [4]. The genuine and imposters by means of uni-modal biometrics and MMB methods are recognized by this method. Different issues namely non-universality, intra-class variations, noise in input data, spoof attacks, along with distinctiveness are possessed by these uni-modal systems. Due to the user’s poor interaction with the sensor, these variations take place [5]. The usage of numerous biometric modalities (i.e., the combination between two or several different biometrics data or combining between the physiological along with behavioral characteristics) within the same system is the solution for overcoming these disadvantages, which is named as a multiple biometric system [6, 7]. Thus, greater attention to MMB was given by most researchers for increasing identification performance and providing more security [8].

The inherent problems of user’s inconvenience along with system inefficiency are solved by the MMB systems [9]. MMB systems can manage the issues of non-universality and can limit imposters from spoofing biometric attributes of authentic people. Thus, it could fulfill challenges [10]. Sometimes, the usage of disparate identities of the same trait rather than utilizing the details of disparate modalities is an MMB. For example, finger shape along with vein, fingerprints as well as finger knuckle point print of an individual human finger are the amalgamation of different traits, which are employed in finger multimodal authentication [11]. A biometric system centered upon the biometric attributes of the traits. The unique features are utilized for comparison along with matching which is sorted out as of the biometric data (raw) [12]. For improving performance, the features from different traits are merged together. Therefore, the main method involved in MMB is a fusion [13]. An amalgamation of disparate features of traits is the definition of multiple modal fusions [14]. This biometric system’s accuracy is considerably affected by the scheme of fusion, which is effective.

Feature-level, sensor-level, Rank-Level Fusions (RLF), matching Score Level Fusions (SLF), along with decision-level fusions are the ‘5’ modules in the MMB system through its fusion [15] [16]. Different classifiers

and predictors with the estimators are utilized for fusing the information largely after fusion and are composed of '3' types namely fusion before matching, fusion at the time of matching along with fusion after matching called, pre-mapping, midst mapping, along with post mapping respectively [17]. Further investigation of different stimulating factors namely privacy, cost, accuracy, selection of powerful biometric traits, system complexity, easy to use, etc. is performed as per the fusion outcome [18]. In the previous ten years, numerous MMB systems centered on conventional traits, namely fingerprint, and iris, are developed. Ear, palm print geometry, finger geometry, as well as retina are the few works about an MMB system. An MMB authentication system centered on the MDLNN algorithm is proposed by this research methodology for improving the recognition's accuracy and making the MMB authentication system quicker. The face, ear, retina, fingerprint, palmprint with palmprint geometry, along with finger geometry traits are integrated by the presented MMB. This paper is categorized as: Section 2 described the top-notch methods of the multi-modal biometric. Section 3 elucidates the proposed methodology of the MMB recognition. Section 4 examined the proposed methodology's performance with the existing research technique. Section 5 completes the paper.

2. RELATED WORK

Nada Alay [20] introduced an MMB aimed at recognizing humans by biometric modes of face, iris, along with the finger vein centred upon a deep learning (DL) algorithm. For increasing training data along with reducing overfitting problems, pre-processing actions were initially executed on the images, namely resized the image utilizing the Visual Geometry Groups (VGG-16) as well as data augmentation was employed. Next, every biometric trait was given to its Convolutional Neural Networks (CNN) design. After that, feature level, along with SLF methods were utilized for fusing the '3' CNN's multi-modal models (iris, face, along with finger vein). The approach had comfortably outshined the top-notch methods as shown by the experimental outcomes. The approach's drawback was not implemented with different level fusion methods as it was inappropriate for complementary traits.

Meryem Regoud *et al.* [21] introduced an MMB system aimed at human identification and security centred on the local textures descriptors. For eradicating the redundant data, normalization along with segmentation was the pre-processing method applied to Electrocardiography (ECG), ear, and iris biometrics. After that, for extracting the significant features as of the ECG signal as well as converting the ear along with iris images to 1D signals, 1D-Local Binary Patterns (1D-LBP), Shifted-1D-LBP, along with 1D-Multi-Resolution-LBP were employed. K-Nearest Neighbours (KNN) and the Radius Basis Function (RBF) were employed for matching to categorize an unidentified user as the genuine or else the impostor. The approach had performed well with the MMB system by different classification methods as shown by the experimental outcomes. The KNN was utilized by the approach which was inappropriate for numerous data. The system's performance was degraded centred on distance calculation.

Gurjit Singh Walia *et al.* [22] presented an MMB system centred on an optimum SLF model. Iris, fingerprint, together with the finger's vein was the integrated '3' complementary biometric attributes. Gabor transform was implemented on the image and the following output was split as non-overlapping rectangle blocks. Aimed at every block, the LBP's histogram was determined and these extracted histograms were concatenated for forming feature vectors. Therefore, the authenticate person was identified by the feature's vectors given to the individual classifier. For acquiring a simultaneous solution, individual classifiers were resolved by Proportional Conflict Redistributions rules (PCR-6). The score dissemination of imposter together with the genuine class was large which results in high accuracy along with reliability for this approach as shown by the investigational results. The information was directly taken by the framework as of the input; then given onto the classifier that could affect the recognition's rate as a consequence of the noises prevalent in the inputted image.

S. Prabu *et al.* [23] specified a multimodal authentication for BR System centred on Intelligent Hybrid Fusion Techniques. Firstly, the image was pre-processed via the median filters for eliminating the noises and smoothening images. The resulting images were inputted to the Discrete Curvelet Transforms for better clarity. For feature detection, Effective linear Binary Patterns (ELBP) along with Scale-Invariant Fourier Transforms (SIFT) were employed. After that, the genuine, and the imposter were verified via the Extreme Learning Machines (ELM) classification method. The authentication system had attained higher accuracy by fusion method as demonstrated by the experiential outcomes. Because of the ELM classification technique, the method's robustness was not efficient.

Gaurav Jaswal *et al.* [24] introduced an MMB Authentication System by Palm Print, Hand Shape, along with Hand Geometry. The image (original) was altered as grey-scale at the pre-processing procedure and then, a 2D Gaussian filter was implemented for reducing the noise and other irregularities. Next, certain rotation along with illumination effects was experienced by the extracted palm's ROI samples that limited the corresponding performance. Utilizing the Speeded-Up Robust Features (SURF) descriptor, the transformed ROI images' local key-points had been taken out. Sub-pattern-centred PCA (SpPCA) and Support Vector Machine (SVM)-centred classification method was utilized for better recognition. A multimodal recognition system centred on the feature-level fusion of the normalized features of the palm print, hand shape, along with the hand geometry traits

had achieved better accuracy as exhibited in the experimental outcomes. This system was a complex process for recognizing the damaged contour parts of the traits.

K. Gunasekaran et al. [25] introduced a deep multimodal BR by the contour-let derivative weighted RLF with fingerprint, human face along with iris. The pre-processing was initially executed using the Contour-let Transform design. Then, the Local Derivative Ternary Patterns was implemented in the pre-processed attributes. The obtained coefficients were utilized for improving the feature discrimination power. The biometric matching scores as of numerous modalities were effectively combined by the multimodal features (extracted) to which the Weighted RLF was implemented. For improving the MMB system's recognition rate within the temporal domain, a DL framework was implemented. The investigational outcomes had demonstrated that the system had attained better results in MMB authentication. The approach's drawback was not efficient as the full image of traits was deemed and then the images were fused. The recognition's rate was affected since the redundant parts prevalent in the traits also were pondered.

3. PROPOSED MULTI-MODEL BIOMETRIC AUTHENTICATION

In the past decennia, one of the major domains of security systems has turned out to be Biometrics. Recently, the utilization of automated biometrics-centred personal recognition systems has turned out to be an omnipresent procedure. Nevertheless, several problems, such as illumination variation, noisy data, spoofing, pose variation, partial occlusion, and non-universality are confronted by the uni-modal biometrics, which brings about less accurateness and security. MMB identification is a propitious alternative to trounce few of these cons and for augmenting the level of security. This research method proposed an MMB authentication system centred on the MDLNN algorithm, which considers the face, ear, retina, fingerprint, palm print, the geometry of the palm together with the fingers. Initially, IPHE enhances all the inputted images. Next, the segmentation process is executed in which the VJA segments the face image, and also the PPWS eradicates the unnecessary information of the ear and finger image in addition to the blood vessels are segmented as of the retina. Additionally, the ROI calculation separates the palm region as of the front hand image. As of every trait, the features are extracted. After that, the KLDA algorithm reduces the features' dimension. After that, the FLSL fuses the features, which are then inputted to the MDLNN. Centred on the generated rules, the person is classified as genuine or imposter by the MDLNN (explicitly, the features are trained and tested centred on that rule). Here, more hidden layers are employed in DL neural network, and the Stain Bowerbird Optimization (SBO) algorithm selects the optimal weight value for attaining higher accuracy. The proposed method's block diagram is exhibited in Figure 1,

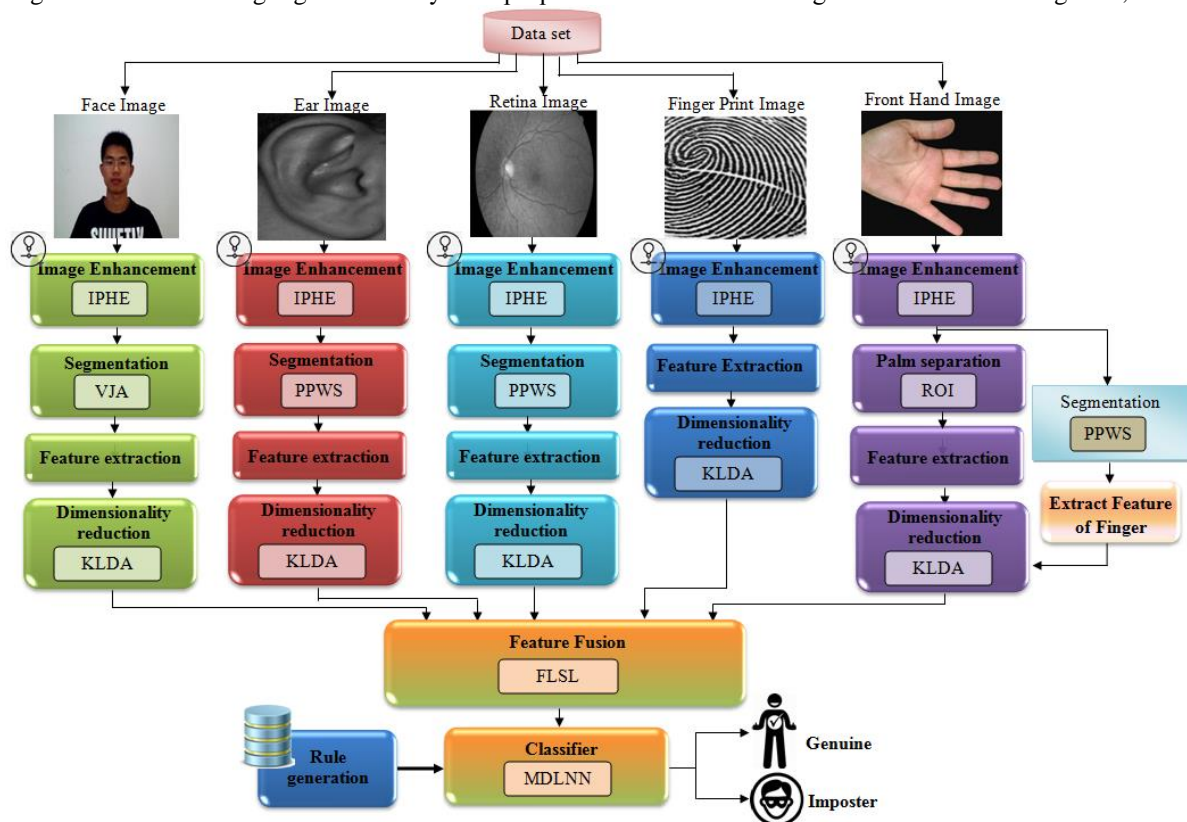


Figure 1: Block diagram for the proposed methodology

3.1 Image Enhancement

Initially, the IPHE algorithm enhances the input trait: face, ear, retina, finger-print, and front hand images. For the extraction of information as of the image, the image's contrast must be ameliorated. Over enhancement is a major issue for most contrast enhancement methods. Thus, the gamma correction function is considered here to evade that issue. The IPHE follows these steps: the images' pixel values are arranged (ascending order). Next, the histogram building will be generated, followed by which, the histogram median value is calculated, which is round of to the nearest integer value (i.e. threshold value). After that, the EXOR operations of equivalent '2' histogram values are performed and the values are considered as Cumulative Distributions Function (CDF). The image's histogram equalization is gauged as:

$$H_e = \frac{CDF}{A_p * N_o} \quad (1)$$

Wherein, H_e signifies the histogram equalization, A_p implies the entire number of pixels, N_o implies the number of output precise. At that moment, the over enhancement problem will occur. Thus, the gamma correction is employed to regulate the intensity that is rendered as,

$$M_{i(out)} = \omega M_{i(in)}^{\chi} \quad (2)$$

Wherein, $M_{i(in)}$ and $M_{i(out)}$ signify the input as well as output image intensities, correspondingly, ω and χ imply '2' parameters that control the transformation curve's shape.

3.2 Segmentation

Segmentation of traits, say, the face, ear, retina, finger, and palm is performed subsequent to image enhancement. Here, the VJA segments the face parts, and the PPWS algorithm takes care of the ear, finger (explicitly, unnecessary information elimination), and retina. In addition, via gauging the ROI, the palm is attained.

3.2.1 Face segmentation by VJA

VJA is robust and its face detection in practical situations is faster, thus it is preferred. Here, only the face parts (left and right eye, nose, lips, as well as eyebrows) are segmented. There are totally '4' section (i) Haar Features Selection, (ii) Generating an integral image, (iii) Adaboost Training, as well as (iv) Cascading.

Haar Feature Selection: Haar features are categorized into: a) '2'-rectangle features, which stands as the difference betwixt the sums of the pixels among '2' rectangular areas, b) '3'-rectangle features gauges the sum of pixels among '2' outside rectangles and is deducted as of the sum of pixels on the centre rectangle and c) '4' rectangle features gauges the difference betwixt the diagonal pairs of the rectangle.

Integral Image Computation: In the image, the integral image value of any point is equivalent to the sum of the entire pixels on the upper left corner of the point. The integral image at u, v encompasses the sum of the pixels above as well as to the left of u, v , inclusive:

$$in(u, v) = \sum_{u' \leq u, v' \leq v} G_i(u', v') \quad (3)$$

Wherein, $in(u, v)$ signifies the integral image and also $G_i(u', v')$ implies the original image, $i = 1, 2, \dots, n$, the face image is signifies as G_1 . The recursion formula is employed in the integral computation, which is described as,

$$cu(u, v) = cu(u, v-1) + G_i(u, v) \quad (4)$$

$$in(u, v) = in(u-1, v) + cu(u, v) \quad (5)$$

Wherein, $cu(u, v)$ signifies the cumulative row sum, the integral image can well be gauged on one pass above the image (original).

Adaboost Training: Adaboost algorithm eradicates redundant features and converts numerous features into a compact one. It stands as a learning classification function. Aimed at representing a face, the most meaningful ones are the chosen features. Several thousands of features can be lessened to a few hundred features by this algorithm.

Cascading: The cascaded classifier stands as a compilation of stages that encompasses a stronger classifier. Every phase verifies whether a specific sub-window is definitely not a face or maybe a face. If a specified phase classified a sub-window as a non-face, it will be discarded; whereas, if it is classified as a maybe face, it is sent to the succeeding stage on the cascade.

3.2.2 Palm separation

As of the improved front hand image, the palm's area is computed by computing the ROI region. The fingers are acquired separately in this computation. Extracting ROI is a necessary task. The ROI is computed

centred on the palm's rotation and also the region's size; consequently, the region's size is computed via the valleys' localization. The ROI's end result is articulated as,

$$ROI_{out} = \left(N - \frac{r_s}{2} \right) + 1 \quad (6)$$

Here, ROI_{out} signifies the ROI's outcome; N implies the novel rotation; r_s implies the novel region's size.

3.2.3 Ear, retina, and finger segmentation using PPWS

The occlusions, together with the other unwanted information (e.g., ear-rings, hair) are removed as of the improved ear image to acquire the actual ear region. The segmentation within the ear region is handled in the ear aimed at eradicating the redundant information. Next, the blood vessels are segmented to acquire the information in the retina's image. The segmentation procedure is implemented in the finger image aimed at the reason of eliminating unnecessary things prevalent in the image (for instance, some individual wears the rings such that the unnecessary things are eliminated). Herein, aimed at segmentation, the PPWS technique is utilized. The correlation calculation is executed in a typical watershed segmentation technique. However, it couldn't attain added information. Consequently, the Pearson correlation's computation is executed here; the over-segmentation issue is evaded via the penalty parameter. The morphological processes, like convolution, and also Pearson's correlation, are applied in this technique aimed at locating the foreground and also background detection. The convolution's arithmetic formulation is:

$$vol(G_i, k) = \sum_p \sum_q G_i \cdot k(p, q) \quad (7)$$

Herein, vol signifies the convolution function; G_i implies the inputted image ($i = 1, 2, \dots, n$); the ear and retina image is signified as G_2 and G_3 ; k symbolizes the kernel. Next, Pearson's correlation is computed. The correlation is nearly alike convolution. It is enumerated as the nearby pixels' weighted summation. The correlation is equated as,

$$rel(G_i, k) = \frac{n \sum_{p,q} (u+p)(v+q) - \left(\sum_p (u+p) \right) \left(\sum_q (v+q) \right)}{\sqrt{\left[n \sum_p (u+p)^2 - \left(\sum_p (u+p) \right)^2 \right] \left[n \sum_q (v+q)^2 - \left(\sum_q (v+q) \right)^2 \right]}} k(p, q) \quad (8)$$

Here, (u, v) implies the inputted image's pixel location; (p, q) signifies the actual image's pixel location.

3.3 Feature Extraction

The features are taken as of every trait past the segmentation and also palm separation. As of the segmented face parts, segmented retina's blood vessels, fingerprint, and also ear, the Local Tetra Pattern (LTrP), Gabor feature, edge, SURF, and also Binary Robust Invariant Scalable Key-points (BRISK) features are extracted. The minute points and also cross-line points are taken out as of the fingerprint utilizing these feature descriptors. The geometric features, LTrP, Discrete Wavelet Transform (DWT), and SIFT are extracted as of the palm. Next, the finger's geometric measurement is taken out as of every finger.

LTrP: The LTrP defines the local texture's spatial structure utilizing the central grey pixel's direction. t_c signifies the G_i image's central pixel; t_h implies t_c 's horizontal neighbour; t_v signify t_c 's vertical neighbour. Next, the 1st-order derivatives prevalent at the t_c is equated as,

$$G_{i(0^\circ)}^{(1)}(t_c) = G_i(t_h) - G_i(t_c) \quad (9)$$

$$G_{i(90^\circ)}^{(1)}(t_c) = G_i(t_v) - G_i(t_c) \quad (10)$$

Next, compute the pixels' magnitude $M_{G_i(t_p)}$ utilizing,

$$M_{G_i(t_p)} = \sqrt{\left(G_{i(0^\circ)}^{(1)}(t_p) \right)^2 + \left(G_{i(90^\circ)}^{(1)}(t_p) \right)^2} \quad (11)$$

Herein, t_p signifies the image's pixels.

Gabor Feature: A '2'-dimensional Gabor function is equated as,

$$ga(u, v) = \exp\left(-\frac{u'^2 + v'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{u'}{\lambda} + \varphi\right) \quad (12)$$

$$\begin{bmatrix} u' \\ v' \end{bmatrix} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \quad (13)$$

Here, $ga(u, v)$ signifies the Gabor result; $\lambda, \theta, \varphi, \sigma, u', v'$ are the wavelet's parameters.

SURF: SURF defines a local FE technique. It utilizes a local invariant fast key-point detector to take out the image's feature key points. It utilizes a unique descriptor to take the image's feature descriptor. SURF's features are in-variant of shifting, scaling and also rotation; it is partly invariant towards illumination and also affine transformation. Herein, the Hessian Matrix (HM) is found regarding the image G_i 's each pixel position; it is arithmetically equated as,

$$H(u, \delta) = \begin{pmatrix} Z_{uu}(R, \delta) & Z_{uv}(R, \delta) \\ Z_{uv}(R, \delta) & Z_{vv}(R, \delta) \end{pmatrix} \quad (14)$$

Here, R signifies the image's point; σ is signified as scale. Generally, $Z_{uu}(R, \sigma)$ implies the convolution of the image's Gaussian 2nd-order derivative at the respective point comprising the coordinates (u, v) .

BRISK: BRISK is stated as a method aimed at scale-space Key-point's detection and also the binary description's creation. The Gauss function is utilized to decrement the grey-scale aliasing in the BRISK feature descriptor. The standard deviation sigma's Gauss function is proportional to the distance betwixt the points on every concentric circle. Picking a pair as of the point pairs created by every sampling point, signified as (L_m, L_n) ; the grey values past the treatment are $G_i(L_m, \rho_m)$ and also $G_i(L_n, \rho_n)$. Thus, the gradient betwixt '2' sampling points $gr(L_m, L_n)$ is,

$$gr(L_m, L_n) = (L_n - L_m) \cdot \frac{G_i(L_n, \rho_n) - G_i(L_m, \rho_m)}{\|L_n - L_m\|^2} \quad (15)$$

Split the pixel sets to '2' sub-sets: short separation pairs (Sh) and long-distance sets (Lo). Hence, the long, as well as short-distance pairs, are equated as,

$$Sh = \{(L_m, L_n) \in A \mid \|L_n - L_m\| < \varepsilon_{\max}\} \subseteq A \quad (16)$$

$$Lo = \{(L_m, L_n) \in A \mid \|L_n - L_m\| < \varepsilon_{\min}\} \subseteq A \quad (17)$$

Here, A signifies the compilation of all sampling points' pairs; ε_{\max} and ε_{\min} signifies the distance thresholds. Generally, the BRISK technique is utilized aimed at solving aimed at the overall pattern's direction gr regarding the gradient betwixt '2' sampling points:

$$gr = \begin{pmatrix} gr_u \\ gr_v \end{pmatrix} = \frac{1}{Lo} \cdot \sum gr(L_m, L_n) \cdot (L_m, L_n) \in Lo \quad (18)$$

Aimed at attaining scale as well as rotation invariance, the sampling pattern has been again sampled past the rotational angle $\theta = \arctan 2(gr_v, gr_u)$. The binary descriptor b_d 's creation is executed by implementing eqn.

(19) on all points' pairs prevalent in set Sh via the short-range sampling points.

$$b_d = \begin{cases} 1 & G_i(L_n^\theta, \rho_n) > I(L_m^\theta, \rho_m) \\ 0 & otherwise \end{cases}, \quad \forall (L_m^\theta, L_n^\theta) \in Sh \quad (19)$$

Edge: Aimed at edge feature, the Canny edge's detection technique is employed. It comprises '5' stages: Smoothing, Finding the gradients, Non-maximal suppression, Thresholding, and then Edge tracking via hysteresis.

The smoothing stage eliminates the noise prevalent in the original image; the Gaussian filter is utilized aimed at this noise removal. After that, the sharpening alters the edge pixels detected by enumerating the image's gradient. The gradient signifies a unit vector that directs in the maximal intensity change's direction.

The gradient's vertical F_u as well as horizontal F_v components are calculated initially; next, the gradient's magnitude and the direction are enumerated; this calculation is executed at the finding gradients stage. The magnitude is enumerated as,

$$\theta = \arctan\left(\frac{|F_u|}{|F_v|}\right) \quad (20)$$

Just the local maxima are signified as edges in the 3rd step. Next, the potential and also the actual edges are specified via thresholding; this is executed at the thresholding phase. In the end step, the edges, which aren't connected to the strong edges, are suppressed.

DWT: The image's decomposition is executed utilizing the wavelet transform. The image has been disintegrating into '2' diverse frequency bands: LH, LL that comprises the horizontal contents and the approximate contents. The wavelet transform's definition is:

$$W_{t(di,tr)} = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{di}} \psi * \left\{ \frac{t-tr}{di} \right\} dt \quad (21)$$

Here, $W_{t(di,tr)}$ signifies the wavelet transform function; di, tr signify the dilation as well as the translation factors; $f(t)$ implies the wavelet transformation function; ψ implies the mother wavelet's dilation.

SIFT: The SIFT technique comprises '4' steps: (a) Scale-Space Extrema's Detection, (b) Key-point's Localization, (c) Orientation Assignment, and then (d) Key-point Descriptor's Generation. Scale-space functions are employed in the detection of the similar object's position as of the diverse dimensions. The scale-space function $r_g(u, v, \beta)$ is articulated as,

$$r_g(u, v, \beta) = Gau(u, v, \beta) * G_i(u, v) \quad (22)$$

Herein, $Gau(u, v, \beta)$ signifies the convolutional Gaussian function. The subsequent stage is the key-point localization that is the procedure of choosing the features section aimed at locating the formerly characterized. Every key point is allotted '1' or else more orientations centred on the local image's gradient directions in the orientation assignment stage. The end process is the Key-point Descriptor's generation that targets the main key-point descriptors' creation.

Geometric Features of Palm and fingers: The palm's and fingers' geometric features are taken out utilizing the Hough Transform function. The Hough transforms detecting a line that is articulated as,

$$d = u \cos\theta + v \sin\theta \quad (23)$$

Here, d signifies the distance as of the origin onto the nearby point on the straight line; θ implies the angle betwixt the u axis and the line linking the origin with the nearby point. The angle's measurement is as of the line to the fingers' rotation alongside the straight direction. Next, the fingers' width is enumerated in '3' diverse positions: in the finger's top, middle and then its bottom. The finger's 1st and end pixel are enumerated to compute the finger's width. The finger's length is partitioned to '3' segments (i.e.) top, middle and then final; after that, the width's measurement is as of the 1st pixel to the end pixel. Next, the palm region's height, width is enumerated by building the bounding box; the palm region's area is computed by utilizing the width and also the height. At last, the features taken out have been equated as,

$$B_f = a_i, \quad i = 1, 2, \dots, n \quad (24)$$

Here, B_f signifies the extracted feature's set; a_i implies the n-number of features.

3.4 Dimensionality Reduction by KLDA

As of the features taken out, necessary features are decremented utilizing the KLDA technique. As of every trait, the features are normalized utilizing Gaussian Kernel's function aimed at the cause of decrementing the errors in the classification stage; the Gaussian Kernel's function is articulated as,

$$\ker(u, v) = \exp\left(-\frac{\|u - v\|^2}{2\xi^2}\right) \quad (25)$$

Here, ξ implies the variable parameter. Past the feature dimension's normalization, the class matrix's mean vector is equated as:

$$\tau_i = \frac{1}{nn_j} \sum_{dd_i \in Dm_i} dd_i \quad (26)$$

Here, τ_i signifies the i^{th} feature's mean; dd_i implies the i^{th} sample; nn_j symbolizes the number of samples prevalent in the j^{th} feature; Dm_i signifies the data matrix. Next, identify all the features' total mean τ . Aimed at every sample of every class, the betwixt-class scatters' matrix bc_s and also the within-class scatters' matrix wc_s is equated as:

$$bc_s = \sum_{i=1}^n (\tau_i - \tau) (\tau_i - \tau)^{Tr} \quad (27)$$

$$wc_s = \sum_{i=1}^n \sum_{j=1}^{n_i} (Y_j - \tau_i) (Y_j - \tau_i)^{Tr} \quad (28)$$

Here, n signifies the number of training samples prevalent in the class i ; τ_i implies the mean vector of samples originating as of class i ; Y_j signifies that class's j^{th} data. wc_s implies the features' scatter about every class's mean; bc_s signifies the features' scatter about the overall mean aimed at every class.

3.5 Feature Fusion

Next, the Feature-level Fusion (FF), together with the Score-Level (SL) fusion is executed. Amidst all fusion techniques, this feature level technique aids in attaining maximal accuracy in-person identification. In FF , unwanted information can be prevalent in the features. The SL is computed aimed at decrementing the information as a single quantity. FF is identified via the easy concatenation of the feature's sets acquired as of the diverse traits. The concatenation procedure is equated as,

$$FF = \{a_i(G_1), a_i(G_2), a_i(G_3), a_i(G_4), a_i(G_5), a_i(G_6)\} \quad (29)$$

Here, $a_i(G_1), a_i(G_2), a_i(G_3), a_i(G_4), a_i(G_5), a_i(G_6)$ implies the facial, ear, retina, finger-print, fingers and also the palm features. After that, the SL computation is centred on score normalization that is vital to change the various systems' scores to a general domain prior to compiling them is provided as,

$$SL = \frac{S(a_i) - \min(S(a_i))}{\max(S(a_i)) - \min(S(a_i))} \quad (30)$$

Here, $S(a_i)$ signifies the features' original score values. Lastly, the end feature is signified as aa_i .

3.6 Rule Generation

The rules are created to test the template image past the feature reduction. The rule generation's combination is: i) every inputted trait is real signifying that the person is real, ii) every inputted trait is fake signifying that the person is the imposter, iii) '3' or else more than '3' inputted traits are fake signifying that the person is the imposter, and also iv) '1' or else '2' inputted traits are fake signifying that the individual is real.

3.7 Identification by using MDLNN

In this identification stage, the features extracted had been inputted into the MDLNN technique that finds the person regarding the rules generated. The technique comprises '3' layers. The 1st layer is the inputted layer (IL) and the final layer is the outputted layer (OL). Betwixt the IL and OL, extra layers of units may exist, termed Hidden Layers (HLs). n - number of HLs is pondered in this methodology. Hence, the NN comprises the accuracy issue owing to the weight updation process such that this study technique utilizes the Satin Bowerbird Optimization (SBO) technique aimed at HL's weight value selection. At first, the selected features' outputs are fed to the IL. The data as of the IL is inputted to the HL; in the HL, the hidden unit is enumerated aimed at the inputted features utilizing the eqn. (31):

$$hid_i = b_s + \sum_{i=1}^n aa_i.l_i \quad (31)$$

Here, b_s signifies the bias value; l_i implies the weight value; aa_i signifies the inputted features. Therefore, the HL's output is inputted to the OL. In the OL, the activation function is equated as,

$$ott_i = b_s + \sum_{i=1}^n hid_i.l_i \quad (32)$$

Here, ott_i signifies the outputted unit. Lastly, the loss function is enumerated utilizing the eqn. (33) as,

$$loss = (tar - ott_i) \quad (33)$$

Here, $loss$ signifies the loss function; ott_i implies the outputted unit; tar implies the network's targeted output. Past the loss computation, examine if the loss obtained matches with the particular threshold value; if it doesn't match then the weight value is optimized utilizing the SBO technique, otherwise the output is signified as the finalized output. Figure 2 exhibits the DLNN's pseudo-code,

Input: Outcome of Fused features aa_i
Output: Genuine (or) Imposter

Begin
Initialize neurons, input layer, hidden layer, output layer, weight value l_i , and loss threshold ll_i .
Calculate the hidden unit and output unit by,

$$hid_i = b_z + \sum_{i=1}^n aa_i \cdot l_i \text{ and } ott_i = b_z + \sum_{i=1}^n hid_i \cdot l_i$$

Check loss function
if ($loss \geq ll_i$) {
 Select the weight value of neurons by SBO
 // **Weight updation by SBO**
 Generate bowers
 Evaluate fitness function
 while the criteria is not satisfied **do**
 Calculate the probability of bowers by, $pb_i = \frac{fit_i}{\sum_{n=1}^N fit_n}$
 if ($co_i \geq 0$) {
 Calculate fitness function by, $\frac{1}{1 + co_i}$
 } **else** {
 Calculate fitness function by, $1 + co_i$
 }
 end if
 Change new position in each iteration
 Calculate fitness function
 Update elite if a bower becomes fitter than the elite
 end while
} **else** {
 Denote the output is the final output
}
end if
End

Figure 2: Pseudocode of MDLNN algorithm

The SBO technique comprises '5' stages: (a) random bowers generation, (b) probability calculation, (c) elitism, (d) position changes, and then (e) mutation.

(a) Random bowers generation: Initially, the set of bowers are created. The 1st population involves a sequence of positions aimed at bowers. Every position is determined as an n-dimensional vector of parameters. These values are initialized arbitrarily such that a uniform distribution is pondered betwixt the lower as well as upper limit parameters. The bower's attractiveness is specified by the compilation of parameters.

(b) Probability calculation: Past initialization, aimed at every population members, the probability is computed. The male, together with the female satin bowerbird, choose the bower centred on the probability calculation. The probability function is enumerated as,

$$pb_i = \frac{fit_i}{\sum_{n=1}^N fit_n} \quad (34)$$

Here, pb_i signifies the probability function; N implies the total number of bowers; fit_i signifies the fitness function that is equated as,

$$fit_i = \begin{cases} \frac{1}{1 + co_i} & co_i \geq 0 \\ 1 + co_i & co_i < 0 \end{cases} \quad (35)$$

Here, co_i signifies the cost function's value in the i^{th} position or else i^{th} bower. The cost function is the function optimized by Eq. (35) that comprises '2' parts. The 1st part computes the final fitness in which values have been greater analogized to or equivalent to '0'; whilst the 2nd part computes the fitness aimed at values lesser than '0'. This eqn. comprises '2' main features.

(c) **Elitism:** Elitism permits the finest solution (solutions) to be conserved at each phase of the optimization procedure. The position of the best bower constructed by birds is proffered as the elite of iteration. The best individual in every iteration is conserved as the elite of iteration. Elites comprise the maximal fitness values and also it can affect other positions.

(d) **Position changes:** In every iteration, any bower's novel alterations can be enumerated as,

$$h_{ie}^{new} = h_{ie}^{old} + v_e \left(\left(\frac{h_{je} + h_{elite,e}}{2} \right) - h_{ie}^{old} \right) \quad (36)$$

Here, h_i signifies the i^{th} bower or else solution vector; h_{ie} implies this vector's e^{th} member; h_{ie}^{old} signifies the bower's old position; h_{ie}^{new} signifies the bower's new position; $h_{elite,e}$ implies the elite's position; h_{je} symbolizes the target solution amidst all the solutions prevalent in the present iteration; the parameter v_e signifies the attraction power prevalent in the goal bower; it specifies the amount of step that is computed aimed at every variable. The v_e is equated as,

$$fit_k = \frac{v}{1 + pb_j} \quad (37)$$

Here, v implies the maximal step size; pb_j implies the probability attained by eqn. (34) utilizing the goal bower.

(e) **Mutation:** At every cycle's end, the random alterations are implemented with definite probability to prevent the male as of attacks, i.e., whilst the males have been busy constructing a bower upon the ground, other animals can attack them. The distribution and mutation procedure are articulated as,

$$h_{ie}^{new} \sim N_d(h_{ie}^{old}, \gamma^2) \quad (38)$$

$$N_d(h_{ie}^{old}, \gamma^2) = h_{ie}^{old} + (\gamma * N_d(0,1)) \quad (39)$$

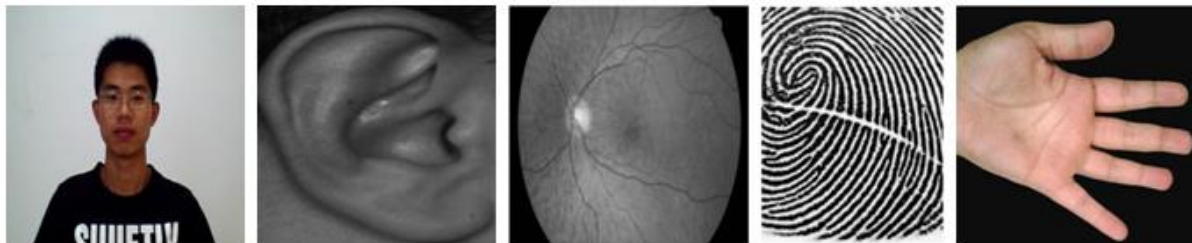
Herein, N_d signifies the normal distribution; the γ implies the proportion of the space width and it is articulated as,

$$\gamma = d(\varepsilon) * (h(\text{var})_{\max} - h(\text{var})_{\min}) \quad (40)$$

Herein, $h(\text{var})_{\max}$ and $h(\text{var})_{\min}$ imply the upper and lower bound allotted to variables; the parameter $d(\varepsilon)$ implies the percentage difference betwixt the upper and lower bounds that is variable. All specified steps are continued till the fit_i is met. At last, the classifier categorised that the person is an imposter or else real centred on the created rule that is specified in the 3.5 section.

4. RESULT AND DISCUSSION

The proposed multi-biometric model's performance is examined. The proposed work is applied in MATLAB. The synthetic dataset is utilized by this work for the performance analysis. The dataset's sample images along with the further process of the image are displayed in Figure 3,



(a)

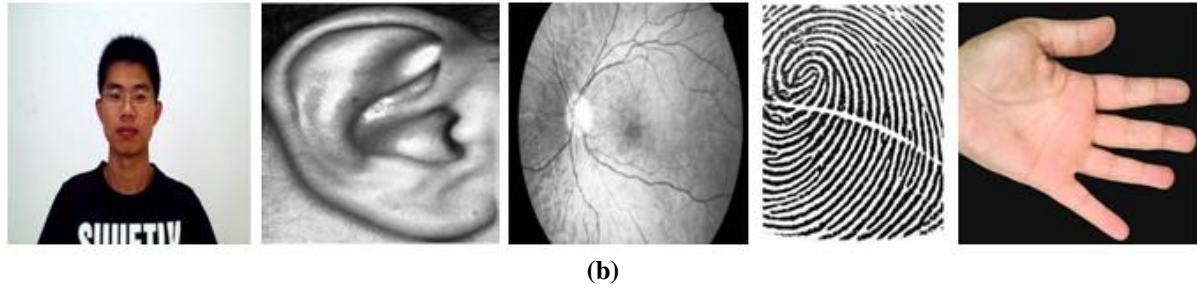


Figure 3: Sample images of all traits, (a) input image, and (b) enhanced image

The sample images of every trait, namely the face, ear, retina, fingerprint, along with front hand image are demonstrated in figure 3. Figure 3 (a) displays the dataset's input image and the enhanced image by utilizing the IPHE algorithm is displayed in figure 3 (b).

4.1 Performance analysis

The proposed MDLNN's performance is examined with the existent DL Neural Network (DLNN), Convolutional Deep Neural Networks (CDNN), along with Artificial Neural Networks (ANN) centered on sensitivity, specificity, accuracy, precision, recall, Negative Predictive Value (NPV), F-Measure, False Positives Rates (FPR), False Negative Rates (FNR), False Rejections Rate (FRR), False Discovery Rates (FDR), along with Matthews Correlations Co-efficient (MCC).

Table 1: Analysis of the performance of the proposed classifier with the existent classifiers based on sensitivity, specificity, and accuracy metrics

Performance Metrics	Proposed MDLNN	DLNN	CDNN	ANN
Sensitivity	0.9111	0.5887	0.2649	0.0056
Specificity	0.9555	0.7943	0.6324	0.8986
Accuracy	0.9407	0.7258	0.5099	0.6009

The MDLNN classifier's performance with the prevailing classifiers, namely DLNN, CDNN, and ANN, concerning sensitivity, specificity, along with accuracy metrics is established in Table 1. The ability to decide the persons rightly is sensitivity, the ability to determine the genuine persons rightly is specificity, and the accuracy metric is differentiating the persons and genuine cases correctly, which is indicated as the recognition rate. Now, the MDLNN algorithm's accuracy is 0.9407, the accuracy of the existent method is 0.7258 for DLNN, 0.5099 for CDNN, and 0.6009 for ANN. The CDNN achieves poor performance analogized to the prevailing methods along with the MDLNN centered upon the accuracy metric. Likewise, the proposed achieve a higher result, i.e. (0.9111) sensitivity, and (0.9555) specificity centered upon the other '2' metrics. The prevailing algorithms attain less performance analogized to the proposed work. It is inferred that the MDLNN centered MMB recognition system attains a better result analogized to the prevailing methods. The graphical depiction of table 1 is demonstrated in Figure 3,

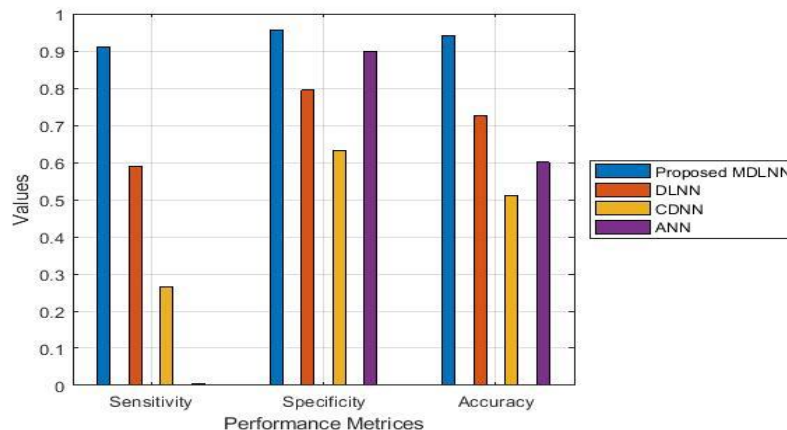


Figure 3: Analyze the performance of the proposed method with the existing methods based on sensitivity, specificity, and accuracy metrics

Table 2: illustrate the performance of the MDLNN classifier with the existing classifiers based on sensitivity, specificity, and accuracy metrics

Performance Metrics	Proposed MDLNN	DLNN	CDNN	ANN
Precision	0.9111	0.5887	0.2649	0.0268
Recall	0.9111	0.5887	0.2649	0.0056
F-Measure	0.9111	0.5887	0.2649	0.0092

Table 2 demonstrated the MDLNN classifier's performance with the DLNN, CDNN, along with ANN classifiers concerning the precision, recall, along with F-Measure metrics. The number of genuine class predictions is measured by the precision metric that belonged to the genuine class, recall measure specifies the total genuine class predictions made out of every genuine example in the dataset together with the amalgamation of both the precision along with recall metrics is the F-measure metric. The precision, recall, F-Measure value of the MDLNN classifier is 0.9111, and the existent methods have 0.5887 for DLNN, 0.2649 for CDNN, and the ANN has (0.0268) precision, (0.0056) recall, and (0.0092) F-Measure in this table 2. The existing ANN algorithm attains worst performance analogized to the existent techniques and also the proposed methods as concluded by this table. The existent DLNN algorithm is better than the CDNN and ANN but, it also gives lower performance than the proposed MDLNN. Therefore, it indicates that a better performance is achieved by means of the proposed MDLNN when contrasted to the prevailing research methods. The pictorial depiction of table 2 is exhibited in Figure 4,

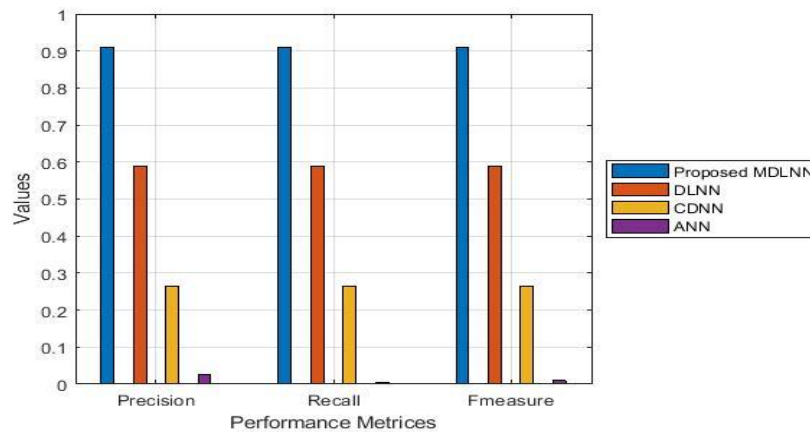


Figure 4: Comparative analysis of MDLNN with the other classifiers based on precision, recall, and F-Measure metrics

Table 3: Performance analysis based on NPV, FPR, and FNR

Performance Metrics	Proposed MDLNN	DLNN	CDNN	ANN
NPV	0.9555	0.7943	0.6324	0.6437
FPR	0.0444	0.2056	0.3675	0.1014
FNR	0.0888	0.4113	0.7350	0.9944

The performance metrics, NPV, FPR, and FNR centered analysis are done for the MDLNN with the prevailing classifiers in table 3. Now the first place is held by the MDLNN classifiers centered on NPV metric and hold the last place centered on FPR and FNR metric. The last place is held by the CDNN algorithm centered upon the NPV metric. The ANN holds first place and CDNN holds first place centered on the FPR metric. The NPV, FPR, and FNR value of the MDLNN is 0.9555, 0.0444, and 0.0888. The NPVE value of the CDNN is 0.6324. Overall, the proposed classifier has higher NPV as of the analysis and low FPR along with FNR result, therefore, it summarized that higher performance is acquired by the proposed MDLNN centered MMB recognition system analogized to the other method-centered recognition. The pictorial demonstration of table '3' is displayed in Figure 5,

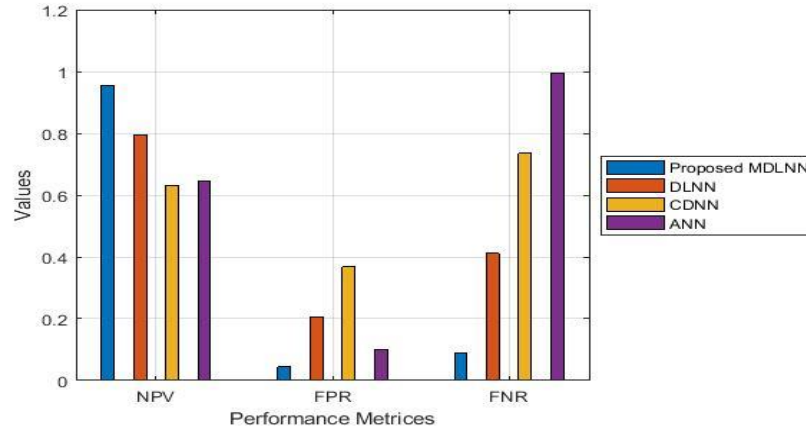


Figure 5: Performance analysis based on NPV, FPR, and FNR metrics

Table 4: Performance analysis based on NPV, FPR, and FNR

Performance Metrics	Proposed MDLNN	DLNN	CDNN	ANN
MCC	0.8667	0.3830	0.1025	0.1776
FRR	0.0888	0.4113	0.7350	0.9944
FDR	0.0888	0.4113	0.7350	0.9731

The proposed MDLNN classifier's performance with the existing DLNN, CDNN, and ANN algorithm concerning MCC, FRR, and FDR metrics is examined in Table 4. The true classes with the predicted classes are measured by the MCC, the FRR specifies the percentage of identification instances in which authorized persons are wrongly rejected, and the total false discoveries in recognition divided by means of the total discoveries in that recognition is FDR. Here, a higher MCC value i.e. 0.8667 is possessed by means of the proposed MDLNN algorithm along with lower FRR and FDR values. The CDNN and ANN possess lower MCC values and higher FRR and FDR values. The DLNN is much better as contrasted with the CDNN and ANN, but the DLNN also has low performance than the MDLNN. Therefore, better results are attained by the MDLNN in multi-biometric recognition. The graphical depiction of table 4 is exhibited in Figure 6,

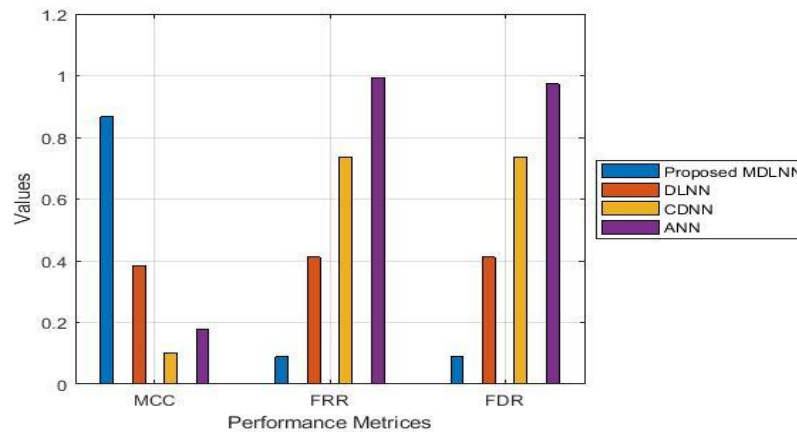


Figure 6: analyze the performance based on MCC, FRR, and FDR metrics

5. CONCLUSION

BR is stated as a budding technique and is attaining higher performance in recent years. This study method proffered an MMB recognition system centred on the FLSL fusion technique and also an MDLNN classifier. This methodology utilizes the face, iris, ear, fingerprint, and also front hand as the inputted traits. Herein, the MDLNN classifier is utilized to make the recognition system become more robust and also yields greater accuracy (also signified as the recognition rate). The synthetic dataset is acquired to examine the proposed multi-model BR system's performance, which is analogized with the existent DLNN, CDNN, and ANN techniques

regarding the accuracy, precision, recall, F-Measure, sensitivity, specificity, NPV, FPV, FNR, MCC, FRR, and also FDR. In this examination, the MDLNN proposed yields an efficient outcome analogized to the other classifiers centred on all performance metrics. The proposed MDLNN accuracy is (0.9407) that is greater analogized to every other classifier. Therefore, this research technique validated that the MDLNN proposed, segmentation and also enhancement centred multi-modal BR attains efficient outcomes analogized to the other techniques. This work can be improved in the upcoming future by implementing the latest technique to increment the performance and make the system more robust.

REFERENCES

1. Kumari P, "A fast feature selection technique in multi modal biometrics using cloud framework" , Microprocessors and Microsystems, vol. 79, pp. 103277, 2020, [10.1016/j.micpro.2020.103277](https://doi.org/10.1016/j.micpro.2020.103277).
2. Sree Vidya B, and Chandra E., "Entropy based Local Binary Pattern (ELBP) feature extraction technique of multimodal biometrics as defence mechanism for cloud storage" , Alexandria Engineering Journal, vol. 58, no. 1, pp. 103-114, 2019.
3. Basma Ammour, Larbi Boubchir, Toufik Bouden, and Messaoud Ramdani, "Face-iris multimodal biometric identification system" , Electronics vol. 9, no. 1, pp. 85, 2020.
4. Ibrahim Omara, Ahmed Hagag, Souleyman Chaib, Guangzhi Ma, Fathi E. Abd El-Samie, and Enmin Song, "A hybrid approach combining learning distance metric and dag support vector machine for multimodal biometric system" , IEEE Access 2020, [10.1016/j.micpro.2020.103277](https://doi.org/10.1016/j.micpro.2020.103277).
5. Milind E Rane, and Deshpande Prameya P, "Multimodal biometric recognition system using feature level fusion" , In International Conference on Computing Communication Control and Automation (ICCUBE), IEEE, pp. 1-5, 2018, [10.1109/ICCUBE.2018.8697821](https://doi.org/10.1109/ICCUBE.2018.8697821).
6. Haider Mehraj, and Ajaz Hussain Mir, "Feature vector extraction and optimisation for multimodal biometrics employing face, ear and gait utilising artificial neural networks" , International Journal of Cloud Computing, vol. 9, no. 2-3, pp. 131-149, 2020.
7. Gunasekaran Raja J K., and R. Pitchai, "Prognostic evaluation of multimodal biometric traits recognition based human face, finger print and iris images using ensembled SVM classifier" , Cluster Computing , vol. 22, no. 1, pp. 215-228, 2019.
8. Mohsen AM El-Bendary, Hany Kasban, Ayman Haggag, and M. A. R. El-Tokhy, "Investigating of nodes and personal authentications utilizing multimodal biometrics for medical application of WBANs security" , Multimedia Tools and Applications, vol. 79, no. 33, pp. 24507-24535, 2020.
9. Shaveta Dargan, and Munish Kumar, "A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities" , Expert Systems with Applications vol. 143, pp. 113114, 2020.
10. Swati K Choudhary, and Ameya K. Naik, "Multimodal Biometric Authentication with Secured Templates—A Review", In International Conference on Trends in Electronics and Informatics (ICOEI), IEEE, pp. 1062-1069, 2019, [10.1109/ICOEI.2019.8862563](https://doi.org/10.1109/ICOEI.2019.8862563).
11. Aman Kathed, Sami Azam, Bharanidharan Shanmugam, Asif Karim, Kheng Cher Yeo, Friso De Boer and Mirjam Jonkman, "An enhanced 3-tier multimodal biometric authentication", In International Conference on Computer Communication and Informatics (ICCCI), IEEE, pp. 1-6, 2019, [10.1109/ICCCI.2019.8822117](https://doi.org/10.1109/ICCCI.2019.8822117).
12. Rachid Chlaoua, Abdallah Meraoumia, Kamal Eddine Aiadi and Maarouf Korichi, "Deep learning for finger-knuckle-print identification system based on PCANet and SVM classifier", Evolving Systems, vol. 10, no. 2, pp. 261-272, 2019.
13. Karthiga R and Mangai S, "Feature selection using multi-objective modified genetic algorithm in multimodal biometric system", Journal of Medical Systems, vol. 43, no. 7, pp. 1-11, 2019.
14. YanTong, Frederick W. Wheeler and Xiaoming Liu, "Improving biometric identification through quality-based face and fingerprint biometric fusion", In IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 53-60, 2010, [10.1109/CVPRW.2010.5543233](https://doi.org/10.1109/CVPRW.2010.5543233).
15. Annamalai Prakash R. Krishnaveni and Ranganayakulu Dhanalakshmi, "Continuous user authentication using multimodal biometric traits with optimal feature level fusion", International Journal of Biomedical Engineering and Technology, vol. 34, no. 1, pp. 1-19, 2020.
16. El mehdi Cherrat, Rachid Alaoui and Hassane Bouzahir, "A multimodal biometric identification system based on cascade advanced of fingerprint finger vein and face images", Indonesian Journal of Electrical Engineering and Computer Science, vol. 18, no. 1, pp. 1562-1570, 2020.
17. Gayatri U Bokade and Rajendra D. Kanphade, "Secure Multimodal Biometric Authentication Using Face, Palmprint and Ear a Feature Level Fusion Approach", In International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, pp. 1-5, 2019, [10.1109/ICCCNT45670.2019.8944755](https://doi.org/10.1109/ICCCNT45670.2019.8944755).

18. Mohammad Haghighat, Mohamed Abdel-Mottaleb and Wadee Alhalabi, "Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition", *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 9, pp. 1984-1996, 2016.
19. Karthiga Rand MangaiS, "Feature selection using multi-objective modified genetic algorithm in multimodal biometric system", *Journal of Medical Systems*, vol. 43, no. 7 pp. 1-11, 2019.
20. Nada Alay and Heyam H. Al-Baity, "deep learning approach for multimodal biometric recognition system based on fusion of iris, face, and finger vein traits", *Sensors*, vol. 20, no. 19, pp. 5523, 2020.
21. Meryem Regouid, Mohamed Touahria, Mohamed Benouis and Nicholas Costen, "Multimodal biometric system for ECG, ear and iris recognition based on local descriptors", *Multimedia Tools and Applications*, vol. 78, no. 16, pp. 22509-22535, 2019.
22. Gurjit Singh Walia, Tarandeep Singh, Kuldeep Singh and Neelam Verma, "Robust multimodal biometric system based on optimal score level fusion model", *Expert Systems with Applications*, vol. 116, pp. 364-37, 2019, 10.1016/j.eswa.2018.08.036.
23. Prabu S, LakshmananM and V. Noor Mohammed, "A multimodal authentication for biometric recognition system using intelligent hybrid fusion techniques", *Journal of Medical Systems*, vol. 43, no. 8, pp. 1-9, 2019.
24. Gaurav Jaswal, Amit Kaul and Ravinder Nath, "Multimodal biometric authentication system using hand shape palm print and hand geometry", In *Computational Intelligence Theories Applications and Future Directions-Volume II*, Springer, Singapore, pp. 557-570, 2019, 10.1007/978-981-13-1135-2_42.
25. Gunasekaran K, RajaJ and R. Pitchai, "Deep multimodal biometric recognition using contourlet derivative weighted rank fusion with human face fingerprint and iris images", *Automatika Časopis Za Automatiku Mjerenje Elektroniku Računarstvo I Komunikacije*, vol. 60, no. 3, pp. 253-265, 2019.

Novel Algorithm For Optimal PMU Placement For Wide Ranging Power System Observability

Nitish Arora
Department of Electrical Engineering
Delhi Technological University
New Delhi, India
nitisharora_2k19psy14@dtu.ac.in

S.T.Nagarajan
Department of Electrical Engineering
Delhi Technological University
New Delhi, India
stnagarajan@dce.ac.in

Abstract— The paper envisages a novel algorithm for optimal location for placement of phasor measurement unit (PMU) ensuring complete power network observability. Based upon the network connectivity information, a two stage method has been proposed to keep the number of PMU's minimum in the network. In stage1, the PMU's are added starting with the highest valency bus whereas in stage 2 they are eliminated from less important buses while ensuring complete system observability. The proposed novel algorithm has been tested on various test systems viz. IEEE-14 bus, 30-bus, 57-bus and New England 39-bus system. On comparison with already known techniques available in literature, it was found that the algorithm designed in the present study is simple to implement and better in requisites of accuracy and computational speed.

Keywords—PMU (Phasor Measurement Unit), Observability, Computational Time

I. INTRODUCTION

Introduced in early 1990's, PMU's have wide applications in power system in recent times [1]. State estimation, power system protection, monitoring and control are some of the thrust areas where it is widely used. State estimation, an essential part of Energy Management System, requires the power system to be fully observable from the available data [2]. Conventional Estimators used bus voltage, real and reactive power flows, power injection data obtained from SCADA (Supervisory Control and Data Acquisition) to estimate the real time state of the power system. However, PMU provides time synchronized voltage phasors measurement of the bus connected and the corresponding current phasor of all the branches which are incident to that bus [3]. This feature makes PMU measurements superior to that of SCADA measurements and hence by placing PMU's at all the buses of the network, direct measurement of the state of system can be done without the necessity of state estimators.

The high cost of the PMU and its communication system, does not make it economically viable to install it at every bus. Hence, there is a necessity of optimal PMU location in power system for complete observability for state estimation. Various methods, mathematical and heuristic, have been proposed to address the issue of optimum PMU placement. Bisecting search and simulated annealing based method [4], integer programming (IP) based approach, genetic algorithm, Tabu search [5-8] are some of the contributions in this field. Binary Particle Swarm Optimization (BPSO), which takes care of maximum measurement redundancy and minimum number of PMU's has been proposed by Ahmadi et al [9]. Heuristics based placement method [10], addresses the issue of zero injection bus as pseudo measurements.

A topological observability based, two stage, optimal PMU placement novel algorithm is proposed in this paper. Initially the PMU's are added in the power system starting with the bus having maximum number of branches connected to it (bus with maximum valency). This stage is referred to as addition stage. This addition of PMU's is continued until the constraint vector function associated with the optimization problem is satisfied. After the addition stage, PMU elimination begins starting from the Radial Bus i.e. the bus connected to only one other bus in the given power system. Maximization of number of observable buses with minimum number of PMU's at the strategic locations, reduction of computation speed and enhancing the accuracy are the crucial contributions of this paper.

This paper is structured as follows: Section 2 describes rules as well as observability analysis of power system. Section 3 illustrates in detail the problem formulation and Section 4 gives the insight about the methodology for problem solution. The results are discussed in Section 5. The overview of the entire paper is concluded in Section 6.

II. POWER SYSTEM OBSERVABILITY ANALYSIS AND RULES

For state estimation of the power system, total network observability is essential from the available measurements. A bus is observable if the voltage and current flows are known. Hence for complete power system observability all the buses must be fully observable either by direct or indirect measurements. Observability can be evaluated by either numerical or topological techniques. For a system to be observable numerically, design matrix H is required to be of full rank [11] whereas a topologically observable system requires at least one spanning measurement tree to be of full rank [12]. Certain rules are followed for estimation of observability of each bus which are:

Rule 1: PMU installation at a bus provides direct measurement of the voltage phasor of the connected bus and the current phasors of the incident branches.

Rule 2: In case of two buses connected by a branch, voltage phasors of any one bus can be calculated. Provided, voltage phasor of any one bus and current phasor of the branch incident to a bus is available.

Rule 3: If measurement of voltage phasors of two buses connected by a branch is known then the current phasors of the incident branch can be calculated.

Complete system observability, using the rules 1, 2 and 3 can be illustrated by IEEE 9-bus system as shown in Fig1.

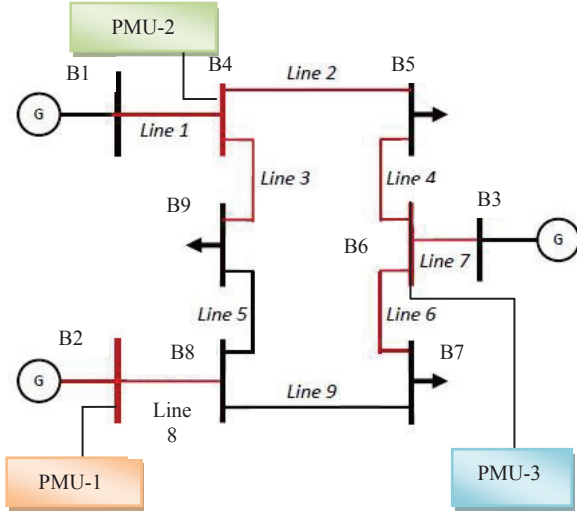


Fig. 1. IEEE 9-bus system

PMU placement at B2 provides the measurements of voltage phasor at B2 and current phasor corresponding to line 8 by Rule 1. Hence B2 is observable directly. By applying Rule 2 voltage phasor of B8 can be measured hence making it indirectly observable. Placing a PMU at B4, incident B1, B5 and B9 will also be observable. Current phasor through line 5 can also be calculated by Rule 3. Similarly by placing a PMU at B6, indirectly observable buses are B3, B5 and B7. Line current through B8 can also be calculated applying Rule 3. It can be observed that by placing 3 PMU's the entire power system becomes entirely observable. However identifying PMU location (in this case B2, B4 and B6) is an optimization problem, which is discussed in next section.

III. PROBLEM FORMULATION

Optimization problem for PMU in N bus system is formulated as:

$$\begin{aligned} & \text{minimize } \sum_{i=1}^N w_i x_i \\ & \text{such that } G(X) \geq b \end{aligned} \quad (1)$$

here 'X' denotes the binary decision variable vector and G(X) represents the constraint function. The present study, for deciding the optimum locations for placement of PMU, can be formulated as zero one problem. The presence or absence of the PMU at a particular location can be denoted by 1 or 0 respectively. The entries of 'X' are as:

$$x_i = \begin{cases} 1, & \text{if the PMU is placed at the } i^{\text{th}} \text{ bus} \\ 0, & \text{otherwise} \end{cases} \quad i=1, \dots, N \quad (2)$$

$b=[111\dots]^T$ depicts a unit vector of size N. Considering the installation cost of all the PMU's installed in the power network same and the value is taken as 1 per unit, then the optimization problem formulation can be done as:

$$\begin{aligned} & \text{Minimize } \sum_{i=1}^N x_i \\ & \text{such that } G(X) \geq 1 \end{aligned} \quad (3)$$

Full network observability is ensured by constraint vector function. A minimum set of x_i is found that satisfies (3). Since PMU installation at a bus can provide current phasor of the branches incident to that bus apart from the voltage phasor, hence voltage phasors of the nearby branches can be computed. Hence the matrix [A], representing bus connectivity information, can be obtained having elements as below:

$$A_{ij} = \begin{cases} 1, & \text{if } i=j \text{ or if bus } i \text{ and } j \text{ are connected} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

The product of matrix [A] and [X] results in constraint function G(X)

$$G(X) = AX \geq 1 \quad (5)$$

$$f_i = a_{i,1}x_1 + \dots + a_{i,i}x_i + \dots + a_{i,N}x_N \quad (6)$$

f_i , the object function, is observable if any x_i that appears in f_i is non zero. The system becomes completely observable if all the f_i appearing in F are non zero.

IV. PROPOSED ALGORITHM FOR PLACEMENT OF PMU

The present study aims to make the complete power system observable by installing least number of PMU's. The placement of the PMU is identified in two stages. The first stage involves placing the PMU's by identifying the bus having maximum number of branches connected to it i.e. maximum valency. This iterative process continues till the constraint vector function (5) is satisfied. To improve the formulation computationally, the number of constraints and variables can be reduced. All the columns of bus connectivity matrix A can be dropped corresponding to $x_i=0$ in (5). Moreover, rows and columns corresponding to variables which have been set to 1 can be dropped. Assume that x_i has been set to '1', now if $a_{ji}=0$ then the value of $x_i=1$ has no significance as $a_{ji}x_i=0$ whether $x_i=1$ or 0.

In the elimination stage, PMU's are removed starting from the radial buses as PMU installed at the radial bus will make two buses observable. Whereas installation of the PMU at a bus connected to a radial bus makes over two buses observable. The iterative process is carried out up to maximum valency bus. This elimination process continues till the elimination of PMU from a bus leads to system being unobservable.

The proposed algorithm provides various options for the placement of PMU with same minimum number of PMU's. However, the one with which highest number of branches are observable is chosen. This provides redundant measurements which helps in good state estimation. Redundancy measurement calculation expression is:

$$\text{Redundancy} = \sum_{i=1}^N \text{sum}(f_{i(k)}) \quad (7)$$

Where f_i denotes the object function, k represents the optimal solution obtained and N denotes the total count of buses in the system.

Flow charts for proposed two stage algorithm are detailed in Fig. 1 and 2.

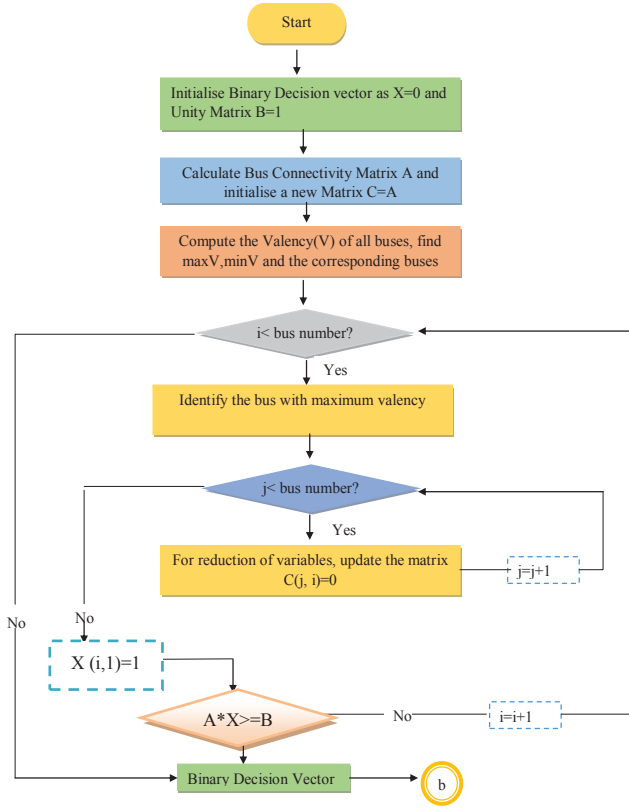


Fig. 2. Flow Chart of Stage 1

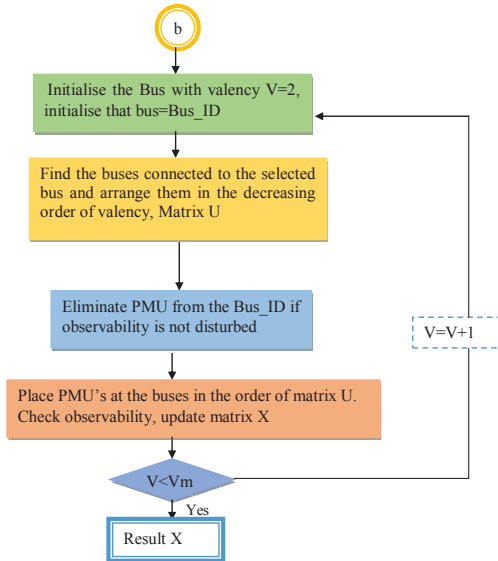


Fig. 3. Flow Chart of Stage 2

V. RESULTS AND DISCUSSION

The effectiveness of the algorithm has been tested on various test systems. The optimal number of PMU's and their corresponding bus locations has been depicted in Table 1. A comparison has been done of the results obtained by the proposed algorithm with the methods described in literature [13],[14] and [15]. Fig.4 depicts the validation of the results on comparison with the available techniques in literature. The computational time taken by the proposed algorithm is depicted in Table 2.

TABLE I. OPTIMAL LOCATION AND NUMBERS OF PMU FOR VARIOUS TEST CASES

Test System	Optimal location of PMU	Optimal numbers
IEEE -14 bus	B2, B6, B7, B9	4
IEEE- 30 bus	B3, B5, B6, B9, B10, B12, B18, B24, B25, B27	10
New England- 39 bus	B2, B6, B9, B10, B13, B14, B17, B19, B20, B22, B23, B25, B29	13
IEEE-57 bus	B1, B6, B9, B15, B19, B22, B25, B27, B29, B32, B36, B38, B39, B41, B47, B50, B53	17

The proposed algorithm also provides various options for PMU placement with same number of PMU's but different number of observable branches. The option with more number of observable branches will provide redundant measurements, resulting in good state estimation. However, with more number of observable branches the requirements for Current Transformers increase, thereby increasing the cost. Hence the proposed algorithm also provides the option for PMU locations when less observable branches are required resulting in less requirement for C.T's. Fig. 5 and Fig.6 illustrate the options and the number of observable branches in case of IEEE 30-bus system and New England 39-bus system.

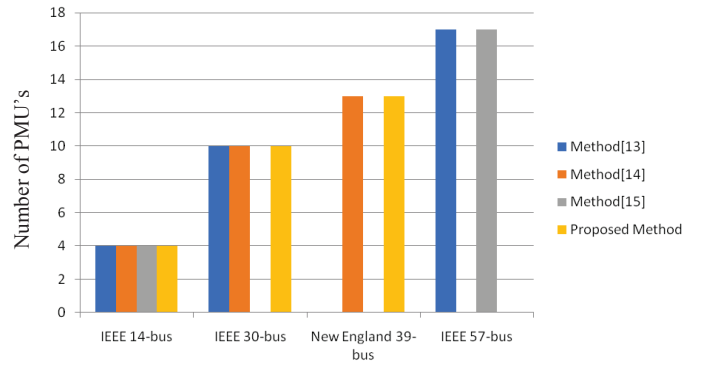


Fig. 4. Comparison of results with available techniques in literature.

TABLE II. COMPUTATIONAL TIME (IN SECONDS)

Test System	Computational Time(in sec)
IEEE 14-bus	0.034308
IEEE 30-bus	0.047533
New England 39-bus	0.076281
IEEE 57-bus	0.22635

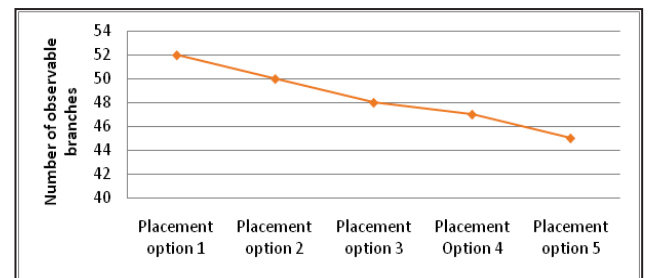


Fig. 5. Various options of PMU placement and number of observable branches for IEEE 30-bus system

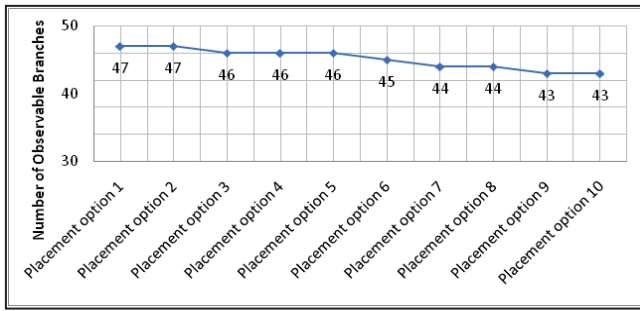


Fig. 6. PMU placement options and number of observable branches for New England 39-bus system

VI. CONCLUSION

The paper presents a novel algorithm to make the complete power system under consideration fully observable by installing minimum count of PMU's, considering their installation cost. The two stage proposed algorithm is fast, uncomplicated and trouble-free to implement. Results obtained for different test systems show the effectiveness of the algorithm for optimal number and location for PMU's for complete system observability and computational efficiency.

REFERENCES


- [1] A. G. Phadke and J. S. Thorp, "Synchronized Phasor Measurements and Their Applications," New York:Springer, 2008.
- [2] Phadke AG., "Synchronized phasor measurements in power systems,"IEEE Trans Comput Appl Power 1993; 6(2):10–5.
- [3] Phadke AG. , "Synchronized sampling and phasor measurements for relaying and control," IEEE Trans Power Deliver 1994; 9(1):442–52.
- [4] Baldwin TL, Mill L, Boisen MB, et al., "Power system observability with minimal phasor measurement placement," IEEE Trans Power Syst 1993; 8(2):707–15.
- [5] Xu B, Abur A, "Observability analysis and measurement placement for systems with PMUs," Proceedings of the IEEE PES power systems conference and exposition 2004;2: p. 943–6.
- [6] Xu B, Abur A, "Optimal placement of phasor measurement units for state estimation", Final project report, PSERC 2005.
- [7] K. S. Cho, J. R. Shin, and S. H. Hyun, "Optimal placement of phasor measurement units with GPS receiver," in IEEE Power Eng. Soc. Winter Meeting, vol. 1, pp. 258–262.
- [8] B. Milosevic and M. Begovic, "Nondominated sorting genetic algorithm for optimal phasor measurement placement," IEEE Trans. Power Syst., , Feb. 2003;vol. 18, no. 1, pp. 69–75.
- [9] Ahmadi A, Beromi YA, Moradi M., "Optimal PMU placement for power system observability using binary particle swarm optimization considering measurement redundancy,"Expert Syst Appl Sci Direct 2011;38(6):7263–9.
- [10] SahaRoy BK, Sinha AK, Pradhan AK., "Optimal phasor measurement unit placement for power system observability – a heuristic approach", Proceedings IEEE Symposium Series on Computational Intelligence; 2011.
- [11] A. Monticelli and F.F. Wu, "Network Observability: Theory", IEEE Transactions on Power Apparatus and Systems, May 1985; Vol.104, No.5, pp1042-1048.
- [12] K.A. Clements, "Observability Methods and Optimal Meter Placement",Electrical Power & Energy Systems, April 1990; Vol. 12, No. 2, pp88-93.
- [13] Xu B, Abur A., "Optimal placement of phasor measurement units for state estimation. Final project report, PSERC 2005.
- [14] Chakrabarti S, Kyriakides E., "Optimal placement of phasor measurement units for power system observability,"IEEE Trans Power Syst 2008;23(3):1433–40.
- [15] Hurtgen M, Maun JC., "Optimal PMU placement using iterated local search," Int J Electr Power Energy Syst 2010;32(8):857–60.

Optical Flow-Based Weighted Magnitude and Direction Histograms for the Detection of Abnormal Visual Events Using Combined Classifier

Gajendra Singh, National Institute of Technology, Jalandhar, India

Rajiv Kapoor, Delhi Technological University, India

Arun Khosla, Dr. B. R. Ambedkar National Institute of Technology, Jalandhar, India

 <https://orcid.org/0000-0001-8571-7614>

ABSTRACT

Movement information of persons is a very vital feature for abnormality detection in crowded scenes. In this paper, a new method for detection of crowd escape event in video surveillance system is proposed. The proposed method detects abnormalities based on crowd motion pattern, considering both crowd motion magnitude and direction. Motion features are described by weighted-oriented histogram of optical flow magnitude (WOHOFM) and weighted-oriented histogram of optical flow direction (WOHOFD), which describes local motion pattern. The proposed method uses semi-supervised learning approach using combined classifier (KNN and K-Means) framework to detect abnormalities in motion pattern. The authors validate the effectiveness of the proposed approach on publicly available UMN, PETS2009, and Avaneue datasets consisting of events like gathering, splitting, and running. The technique reported here has been found to outperform the recent findings reported in the literature.

KEYWORDS

Abnormality Detection, K-Means, KNN, Optical Flow, Semi-Supervised, Surveillance Video, WOHOFM

INTRODUCTION

Security is a major concern for everyone at public places and hence there is an increase in demand of video surveillance systems. These cameras based video surveillance systems generate a huge amount of data but there are limited number of skilled persons to watch and analyze this data. One cannot solely rely upon human observer because a long time may pass before a suspicious event takes place and human attention may not have remained focus on task in such situations which can lead to an event of interest being missed. So to avoid such situations an automated system is needed that can analyze such huge amount of data and trigger alarm in abnormal events.

Abnormal event detection algorithms generally two step process viz. extraction of features and identification of pattern on extracted features. Feature extraction is a process of transforming raw data into different primitives that describes the scene characteristics in more discriminative way. Feature extraction methods are mainly classified into two classes: an object based and pixel based.

DOI: 10.4018/IJCINI.20210701.oa2

This article, published as an Open Access article on April 23rd, 2021 in the gold Open Access journal, the International Journal of Cognitive Informatics and Natural Intelligence (converted to gold Open Access January 1st, 2021), is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Object based methods extract information about arrangement of pixels and trajectory based methods are one of the object based methods (Basharat *et al.*, 2008), where abnormal events are identified by objects trajectories. Trajectories based methods extract trajectories of each object in the scene and make model based on the trajectory statistics. But these type of methods are not able to perform well in crowded scene because there are too many objects in the scene and tracking of each object is difficult due to occlusion.

Pixel based methods describes the scene by low level features. One such method is based on crowd flow which uses information about motion of crowd. In (Wang *et al.*, 2014; Mishra *et al.*, 2020) abnormal event detection is carried out by modelling optical flow, which tracks change in arrangement of pixel and does not require object tracking. Optical flow based algorithms are capable of easily and precisely modeling highly crowded scene. Feature extraction is followed by identification of patterns to differentiate normal and abnormal events. Pattern identification broadly classified into supervised, semi supervised and unsupervised techniques.

In this paper, a new anomaly detection framework is proposed for video surveillance systems using semi- supervised learning process. For feature extraction optical flow is used. Feature vector consists of both magnitude and direction weighted by and energy function, which robustly describe even small change in movement or in direction and the distribution of activity is modelled by combined classifier (KNN and K- Means). Due to robust features and combined classifier our method works well in challenging conditions with respect to other state of the art methods. Rest of the paper is organized as follows, related work provides an overview of related work on abnormality detection in video surveillance systems. Proposed method section gives an overview of proposed method and describe the process of extraction of useful information from the scene by using weighted oriented histogram of optical flow direction (WOHOFD) and weighted oriented histogram of optical flow magnitude (WOHOFM). This section also describes, how to train our model for detection of abnormal events in the scene. The performance results of our proposed method are described in result section. Finally paper is concluded in last conclusion section.

RELATED WORK

First thing in suspicious event detection is to extract features which can robustly describe the scene statistics e.g. low level features and high level features. After extracting features from the scene, event modelling or classification of data is done based on extracted features. In event modelling, algorithm learns the behavior or pattern of extracted features and classify whether scene contains an anomalous event or not. Event modeling is generally known as machine learning and can be classified into three major categories: supervised techniques, semi-supervised techniques and unsupervised techniques.

The anomaly detection based on supervised technique requires the labeling of samples for both normal samples and abnormal samples to train the model and give prediction on test samples. These methods are generally train model for specific abnormal state whose features are previously known or set, such as 'U' turn detection in traffic surveillance scene (Zen & Ricci, 2011; Z *et al.*, 2005) .

Semi- supervised techniques train model only for normal samples and these techniques can be further categorized into two sub categories viz. rule based and model based. In rule-based methods, predefined rules are set in during training phase based on the characteristics of scene or extracted features, if a test sample do not follow predefined rules, would be classified/labelled as abnormal/irregular e.g., sparse coding (Cong, *et al.*, 2011), online dictionary updating (Zhao *et al.*, 2011) and sparse combination learning (Lu *et al.*, 2013) etc. In Sparse coding method (Cong, *et al.*, 2011), a reconstruction cost based method for abnormality detection was proposed, if a sample having larger reconstruction cost will be classified as abnormal sample. In (Boiman & Irani, 2007; Saligrama & Chen, 2012; Hamid *et al.*, 2005; Javan & Levine, 2013) similarity based methods were proposed, where score of a test data is calculated based on how much test sample is similar to the training sample, higher similarity means low abnormality score and vice versa. While on the other hand

model based method attempts to build a model only for normal scene, probability of test samples are calculated with respect to the train model, if probability is low then test sample will be considered as anomalous samples. The commonly used models are Hidden Markov Model (HMM) (Kratz & Nishino, 2009; Zhang et al., 2005; Andrade et al., 2006; Ouivirach et al. 2006) and Markov Random Field (MRF) model (Kim & Grauman, 2009; Benezeth et al., 2009), which are used in a wide variety of applications including anomaly detection. Kim et al. (Kim & Grauman, 2009) proposed a Space Time Markov Random Field (STMRF) approach illustrating distribution of regular motion behavior. (Kratz et al., 2009) proposed a distribution based HMM model using the local spatial and temporal motion behavior. Andrade et al. (Andrade et al., 2006) proposed a Multiple Observation Hidden Markov Model (MOHMM), which groups the video into different cluster using spectral clustering and trained the model for each cluster. (Mehran et al., 2009) proposed Social Force Model (SFM) for detection and localization of abnormal behavior of crowd based on interaction forces. (Adam et al., 2008) presented a method for abnormal behavior detection, which observes histogram of movement using optical flow for normal behavior at several fixed-locations, (Wu et al., 2010) modelled chaotic invariants of Lagrangian particle trajectories for normal event to characterize crowded scene. (Cui et al., 2011) proposed an interaction energy potential function which describes the action by spatial arrangement with the surroundings of normal objects changing over time and if a test sample having abrupt fluctuations in function is categories as abnormal. (Sharif et al., 2012) proposed an entropy based method, which calculate entropy of the spatiotemporal data of the interest points to measure randomness in video frame. (Kwon & Lee, 2014) proposed a method for the detection of abnormal events based on predefined energy function whose parameters reflect frequency, causality, and significance of events. (Gu et al., 2014) proposed a method for detection of abnormal events which uses particle entropy to describe the distribution of objects in crowded scenes. (Wang et al., 2019) proposed a semi supervised method based on deep network for the detection of abnormal events. Deep network is used for the extraction of features and SVDD (Support Vector Data Descriptor) is used for classification. The performance of deep network based methods is very good but these methods require high computational power.

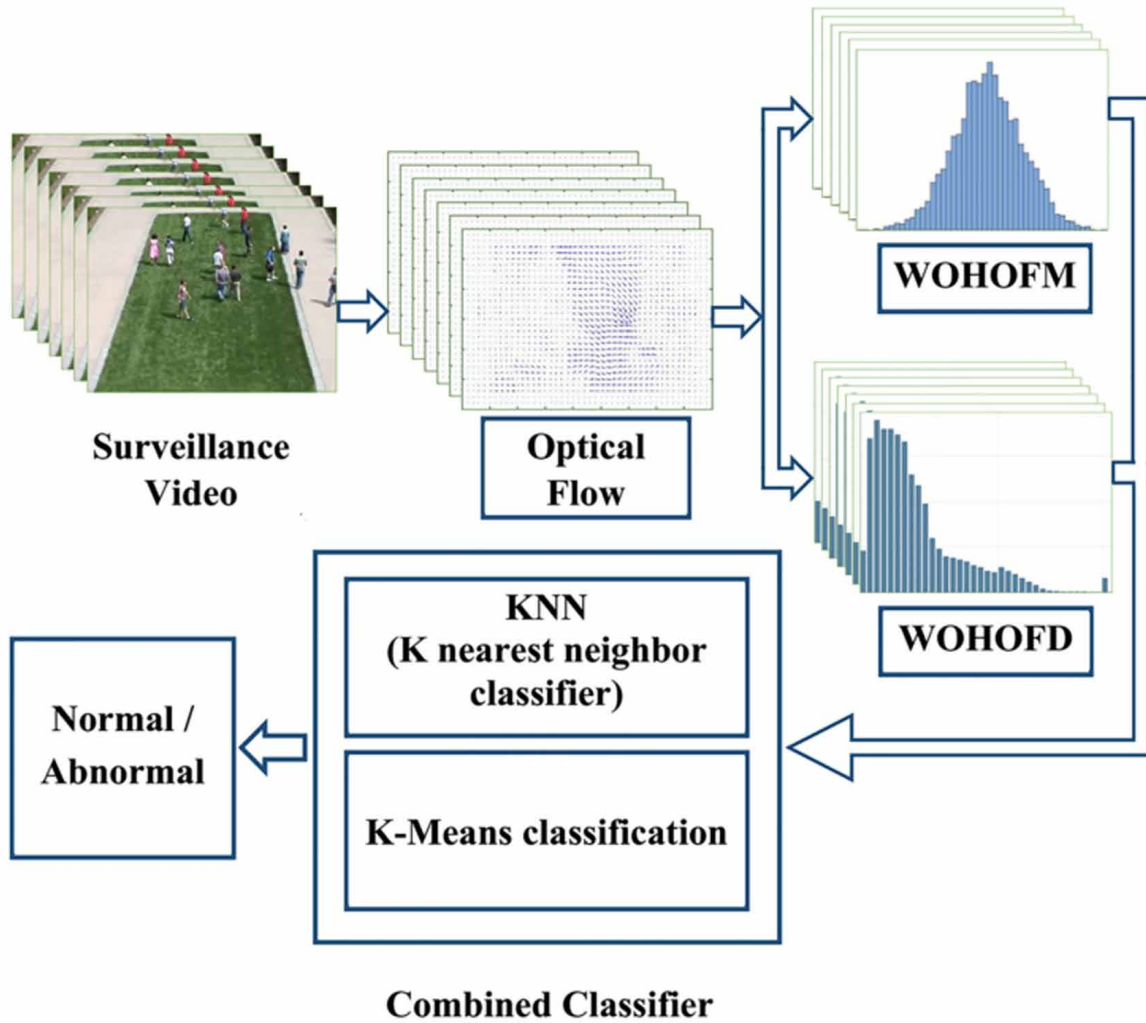
Unsupervised techniques do not require labeling of normal or abnormal data samples in advance and abnormal event detection is carried out based on the statics that abnormalities are rare with respect to normal situation. Many of the techniques are based on clustering of data and detect abnormal events based on the calculation of distance between the test data sample and the nearest cluster center, if the distance is more the predefined threshold than test sample is labelled as abnormal.

PROPOSED METHOD

This section represents our crowd anomaly detection framework. Our proposed method works in two stages: feature extraction and classification. In this method two main features are extracted: optical flow magnitudes and optical flow directions. Optical flow refers to the visible motion of an object in an image, and the apparent flow of pixels with respect to its neighborhood in an image. It is the result of 3-D motion being projected on a 2-D image plane. In normal crowded scenes, people density is high, the movement of individual is constrained by other people's movement and hence movement of individuals are generally slow. The speed and direction of individuals do not change so much within a short period. However, in abnormal situations with respect to normal situations, the speed of movement of crowd is very high and the direction also changes very rapidly due to the fear. Hence, both direction and magnitude of optical flow become significant features to describe the crowd movements. People started running or diverge from the places where they were. When we want to consider the crowd area to detect abnormal situation, the view field of cameras are kept quite large.

Flow chart of proposed algorithm based on weighted magnitude and direction histogram with combined classifier for abnormal event detection is shown in Figure 1.

Figure 1. Flow chart of proposed algorithm

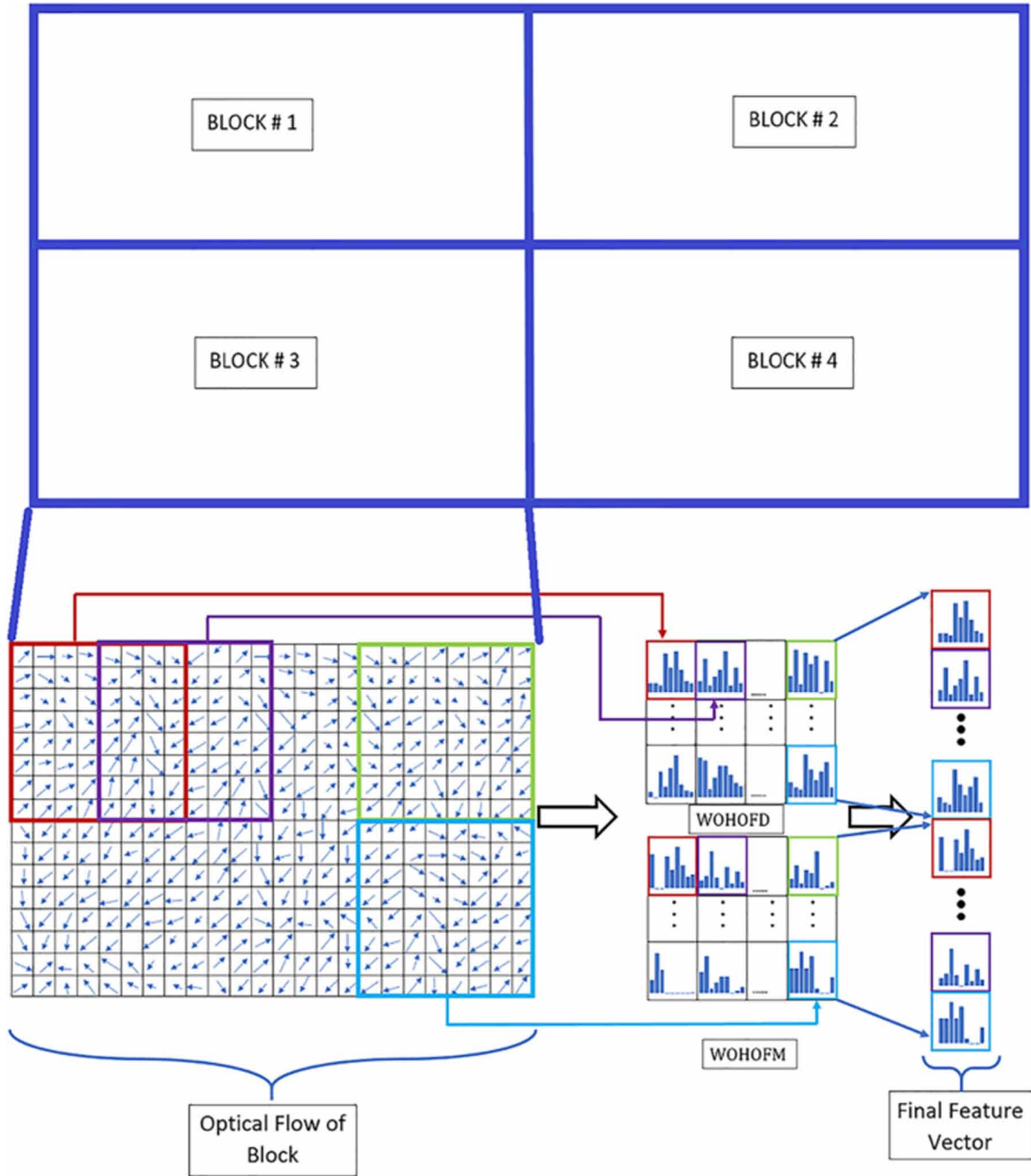


Optical flow is calculated for each frame of video by Lucas-Kanade derivative of Gaussian method proposed by Bruce D. Lucas and Takeo Kanade (Lucas & Kanade, 1981). It assumes that the displacement of patches or objects between consecutive frames is constant or small in a local neighborhood of a point under consideration. Lukas Kanade method divides the image into small sectors and assumes constant velocity in each sector and uses least squares criterion to solve the basic optical flow equations for all the pixels. It computes an estimate of the horizontal and vertical velocity component $[U \ V]^T$ that minimizes the Eq.(1).

$$\sum_{k \in \odot} W^2 \left[E_x u + E_y v + E_t \right]^2 \dots \quad (1)$$

Where W is a window function that emphasizes the constraints at the center of each section and E_x , E_y and E_t are derivatives of image brightness in spatial and temporal dimensions and u is horizontal optical flow and v is vertical optical flow. Solution to the minimization problem is given in Eq.(2)

Figure 2. Extraction of Feature Vector



$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_i W^2 E_{xi}^2 & \sum_i W^2 E_{xi} E_{yi} \\ \sum_i W^2 E_{xi} E_{yi} & \sum_i W^2 E_{yi}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i W^2 E_{xi} E_{ti} \\ -\sum_i W^2 E_{yi} E_{ti} \end{bmatrix} \dots \quad (2)$$

Magnitude and direction of optical flow for each pixel is calculated according to Eq.(3) and Eq.(4) respectively.

$$M(x, y, t) = \sqrt{u_{x,y,t}^2 + v_{x,y,t}^2} \quad (3)$$

$$D_{(x,y,t)} = a \tan 2 \left(\frac{u_{x,y,t}}{v_{x,y,t}} \right)$$

Where $M(x,y,t)$, $D(x,y,t)$ are optical flow magnitude and direction value at each spatial location $u_{x,y,t}$ and $v_{x,y,t}$ are horizontal and vertical component of optical flow vector respectively.

Initially, each frame of video is divided into b blocks with 50% block overlapping and further each block

is divided into c cells e.g. an image is divided into 4 blocks/frame and each block is further divided into smaller parts or cells of some group of pixels. Then, for each cell, it builds separate magnitude and direction representation. Total feature vector length will be equal to $2*b*c*9$ (9 no of bins for each cell).

Figure 2 shows a representation how feature vector is extracted for an image, each small block has one arrow in some direction, arrow represents the optical flow value for particular pixel, arrow length shows the magnitude of optical flow value, more the length more optical flow value (more displacement in arrangement of pixels). Arrow direction shows the direction of displacement of pixel arrangement. WOHOFD and WOHOFM are extracted for each block as described below and histograms of all blocks are concatenated for final feature vector.

Weighted Oriented Histogram of Optical Flow Direction (WOHOFD)

The extraction of WOHOFD provides a histogram $HD_{b,t} = [h_1, h_2, \dots, h_c]$ at each time instant t , for b^{th} block in the frame (where c is c^{th} cell of b^{th} block), in which each flow vector is binned according to its primary optical flow direction from the horizontal axis and weighted according to its optical flow magnitude. For each cell in frame, Direction angle D_c is binned into 9 bins and weighted (Qian et al., 2011; Kaltsa et al., 2015) according to magnitude $MWeight$ given by Eq.(5)

$$MWeight_{x,y,c} = \left(\frac{M_{x,y,c} * 10^2}{\pi} \right) + \left(\frac{\max M_{x,y,c} * 10^2}{\pi} \right) \quad (4)$$

Where

$$\max M_{x,y,c} = M_{x,y,c} - \max(M_{x,y,c})$$

And

$$\overline{M}_{x,y,c} = M_{x,y,c} - \overline{M_{x,y,c}}$$

where $\overline{M}_{x,y,c}$ is average value of optical flow magnitude at all pixel positions within a cell. Then for each block, weighted oriented histogram of optical flow direction is obtained by concatenating histogram of cells.

Finally, a feature vector WOHOFD of frame is obtained by concatenating histogram of blocks $HD = [HD_1, HD_2, \dots, HD_b]$. Feature vector describes the global movements into successive frames.

Weighted Oriented Histogram of Optical Flow Magnitude (WOHOFM)

The WOHOFM provides a histogram $HM_{b,t} = [h_1, h_2, \dots, h_c]$ at each time instant t , for each block b in the frame, in which each flow vector is binned according to its optical flow magnitude and weighted according to its optical flow direction angle. For each cell c in frame, Magnitude M_c is binned into 9 bins and weighted according to magnitude $DWeight$ given by Eq.(6)

$$DWeight_{x,y,c} = \left(\frac{D_{x,y,c}}{\pi} * 10^2 \right) + \left(\frac{\max D_{x,y,c}}{\pi} * 10^2 \right) \quad (6)$$

Where

$$\max D_{x,y,c} = D_{x,y,c} - \max(D_{x,y,c})$$

And

$$D_{x,y,c} = D_{x,y,c} - \overline{D_{x,y,c}}$$

where $\overline{D_{x,y,c}}$ is average value of optical flow direction orientation at all pixel positions within a cell. Then for each block weighted oriented histogram of optical flow magnitude is obtained by concatenating histogram of cells. Finally, a WOHOFM of frame is obtained by concatenating histogram of blocks $HM = [HM_1, HM_2 \dots HM_b]$. Final feature vector for each frame is obtained by concatenating both histograms $[HD \text{ } HM]$ that describes the global movements into successive frames.

Classification

After feature extraction, pattern learning is required to analyze the extracted features and find pattern to classify between anomalous event and normal event or expected behavior. Real world crowd behaviors have much more complex distribution, that may not be modeled by single classifier (Gaussian model, SVM (Support vector Machine), and SVDD (Support Vector Data Descriptor) etc.), but combining multiple classifiers can work better for complex scene or data. Because each classifier focus on a selected features or distribution of data and by combining two or more classifiers will combine the strong points of each classifier (Tax & Duin, 2001). Based on this fact, our proposed work combines two one class classifiers: KNN (K-Nearest neighbor) and K-Means. Our training model is trained only on normal instances or normal behavior data. To make the outputs of the different classifiers comparable, confidence or posterior probabilities are estimated. Posterior probabilities ($P_m(t)$) for each test object t for each of the N classes on which the classifiers are trained and limited in the vicinity of 0 and 1. The posterior probabilities are normalized such that:

$$\sum_m^N P_m(t) = 1 \dots \quad (7)$$

In K-means, data points are described by k clusters and arranged in such way that the mean distance to a cluster is minimized. The objective function can be given as:

$$O(s) = \min_i (s - c_i)^2 \dots \quad (8)$$

where s is data points and c_i is i^{th} cluster center. While in KNN test objects are given to the class of the closest mean. Posterior probabilities are calculated by a sigmoid function. This is optimized over the training set using the maximum likelihood rule. KNN is an instance based learning classifier, it is used as base classifier and that performs classification based on the closest data point in feature space.

As the posterior probabilities $P_{nm}(t)$, $m = 1:N$, $n = 1:K$, for N classes and K classifiers are calculated, probabilities are combined to a new set of confidence $Q_m(t)$ for m class, which are used for final classification. The new probabilities $Q_m(t)$ for m class can be given as:

$$c_m(t) = \text{mean}(P_{nm}(t)) \dots \quad (9)$$

$$Q_m(t) = \frac{C_m(t)}{\sum_m Q_m(t)} \dots \quad (10)$$

Final decision is made by:

$$D(t) = \text{argmax}_m (Q_m(t)) \dots \quad (11)$$

RESULTS AND DISCUSSIONS

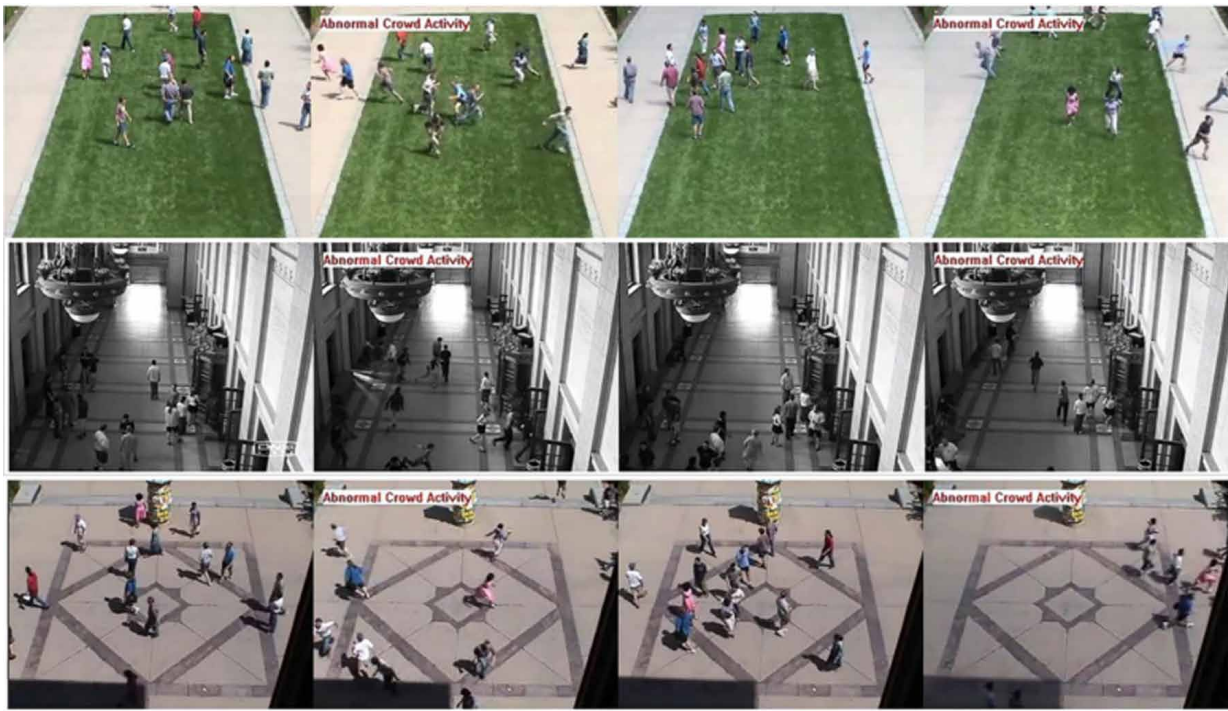
In this section, results of experiments based on the proposed method are presented. To analyze the performance of proposed method UMN dataset, PETS2009 dataset and Avenue Benchmark Dataset (Lu et al., 2013) are used. All simulation is performed in MATLAB and Intel XEON processor with 8GB RAM. The receiver operating characteristic plot (ROC) and Area Under the Curve (AUC) are used as a perform evaluation parameters. ROC curve is plotted between the true positive rate (Sensitivity- the probability of target is 1 same as it's true value which is also 1) on y axis and the false positive rate (100%-Specificity, the probability of target is 1 however its true value is 0) on x axis, for different cut off values. Each point on the ROC plot shows a sensitivity/specificity pair corresponding to a fixed threshold. A test with ideal segregation (no overlapping between two distributions) has a ROC curve that passes 100% sensitivity and 100 specificity. AUC value generally tells the quality of a classifier. AUC value equal to 0.5 suggest that classifier is randomly predicting the output, while the perfect classifier will have the AUC value 1. For real world application most of the classifier have an AUC value for 0.5 to 1. Our proposed method is also compared with the recent finding reported in the literature (Wang & Snoussi, 2014; Kaltsa et al., 2015; Shi et al., 2010)

UMN Dataset

UMN dataset have three different scenes of crowd escape behavior. This dataset is recorded by still camera, which is mounted at some elevation. One scene is of lawn, another is of indoor, and last one is plaza scene. First scene is picturized outdoor in a lawn; it contains total 1452 frames. Out of total frames, 1156 frames show normal behavior while 296 frames show crowd escape behavior. Indoor

scene contains total 4144 frames, 2986 frames showing normal situation while 1158 frames shows panic situation, Plaza scene contains 1836 normal situation frames, while 306 frames shows abnormal situation. In all three scenes, initially all persons are roaming all over the area at normal speed but after some time they started running due to some abnormal situation. Some frames of UMN dataset are shown in Figure 3, where 1st and 3rd column of figures showing normal activity while 2nd and 4th column of figures showing abnormal activity.

Figure 3. UMN dataset: Lawn, indoor and plaza scene (first and third column shows normal activity; Second and Forth column Abnormal activity)



Our model is trained only on normal activity frames and tested on rest of the frames. AUC value for all the three scenes of UMN dataset is compared with HOFO (Wang & Snoussi, 2014), HOS (Kaltsa et al., 2015) STCOG (Shi et al., 2010) and given in table 1. ROC curve and higher value of AUC suggest that our proposed method outperforms over other state of the art methods. ROC curves for all three scene are shown in Figures 4, 5, 6.

PETS2009 Dataset

The PETS2009 datasets contain three different sequences encompassing crowd situations with increasing complexity of scene. Dataset S1 is mainly for person count and density estimation. Dataset S2 involves people tracking. Dataset S3 addresses crowd flow analysis and abnormal event detection. The original resolution of the PETS2009 dataset is 728 x 576. For obtaining feature each frame is divided into smaller size blocks with 50% overlapping and each block is further divided in smaller size cells. The experiments are performed on view 1 for time sequence 14-17, 14-16, 14-06 and 14-55. Some frames of dataset are shown in Figure 7.

The detection results on PETS2009 (time sequence 14- 16) are shown in Figure 8. Where persons are running or walking from left to right or right to left direction. A normal situation corresponds to individuals walking at normal speed. While aberrant situation corresponds to the persons started running.

Figure 4. ROC curve for UMN Lawn

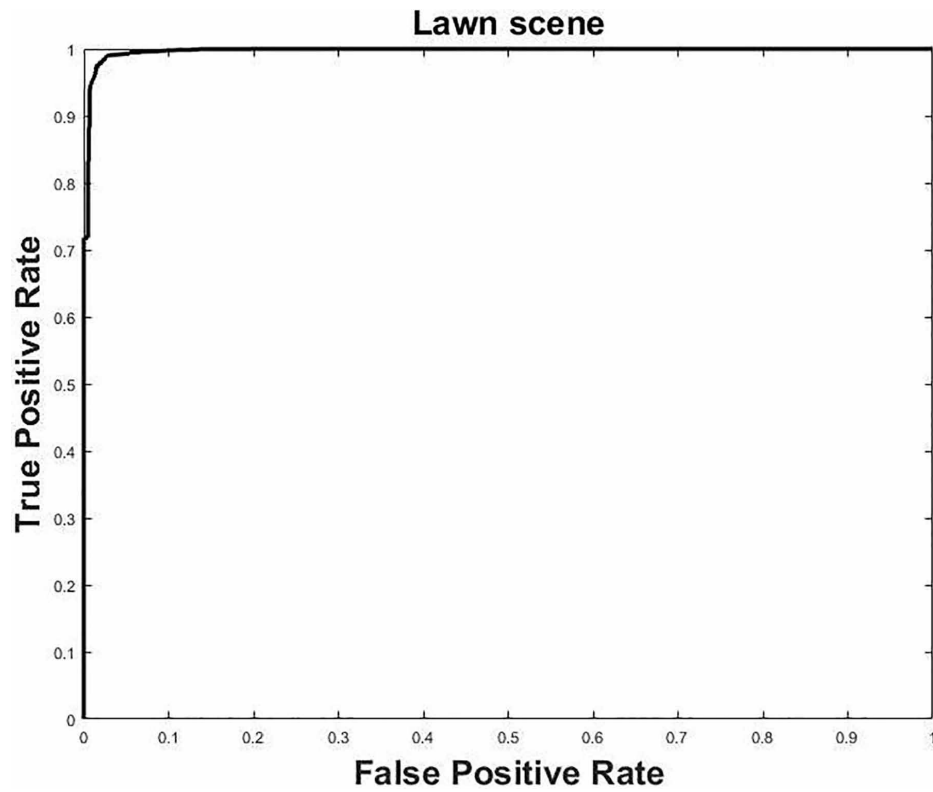


Figure 5. ROC curve for UMN indoor scene

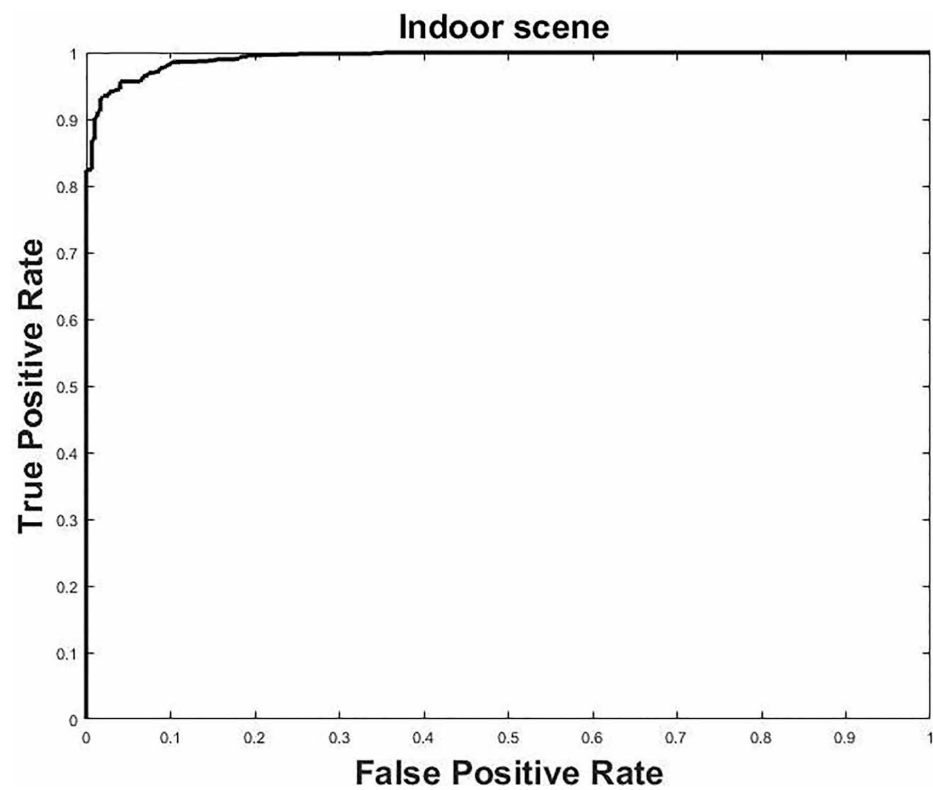


Figure 6. ROC curve for UMN Plaza scene

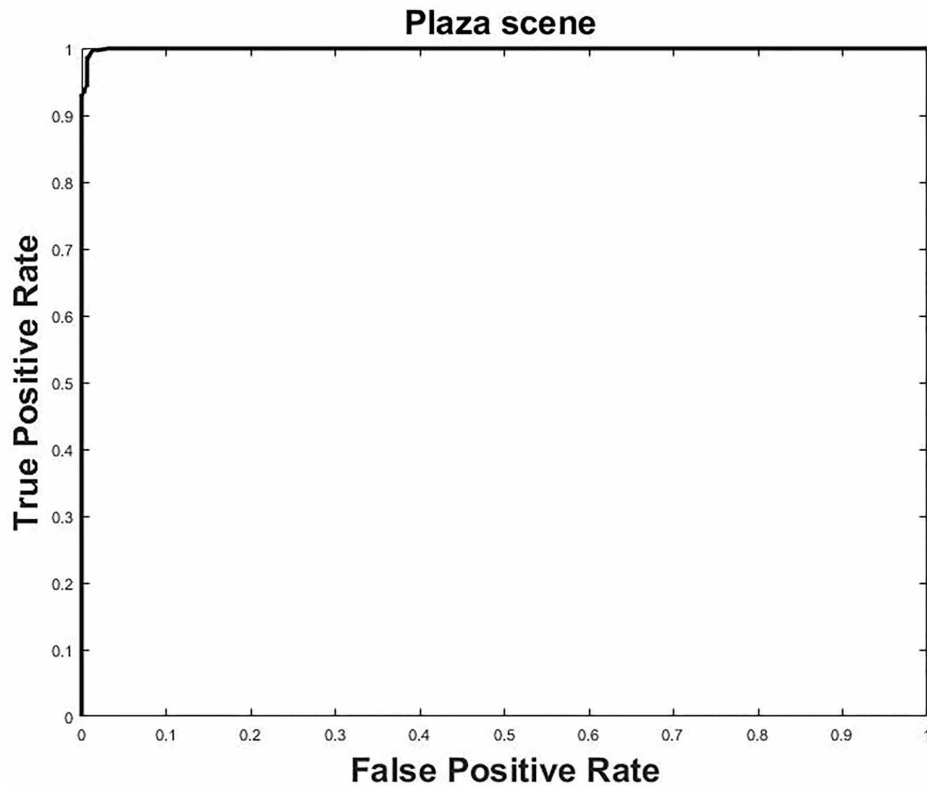


Table 1. Comparison results of our method with state of art methods (AUC performance of anomaly detection on UMN dataset)

S. No	Dataset	HOFO	HOS	STCOG	Proposed method
1	UMN Lawn Scene	0.9845	0.995	0.9362	0.9928
2	UMN Indoor Scene	0.9037	0.933	0.7759	0.9885
3	UMN Plaza Scene	0.9815	0.980	0.9661	0.9995

Classifier is trained only for normal situation where individuals are walking, training frames are chosen from time sequence 14-16 and 14-06. By our proposed method 94.34% detection accuracy is achieved. ROC curve for detection result on Time 14-16 is shown in Figure 9, AUC value for roc curve of time 14-16 is 0.9788.

The detection result on PETS2009 (time sequence 14-17) are shown in Figure 10, where walking is taken as normal situation and running is taken as abnormal situation.

Detection accuracy for time 14-17 is 91.78% and AUC value is 0.9543. ROC curve is shown in Figure 11.

Detection results for sequence (Time 14-06 and Time 14-55) are shown in Figure 12, where walking on the pedestrian way is taken as normal situation while walking on the grass or other than the pedestrian way is taken as abnormal situation.

Training frames are chosen from sequence (Time 14-06) and tested on sequence (Time 14-55). our method achieves 91.99% accuracy and AUC value is 0.9721, roc curve is shown in Figure 13.

Figure 7. PETS2009 dataset:(left column: normal scene right column: abnormal scene)

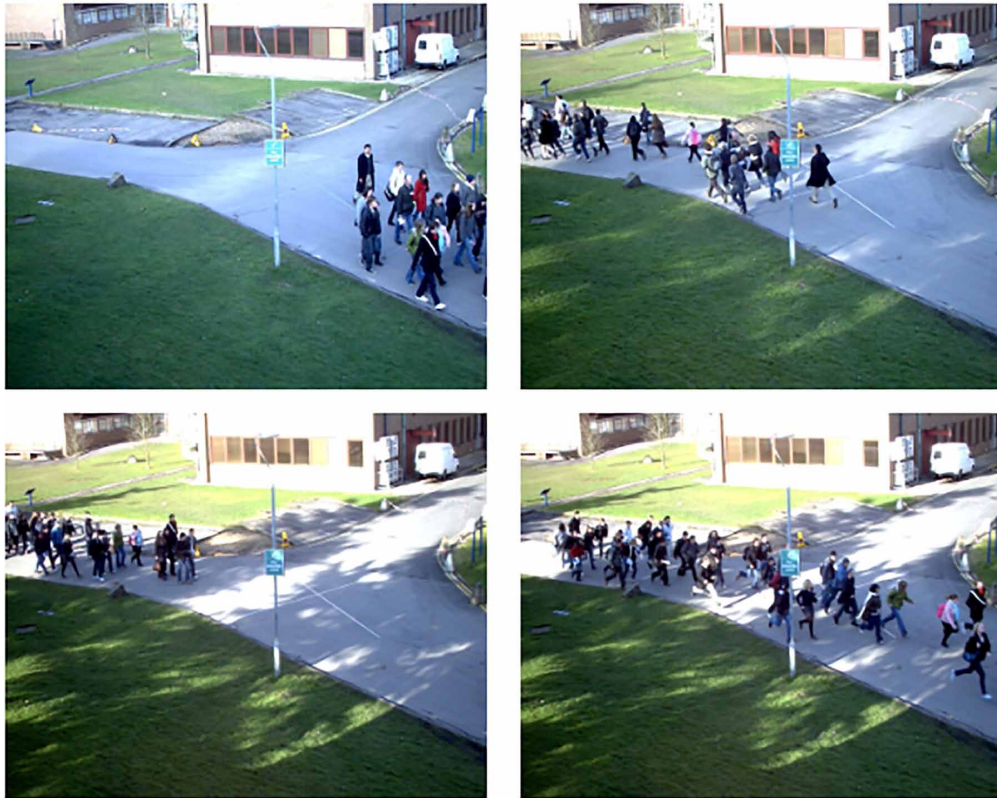


Figure 8. Detection results on PETS2009 (Time 14- 16)

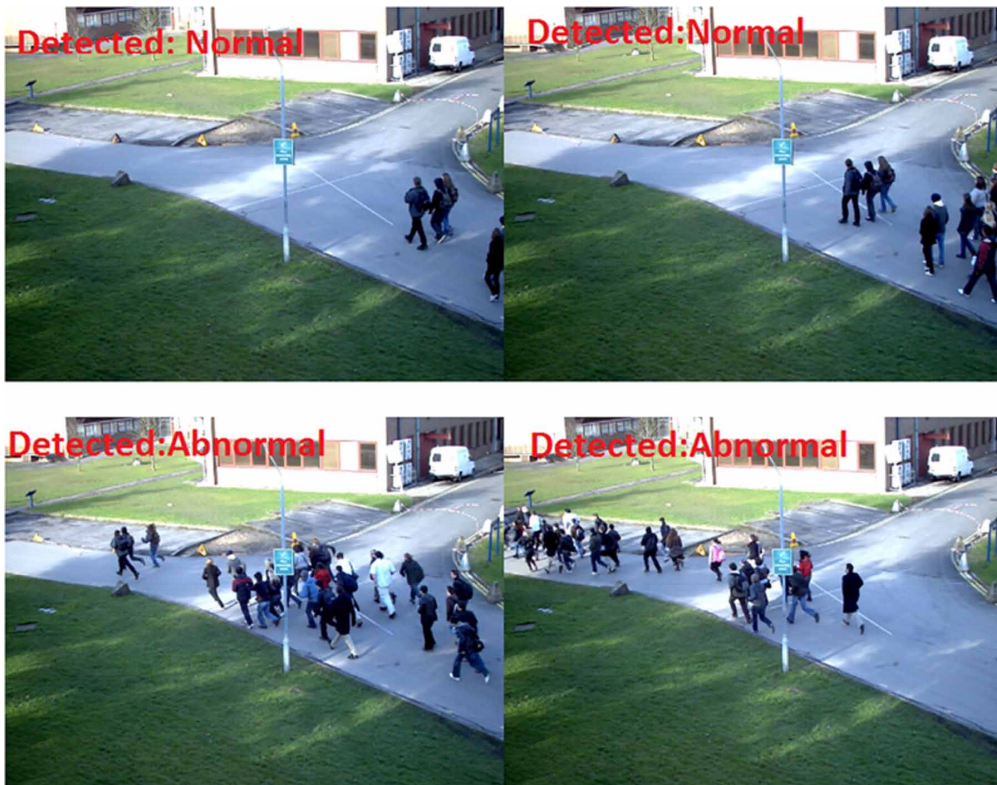


Figure 9. ROC curve for PETS2009 (Time 14- 16)

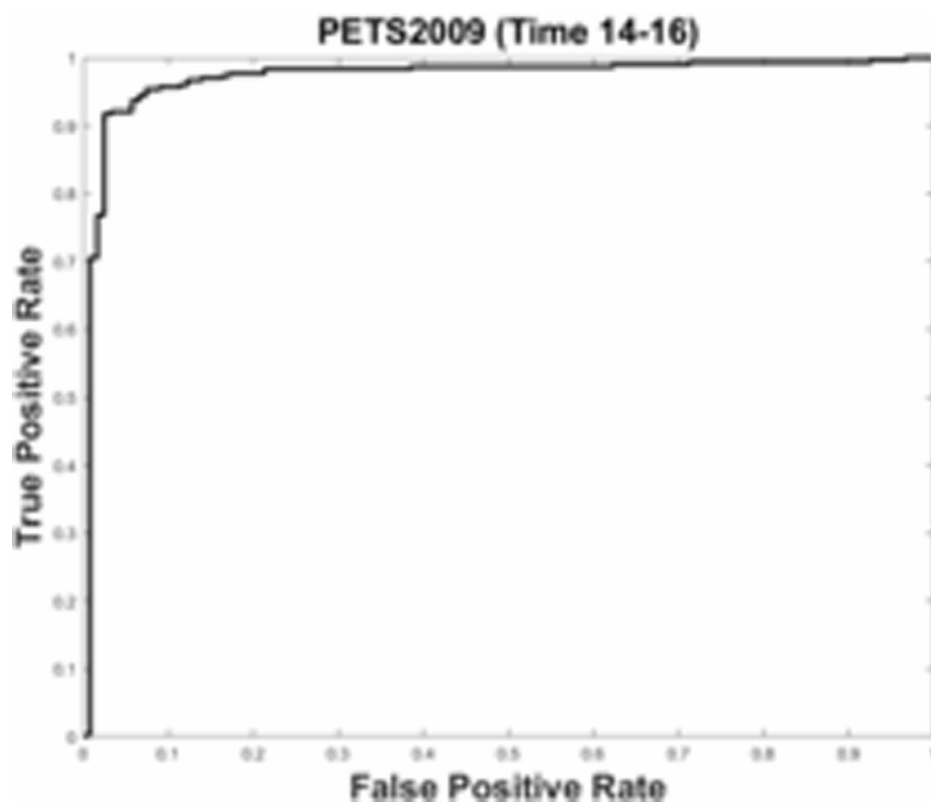


Figure 10. Detection results on PETS2009 (Time 14- 17)



Figure 11. ROC curve for PETS2009 (Time 14- 17)

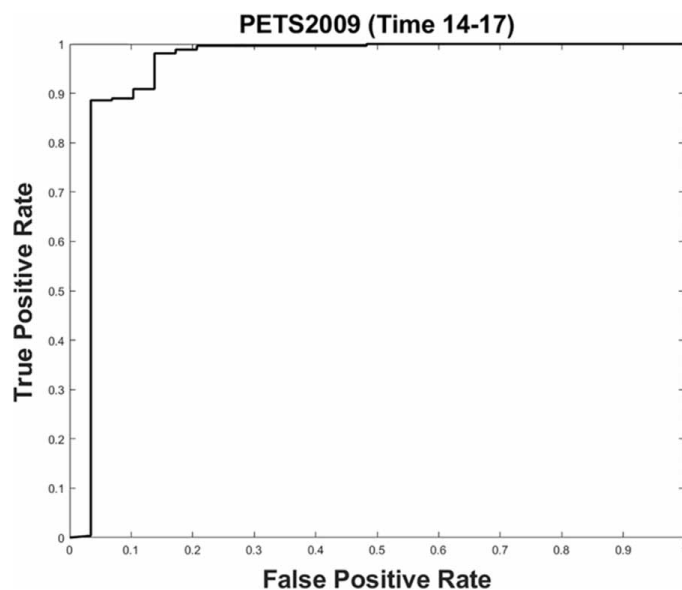
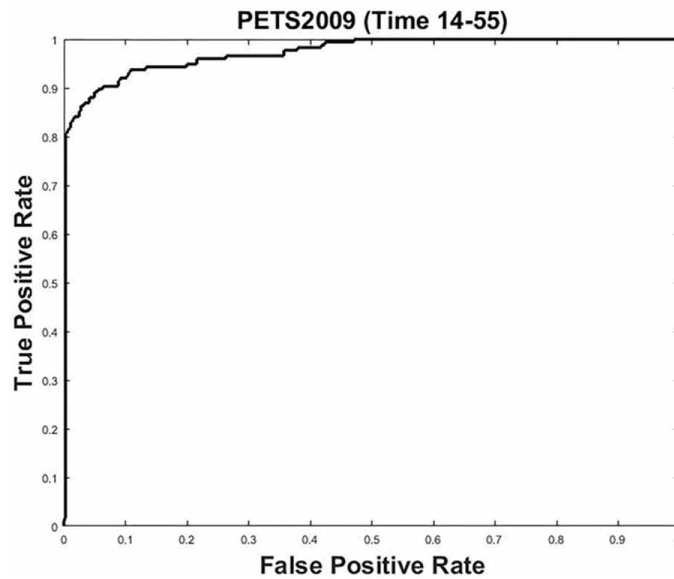


Figure 12. Detection results on PETS2009 (Time 14- 55)



Figure 13. ROC curve for PETS2009 (Time 14- 55)



Avenue Dataset

The Avenue dataset contains 16 training videos and 21 videos for testing. The videos include a total of 15183 frames for training and 15324 frames for testing. The resolution of video is 640*360. The training videos capture normal situations. Testing videos include both normal and abnormal events. This dataset contains abnormalities like persons running to and fro, persons are going towards such area from wrong direction, where only one way movement is allowed and persons are jumping. Three abnormal detected samples are shown in Figure 14.

This dataset also contains the challenges like minor camera shake (testing video 2, frame 1051 - 1100) presents, a few outliers are included in training data and some normal patterns seldom appear in training data. ROC Curve of detection result on Avenue dataset are shown in Figure 15 and comparison result of our method with (Lu et al., 2013) in table 2.

CONCLUSION

In this paper, a new and effective algorithm for suspicious event detection based on direction and magnitude is proposed. Crowd movement information is represented by movement direction and speed and model is trained on combined classifier (Simple but effective), which utilizes strong points of both classifier and perform very great in anomaly detection. ROC curve and high value of AUC of our model on different datasets show good performance of proposed algorithm over state-of-the-art methods. High detection rate on UMN, PETS2009 datasets and Avenue dataset demonstrate that our proposed algorithm can be used in challenging conditions specially in little bit noisy conditions. It's effectiveness in challenging crowded scenes, make our method very suitable for a wide variety of video surveillance applications. Our proposed method achieve good accuracy over different crowd datasets, however our algorithm is not suitable for real time application because optical flow calculation of each frame consumes a lot of time of this algorithm. So some different approach to obtain motion information can be used in future to make this algorithm work in real time. So in future, some other addition information can also be utilize e.g. deep learning, shape information etc., to make this algorithm work more efficiently and effectively in more challenging conditions.

Figure 14. Abnormality Detection (Wrong Direction, Running, Jumping) on Avenue Dataset



Figure 15. ROC curve for Avenue Dataset

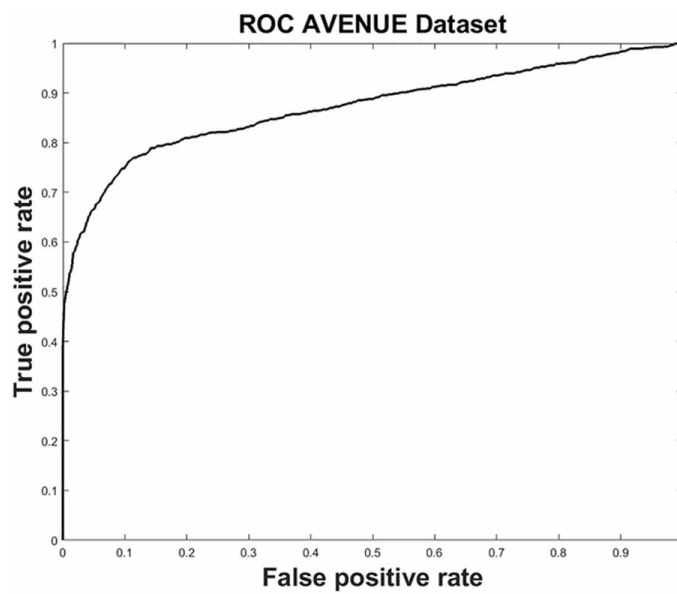


Table 2. Comparison results of our method with state of art methods (AUC performance of anomaly detection on the Avenue dataset)

S.No	Method	AUC
1	(Lu et al. 2013) given in (Del Giorno et al. 2016)	0.809
2	Proposed Method	0.871

REFERENCES

- Adam, A., Rivlin, E., Shimshoni, I., & Reinitz, D. (2008). Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3), 555–560. doi:10.1109/TPAMI.2007.70825 PMID:18195449
- Andrade, E. L., Blunsden, S., & Fisher, R. B. (2006, August). Modelling crowd scenes for event detection. In *18th international conference on pattern recognition (ICPR'06)* (Vol. 1, pp. 175-178). IEEE. doi:10.1109/ICPR.2006.806
- Basharat, A., Gritai, A., & Shah, M. (2008, June). Learning object motion patterns for anomaly detection and improved object detection. In *2008 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-8). IEEE. doi:10.1109/CVPR.2008.4587510
- Benezeth, Y., Jodoin, P. M., Saligrama, V., & Rosenberger, C. (2009, June). Abnormal events detection based on spatio-temporal co-occurrences. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2458-2465). IEEE. doi:10.1109/CVPR.2009.5206686
- Boiman, O., & Irani, M. (2007). Detecting irregularities in images and in video. *International Journal of Computer Vision*, 74(1), 17–31. doi:10.1007/s11263-006-0009-9
- Cong, Y., Yuan, J., & Liu, J. (2011, June). Sparse reconstruction cost for abnormal event detection. In *CVPR 2011* (pp. 3449–3456). IEEE. doi:10.1109/CVPR.2011.5995434
- Cui, X., Liu, Q., Gao, M., & Metaxas, D. N. (2011). Abnormal detection using interaction energy potentials. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 3161–3167.
- Del Giorno, A., Bagnell, J. A., & Hebert, M. (2016, October). A discriminative framework for anomaly detection in large videos. In *European Conference on Computer Vision* (pp. 334-349). Springer. doi:10.1007/978-3-319-46454-1_21
- Fu, Z., Hu, W., & Tan, T. (2005, September). Similarity based vehicle trajectory clustering and anomaly detection. In *IEEE International Conference on Image Processing 2005* (Vol. 2, pp. II-602). IEEE.
- Gu, X., Cui, J., & Zhu, Q. (2014). Abnormal crowd behavior detection by using the particle entropy. *Optik (Stuttgart)*, 125(14), 3428–3433. doi:10.1016/j.ijleo.2014.01.041
- Hamid, R., Johnson, A., Batta, S., Bobick, A., Isbell, C., & Coleman, G. (2005, June). Detection and explanation of anomalous activities: Representing activities as bags of event n-grams. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 1031-1038). IEEE. doi:10.1109/CVPR.2005.127
- Javan Roshtkhari, M., & Levine, M. D. (2013). Online dominant and anomalous behavior detection in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2611-2618). doi:10.1109/CVPR.2013.337
- Kaltsa, V., Briassouli, A., Kompatsiaris, I., Hadjileontiadis, L. J., & Strintzis, M. G. (2015). Swarm intelligence for detecting interesting events in crowded environments. *IEEE Transactions on Image Processing*, 24(7), 2153–2166. doi:10.1109/TIP.2015.2409559 PMID:25769154
- Kim, J., & Grauman, K. (2009, June). Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2921-2928). IEEE. doi:10.1109/CVPR.2009.5206569
- Kratz, L., & Nishino, K. (2009, June). Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1446-1453). IEEE. doi:10.1109/CVPR.2009.5206771
- Kwon, J., & Lee, K. M. (2014). A unified framework for event summarization and rare event detection from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9), 1737–1750. doi:10.1109/TPAMI.2014.2385695 PMID:26353123
- Lu, C., Shi, J., & Jia, J. (2013). Abnormal event detection at 150 fps in matlab. *Proceedings - IEEE International Conference on Computer Vision*, 2720–2727. doi:10.1109/ICCV.2013.338

- Lu, C., Shi, J., & Jia, J. (2013). Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision* (pp. 2720-2727). doi:10.1109/ICCV.2013.338
- Lucas, B. D., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *IJCAI'81 Proceedings of the 7th international joint conference on Artificial intelligence*, 674–679.
- Mehran, R., Oyama, A., & Shah, M. (2009, June). Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 935-942). IEEE. doi:10.1109/CVPR.2009.5206641
- Mishra, S. R., Mishra, T. K., Sarkar, A., & Sanyal, G. (2020). Detection of Anomalies in Human Action Using Optical Flow and Gradient Tensor. In *Smart Intelligent Computing and Applications* (pp. 561–570). Springer. doi:10.1007/978-981-13-9282-5_53
- Ouivirach, K., Gharti, S., & Dailey, M. N. (2013). Incremental behavior modeling and suspicious activity detection. *Pattern Recognition*, 46(3), 671–680. doi:10.1016/j.patcog.2012.10.008
- PETS. (2009). *Benchmark dataset*. <http://cs.binghamton.edu/~mrldata/pets2009>
- Qian, H., Wu, X., & Xu, Y. (2011). Dynamic analysis of crowd behavior. In *Intelligent Surveillance Systems* (pp. 119–154). Springer. doi:10.1007/978-94-007-1137-2_8
- Saligrama, V., & Chen, Z. (2012, June). Video anomaly detection based on local statistical aggregates. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2112-2119). IEEE. doi:10.1109/CVPR.2012.6247917
- Sharif, M. H., & Djeraba, C. (2012). An entropy approach for abnormal activities detection in video streams. *Pattern Recognition*, 45(7), 2543–2561. doi:10.1016/j.patcog.2011.11.023
- Shi, Y., Gao, Y., & Wang, R. (2010, August). Real-time abnormal event detection in complicated scenes. In *2010 20th International Conference on Pattern Recognition* (pp. 3653-3656). IEEE. doi:10.1109/ICPR.2010.891
- Tax, D. M., & Duin, R. P. (2001, July). Combining one-class classifiers. In *International Workshop on Multiple Classifier Systems* (pp. 299-308). Springer. doi:10.1007/3-540-48219-9_30
- UMN. (n.d.). *Unusual event datasets of University of Minnesota*. <http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi>
- Wang, T., Miao, Z., Chen, Y., Zhou, Y., Shan, G., & Snoussi, H. (2019). AED-Net: An Abnormal Event Detection Network. *Engineering*, 5(5), 930–939. doi:10.1016/j.eng.2019.02.008
- Wang, T., & Snoussi, H. (2014). Detection of abnormal visual events via global optical flow orientation histogram. *IEEE Transactions on Information Forensics and Security*, 9(6), 988–998. doi:10.1109/TIFS.2014.2315971
- Wu, S., Moore, B. E., & Shah, M. (2010, June). Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 2054-2060). IEEE. doi:10.1109/CVPR.2010.5539882
- Zen, G., & Ricci, E. (2011, June). Earth mover's prototypes: A convex learning approach for discovering activity patterns in dynamic scenes. In *CVPR 2011* (pp. 3225–3232). IEEE. doi:10.1109/CVPR.2011.5995578
- Zhang, D., Gatica-Perez, D., Bengio, S., & McCowan, I. (2005, June). Semi-supervised adapted hmms for unusual event detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 611-618). IEEE. doi:10.1109/CVPR.2005.316
- Zhao, B., Fei-Fei, L., & Xing, E. P. (2011, June). Online detection of unusual events in videos via dynamic sparse coding. In *CVPR 2011* (pp. 3313–3320). IEEE. doi:10.1109/CVPR.2011.5995524


Gajendra Singh received the Bachelor degree in Electronics and communication engineering from GLAITS, Mathura and his post-graduation degree in Electronics and communication engineering from NIT, Jalandhar. His research interests are Image Processing, Deep Learning and Machine Learning.

Rajiv Kapoor is professor at Delhi Technological University, Delhi. He was also Principle of AI&T, Delhi. He has published hundreds of research papers on Image and signal analysis in journal and conferences. His research interest includes Vision/Speech based Tracking, Activity Recognition Vision/Speech based, Signal Processing, Pattern Recognition, Cognitive Radio.

Arun Khosla is Professor at NIT Jalandhar. His areas of interest are Image Processing, Assisted Technologies and Games for Health. He is Fellow of IETE and Life Member of ISTE and Computer Society of India.



Optimal design of FOPID Controller for the control of CSTR by using a novel hybrid metaheuristic algorithm

NEHA KHANDUJA* and BHARAT BHUSHAN

Department of Electrical Engineering, Delhi Technological University, Delhi, India
e-mail: nehakhanduja.dce@gmail.com

MS received 20 October 2020; revised 2 April 2021; accepted 12 April 2021

Abstract. The escalating complexity in the process control industry emanates the demand for novel and advanced control techniques, which results in enhanced performance indices. A hybrid optimal control method i.e., FOPID control using chaotic state of matter search with elite opposition-based learning for controlling CSTR is proposed in this paper. Fractional order PID is a generalized form of PID Controller. It uses fractional calculus, resulting in a more flexible and better response accompanying rigorous adoption for substantially closed-loop system stability. Hybridization of SMS with chaotic maps and elite oppositional-based learning results in enhanced exploration capability along with randomization. In this paper, the results show that the CSMSEOB L tuned FOPID controller provides superior and optimum performance when compared to other metaheuristic algorithms.

Keywords. State of matter search algorithms; chaotic maps; elite opposition based learning; continuously stirred tank reactor.

1. Introduction

A literature review of process control depicts that in contempt of many new advanced, adaptive, and optimal control methodologies, the use of PID controllers has been stagnated, especially in the areas where reference tracking and disturbance rejection are the major tasks. Some key features which make PID popular are robust performance, self-explanatory, diversified application areas, simple implementation validation, and many more [1]. Although a simple PID controller provides the least impenetrable, most productive, and effortless tuning of controller parameters for the practical process. But with advantages, PID controllers have limitations also like the less optimal solution for a system loaded with non-linearity, time delay, high order disturbances, noise, etc. These limitations lead to introducing new and advanced tuning methods like Fuzzy Logic, Neural Network, Adaptive Control, Internal Model Control, etc. which ameliorate the capability and performance of the traditional PID controller [2] along with enhanced flexibility of conventional PID controller.

FOPID as an alternative can be adopted with five parameters to tune, whereas a conventional PID Controller has only three parameters. Although it increases the complexity of parameter tuning to some extent at the same time, resulting in comparatively fine-tuning [3]. FOPID is an advanced form of PID controller which is proposed by

Podlubny [4] as $PI^{\lambda}D^{\mu}$ controller, where λ and μ are non-integer order of integral and differential term, respectively. A literature review shows that FOPID gives better performance as compared to conventional PID Controller [5]. In continuation of this, presently many metaheuristic algorithms are in great demand for control tuning parameters of the PID controller [6]. The rising complexities in the research area result in limiting the mathematical methods of finding optimal solutions and this necessity results in the investigation of metaheuristic optimization algorithms. Major limitations with traditional methods of optimization are time-consuming, tedious, less efficient, and less accurate [7]. The imperative feature of the metaheuristic algorithm which makes it prominent among researchers is its adaptability and versatility. It can adapt to the problem and determine the optimal solution of different types of problems, whether it is related to mathematics, engineering, process industry, etc. [8]. Other features of the metaheuristic algorithm which makes it popular are:

- Can be easily integrated with the already existing implementation
- Wide applicability area
- Gradient information is not required.
- Decision-making is easy [9].

A complex problem can be solved in a reasonable time and may give an acceptable solution by using metaheuristic which is based on trial and error. The main objective is to produce an attainable solution in a reasonable time frame.

*For correspondence
Published online: 21 May 2021

Whenever a metaheuristic is chosen for a problem, it never guarantees the best solution and even, we are not known if it will give the optimal solution or not. The main point of selecting an algorithm is to give acceptable or accurate solutions most of the time with minimum deviation. Exploration (diversification) and exploitation (intensification) are two key components of any metaheuristic algorithm with exploration searches in undetected areas while exploitation searches in other promising territories in the sample space. Therefore, the success of an algorithm depends on a good balance between exploration and exploitation which leads to the assurance of convergence to optimality [10].

Most of the chemical processes such as continuous stirred tank reactor (CSTR), biochemical reactor, and conical tank systems persist dynamic and highly nonlinear behaviour as they consist of multiple process variables to be manipulated. Many advanced controlling and optimization methods are proposed to control such types of MIMO (Multi-input Multi-output) systems. Extensive Literature review shows that evolutionary techniques like PSO (Particle Swarm Optimization) [11, 12], IWO (Invasive Weed Optimization) [13], FS (Stochastic Fractal Search) [14] FFA (Firefly Algorithm) [15], GWO (Grey Wolf Optimizer) [16], CSO (Cat Swarm Optimization) [17] TLBO (Teacher-Learner based Optimization) [18, 19], SMS (State of Matter Search) [20], CKH (Chaotic Krill Herd) [21], RDO (Red Deer Optimization Algorithm) [22], SOA (Sailfish Optimization Algorithm) [23], and many more have proved their superiority as compared to traditional controllers like Z-N tuned PID, refined Ziegler-Nichols rule [24], intelligent controllers Fuzzy-PID [25], Neural-PID [26], Model-based controllers MRAC (Model reference adaptive control) [27] and Internal model control (IMC) [28].

The proposed methodology is used for concentration and temperature control of continuously stirred tank reactors (CSTR). A vast literature is available for controlling methodologies of CSTR but as it is highly nonlinear and its complex dynamics properties make it a complex problem. Therefore, it is a tedious task to control CSTR by the conventional controller [29]. Nowadays optimization-based control is preferred over the conventional or intelligent controller and to achieve this a hybrid CSMSEOB method is proposed. It is a modified form of SMS algorithm (state of matter search) in which, Chaotic Maps and Elite opposition-based learning (EOBL) are embedded with SMS to enhance the efficiency and efficacy of the SMS algorithm. The basic principle of the SMS algorithm lies in the heart of the thermal energy motion system. The whole algorithm is divided into three states of matter solid, liquid, and gas and each state persist of a different diversification-intensification ratio. The algorithm starts with the gas state and modifying the diversification-intensification ratio and ends at a solid state [20]. The chaotic concept is used for the systems which have high sensitivity towards

the initial condition, and also it increases the randomness because the range of random numbers is limited. The chaotic theory has been used with many evolutionary algorithms like PSO, Krill herd, BFO, etc. [21]. This concept of chaotic SMS algorithms is used to define some random variables to stimulate the convergence of SMS. Further, chaotic SMS is merged with elite opposition-based learning. The concept of OBL was introduced by Tizhoosh in 2005 which increases the exploration capability of the existing algorithm by combining two main properties of OBL which are a global search and good convergence rate [2]. EOBL is the superior form of OBL which gives better global search and a higher convergence rate [30]. A fractional-order PID control of CSTR using a hybrid metaheuristic algorithm CSMSEOB is implemented on MATLAB and results obtained from this hybrid algorithm prove the excellence of the proposed methodology.

The rest of the paper is organized as follows. Section 2 describes a non-linear problem of CSTR, section 3 elaborates the FOPID Controller, the considered Metaheuristic optimization techniques have been described in section 4. Results and discussion are illustrated in section 5, and the conclusions along with future scope are detailed in section 6.

2. Continuously stirred tank reactor (CSTR)

A Continuous Stirred Tank Reactor (CSTR) is one of the most significant unit tasks in the Chemical process industries. It shows a profoundly nonlinear nature and for the most part, has wide working ranges. Chemical responses in a reactor are either exothermic or endothermic and consequently necessitate that heat either be evacuated or added to the reactor to keep up a steady temperature [31]. A jacket encompassing the reactor additionally has fed and leaves streams. The jacket is thought to be entirely blended and at a lower temperature than the reactor, energy at that point goes through the reactor walls into the jacket, to evacuate the heat produced by the chemical reaction. Consider for uniform volume, exact blending, and uniform values of the parameter inside the reactor, the mass-energy balance condition is given by

$$f_1(C_A, T) = \frac{dC_A}{dt} = \frac{F}{V}(C_{Af} - C_A) - r \quad (1)$$

$$\begin{aligned} f_2(C_A, T) &= \frac{dT}{dt} \\ &= \frac{F}{V}(T_f - T) + \left(\frac{-\Delta H}{\rho C_p} \right) r - \frac{UA}{V\rho C_p}(T - T_j) \end{aligned} \quad (2)$$

where C_A stands for concentration, r stands for Arrhenius expression for a chemical reaction is given by

$$r = k_0 \exp\left(\frac{-\Delta E}{RT}\right) C_A$$

Figure 1 shows an irreversible and exothermic compound response that happens inside the reactor where a solitary coolant stream cools a consistent volume reactor [32].

The description of CSTR parameters is given in table 1 [33]. The main objective is to control the reactor temperature and concentration by controlling the cooling rate.

3. FOPID controller

The most common form of PID controller combines three kinds of corrective measures to the error signal, which is the representation of closeness or distance of the desired output from the actual one. In general, these three corrective measures are termed proportional, integral, and derivative. The general form of a PID Controller is given by [34].

$$u(t) = k_p e(t) + \frac{1}{k_i} \int_0^t e(\tau) d\tau + k_d \frac{de(t)}{dt} \quad (3)$$

Podlubny [35] proposed FOPID Controller in 1999 as an extended form of PID controller which has a comparatively wider range for controlling. The FOPID Controller is shown in figure 2 and represented as

$$u(t) = k_p e(t) + k_I D^{-\gamma} e(t) + k_D D^{\mu} e(t) \quad (4)$$

where γ and μ are real numbers with $\gamma > 0, \mu > 0$ [5], D is a fractional calculus operator which is defined by Riemann–Liouville as (n is general non-integer order and $\Gamma(n)$ is Euler's gamma function)

$$D^{-n} f(t) = \frac{1}{\Gamma(n)} \int_0^t f(y) (t-y)^{n-1} dy \quad (5)$$

The FOPID controller also takes current error, accumulated error, and predicted error into account same as classical PID controller but fractional operators are non-local in FOPID which gives a modified definition to the integral as well as derivative action [34]. For the analysis purpose,

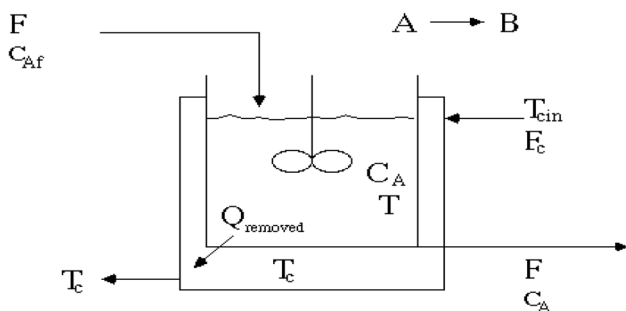


Figure 1. Schematic representation of Jacketed CSTR system.

fractional calculus equations must be transferred into algebraic equations. The Laplace transform of the equation for $D^{-n}f(t)$ can be expressed as

$$\int_0^{\infty} e^{-st} f(t) dt = s^{-n} F(s) \quad (6)$$

Here, it is assumed that all initial conditions are zero [36]. Case I, if $\gamma = 1$ and $\mu = 1$ results in PID controller. Case II, if $\gamma = 1$ and $\mu = 0$, results in PI Controller. Case III, if $\gamma = 0$ and $\mu = 1$, resulting in PD controller. Case IV, if $\gamma = 0$ and $\mu = 0$ results in gain controller only. The transfer function of FOPID Control [37] is represented as,

$$G_C(S) = \frac{U(S)}{E(S)} \left(k_p + k_I \frac{1}{S^{\gamma}} + k_D S^{\mu} \right) \quad (7)$$

Use of FOPID Controller results not only in enhanced performance of the control system, better adaptability but fine control of the dynamical system as well as very fewer variations in parameters of a control system [38].

- Fractional-order linear matrix diversity framework
- The powerful interim check technique
- Fractional-order Lyapunov disparity technique [39].

4. Metaheuristic optimization algorithms

Figure 3 gives the flow of the proposed work with parameters of the FOPID Controller which are optimized by the metaheuristic optimization and controlled parameters are fed into the process.

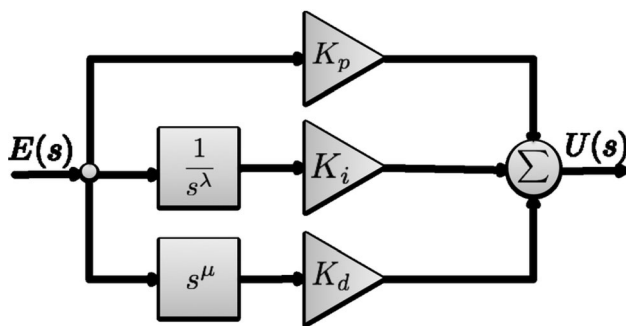
4.1 Particle swarm optimization (PSO)

PSO is a swarm intelligence-based optimization algorithm that was proposed by Kennedy and Eberhart in 1995. It simulates the concept of cooperation, communication, and social behavior in fish and bird schooling. Literature [40] reveals that extensive research has been done on PSO to demonstrate its efficiency in solving real-valued complex, non-linear, non-differentiable optimization problems. However, since the search space dimension can be sufficiently increased, PSO is sensitive to the trend of falling into local optima. To solve this limitation with traditional PSO some improved and hybridized version of PSO has been introduced from time to time to enhance its convergence performance [41]. It is a population-based optimization technique that gives rise to high-quality results within a more concise time and shows stable converge characteristics [33].

PSO is an iterative process. On each iteration of the PSO's main processing loop, each particle's current velocity is first updated based on the particle's current

Table 1. CSTR Parameters.

Reactor Parameter	Description	Values
F/V (hr-1)	Flow rate*reactor volume of the tank	1
K_o (hr-1)	Exponential factor	$10e^{15}$
$-\Delta H$ (kcal/kmol)	Heat of reaction	6000
E (kcal/kmol)	Activation energy	12189
ρC_P (BTU/ ft ³)	Density*heat capacity	500
T_f (°K)	Feed temperature	315
C_{Af} (lbmol/ft ³)	The concentration of feed stream	1
$\frac{UA}{V}$	Overall heat transfer coefficient/reactor volume	1451
T_j (K)	Coolant Temperature	300

**Figure 2.** FOPID Controller.

velocity, the particle's local information, and global swarm information. Then, each particle's position is updated using the particle's new velocity. In math terms the two update equations are:

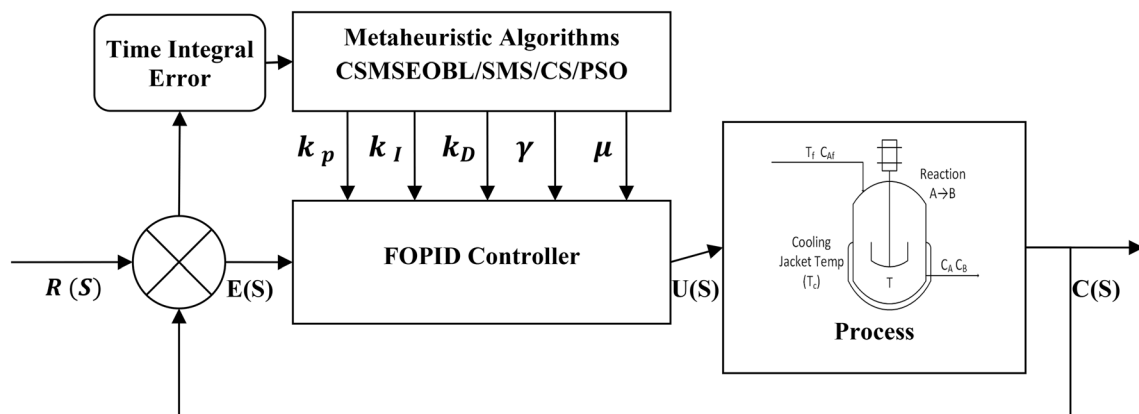
$$\begin{aligned} v(t+1) = & (w * v(t)) + (c_1 * r_1 * (p(t) - x(t)) \\ & + (c_1 * r_2 * (g(t) - x(t))) \end{aligned} \quad (8)$$

$$x(t+1) = x(t) + v(t+1) \quad (9)$$

where r_1 and r_2 are random numbers with a value between $[0, 1]$, c_1 and c_2 are two acceleration constants, w is inertia weight, $x(\cdot)$ is the position of the particle, $p(t)$ is the personal best position of the particle, $g(t)$ is the global best position of the group. The term $v(t+1)$ means the velocity at time $(t+1)$. Once the new velocity, $v(t+1)$, has been determined, it is used to compute the new particle position $x(t+1)$ [42].

4.2 Cuckoo search algorithm (CS)

Yang and Deb [43] developed novel meta-heuristic calculations cuckoo search in 2009. The CS depends on the brood parasitism of some cuckoo species. Moreover, the calculation is upgraded by the purported Lévy flights, as opposed to basic isotropic irregular strolls. Cuckoos are interesting flying creatures, not just as a result of the excellent sounds they can make yet additionally on account of their forceful generation methodology. A few animal types, for example, the ani and guira cuckoos lay their eggs in shared homes, however, they may evacuate others' eggs

**Figure 3.** Optimized FOPID for Control of CSTR.

to expand the bring forth likelihood of their eggs. A lot of animal varieties connect with the committed brood parasitism by laying their eggs in the homes of other host winged animals [30]. In cuckoo search calculation cuckoo egg speaks to a potential answer for the structure issue which has an objective function. The calculation utilizes three glorified guidelines:

- Each cuckoo lays each egg in turn and dumps it in a haphazardly chosen home.
- The best home with great eggs will be extended to the next generation.
- The quantity of accessible host homes is fixed and a host winged animal can find an outsider egg with a probability of $P_a \in [0, 1]$ [44].

4.3 State of matter search algorithm (SMS)

SMS algorithm is a nature-inspired algorithm that lies in the category of evolutionary algorithm and can be used to solve MIMO type global optimization problems. It is based on a thermal energy motion mechanism. Three states of matter i.e., solid, liquid, and gas are simulated in this algorithm and each state has a different exploration-exploitation ratio. The algorithm begins with the gas state which is purely exploration, then after reforming the exploration and exploitation ratio it reaches a liquid state in which a moderate transition takes place between exploration and exploitation, and this reforming is continued till solid-state i.e., pure exploitation is reached. This entire process results in the enhancement of population diversity and simultaneously escapes the particles to concentrate within local minima [20]. The complete SMS Algorithm can be a divided into four stages:

Stage 1: Initialization state and general procedure:

- Find the best element from population P

$$P^{best} \in \{P\} | f(P^{best}) = \max\{f(P_1), f(P_2), \dots, f(P_{N_r})\} \quad (10)$$

- Calculate initial velocity magnitude

$$v_{st} = \frac{\sum_{j=1}^n (b_j^h - b_j^l)}{n} * \beta \quad (11)$$

where, b_j^h is the upper bound of j parameter, b_j^l is lower bound of j parameter, β is a factor ranging [0,1]

- Update the direction vector to control the movement of the particle

$$d_i^{k+1} = d_i^k \left(1 - \frac{k}{gen}\right) 0.5 + a_i \quad (12)$$

$$a_i = \frac{(p_{best} - p_i)}{\|p_{best} - p_i\|}$$

where, a_i attraction unitary vector, p_{best} is the best molecule in population P , p_i is molecule i of population P , k is current iteration number; gen -total number of iterations.

- Calculate velocity, v_i of each molecule

$$v_i = d_i * v_{st} \quad (13)$$

- Calculate collision radius, r and $0 \leq \alpha \leq 1$

$$r = \frac{\sum_{j=1}^n (b_j^h - b_j^l)}{n} * \alpha \quad (14)$$

- Then update the Position of each molecule, which is given by (H is a threshold limit)

$$\begin{aligned} &= p_{i,j}^k + v_{i,j} * rand(0, 1) * \rho * (b_j^h - b_j^l); & \text{if } rand \leq H \\ p_{i,j}^{k+1} &= p_{i,j}^k; & \text{if } rand > H \end{aligned}$$

(15)

Stage 2: Gas state

- Set the parameters for the gas state: $\rho \in [0.8, 1]$, $\beta = 0.8$, $\alpha = 0.8$ & $H = 0.9$.
- Apply the general procedure as described in Stage 1.
- If the no. of iteration=50% of total no. of iterations then the process shifted to liquid state otherwise the general procedure is repeated.

Stage 3: Liquid State

- Set the parameters for the liquid state: $\rho \in [0.3, 0.6]$, $\beta = 0.4$, $\alpha = 0.2$ & $H = 0.2$.
- Apply the general procedure as described in Stage 1.
- If no. of iteration=90% of total no. of iterations then process shifted to solid-state otherwise the general procedure is repeated.

Stage 4: Solid State

- Set the parameters for solid-state: $\rho \in [0.0, 0.1]$, $\beta = 0.1$, $\alpha = 0$ & $H = 0$.
- Apply the general procedure as described in Stage 1.
- If the total no. of iteration=100% then the process is finished otherwise the general procedure is repeated.

4.4 CSMS-EOBL algorithm

A hybrid metaheuristic approach is used to enhance the balance between exploration and exploitation capability of the existing algorithm along with accelerated convergence rate. The benefits of all three algorithms are combined to form this hybrid algorithm. Chaotic Maps are used to calculate the random variable of the SMS algorithm and increase the exploitation capability. Further, the inclusion of EOBL enhances the exploration capability of the SMS Algorithm.

Table 2. Chaotic maps [45].

Name	Chaotic Map	Range
Chebyshev	$x_{k+1} = \cos(k \cos^{-1}(x_k))$	(0,1)
Circle	$x_{i+1} = \text{mod}\left(x_i + b - \left(\frac{a}{2\pi}\right) \sin(2\pi x_k), 1\right) a = 0.5, b = 0.2$	(0,1)
Gauss	$x_{i+1} = \begin{cases} 1, & x_i = 0 \\ \frac{1}{\text{mod}(x_i, 1)}, & \text{otherwise} \end{cases}$	(0,1)
Iterative	$x_{k+1} = \sin\left(\frac{ax}{x_k}\right), a \in (0,1)$	(0,1)
Logistic	$x_{i+1} = ax_i(1 - x_i), a = 4$	(0,1)
Piecewise	$\begin{aligned} &\frac{x_k}{P}; 0 \leq x_k < P \frac{x_{k-} - P}{0.5 - P}; \\ &P \leq x_k < 0.5 \frac{1 - P - x_{k-}}{0.5 - P}; \\ &0.5 \leq x_k < 1 - P \frac{1 - x_{k-}}{P}; \\ &1 - P \leq x_k < 1 \end{aligned}$	(0,1)
Sine	$\frac{a}{4} \sin(\pi x_k); 1 < a < 4$	(0,1)
Singer	$\begin{aligned} x_{i+1} &= \mu(7.86x_i - 23.31x_i^2 + 28.75x_i^3 - 13.302875x_i^4) \\ \mu &= 1.07 \end{aligned}$	(0,1)
Sinusoidal	$x_{i+1} = ax_i^2 \sin(\pi x_i), a = 2.3$	(0,1)
Tent	$x_{i+1} = \begin{cases} \frac{x_i}{0.7} & x_i < 0.7 \\ \frac{10(1 - x_i)}{3} & x_i \geq 0.7 \end{cases}$	(0,1)

4.4.1 Chaotic theory and maps Chaos is a deterministic concept that shows irregular motions and can be used in numerous applications like non-linear control, automobile, industrial applications, etc. It is a randomly based optimization algorithm that uses chaotic variables instead of random variables. The concept of chaos possesses three important properties of non-recurrence, randomness, and dynamic [21]. These features of chaos ensure that various solutions produced by the algorithm can

search on the complex multimodal landscape at a higher speed with various movement patterns. Hundreds of metaheuristic algorithms have been designed to achieve a good balance between exploration and exploitation, according to the literature on metaheuristic algorithms. In this thread, chaotic theory or COA (chaotic-based optimization algorithm) can be understood as a system that is nonlinear, highly sensitive to initial conditions, and possesses the properties of randomness and non-

Table 3. Parameter Setting of Metaheuristic Algorithms.

Algorithm and Parameters	Parameter Value	Algorithm and Parameters	Parameter Value
PSO		CS	
Population	50	Population	50
Iteration	25	Iteration	25
Weight Function	[0.2,0.9]	Pa	0.25
Acceleration constants	2	Beta	1.5
The dimension of search space	5	The dimension of search space	5
Iteration	25	Iteration	25
SMS		CSMS-EOBL	
Vector Adjustment, ρ	1	Vector Adjustment, ρ	1
Beta	[0.8, 0.4, 0.1]	Beta	[0.8, 0.4, 0.1]
Alpha	[0.8, 0.2, 0]	Alpha	[0.8, 0.2, 0]
Threshold Probability, H	[0.9, 0.2, 0]	Threshold Probability, H	[0.9, 0.2, 0]
Phase Percent	[0.5, 0.1, -0.1]	Phase Percent	[0.5, 0.1, -0.1]
Adjustment Parameters	[0.85 0.35 0.05]	Adjustment Parameters	[0.85 0.35 0.05]
Iteration	25	Iteration	25

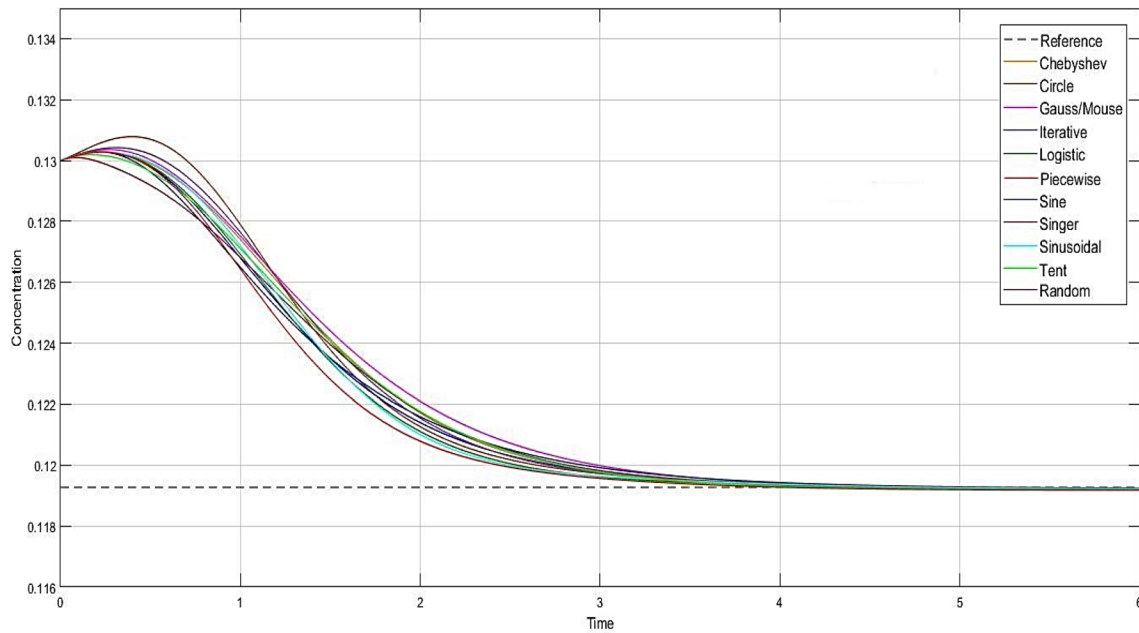


Figure 4. Simulation result for concentration control of CSTR with CSMSEOBL algorithm for different types of Chaotic Maps.

recurrence [45]. Whenever any metaheuristic optimization algorithm is embedded with chaotic maps it has three steps: Initialization, operator, and random generator (table 2).

4.4.2 Elite opposition based learning algorithm (EOBL) Tizhoosh proposed opposition-based learning (OBL) in 2005 [46] and from the day of its introduction this learning has been embedded with many metaheuristic optimization algorithms to enhance the exploration capability along with the accelerated convergence rate of

the existing metaheuristic. The main concept of OBL lies in the fact that the current population and its opposite population are considered at the same point in time. In 2013, Wang *et al* proposed a modified form of OBL Strategy called Elite Opposition Based Learning (EOBL). It uses dynamic bounds instead of fixed bounds which make the search space shortening also, elite oppositional numbers are defined at the center point of search space which results in better convergence and exploration capability [47]. As EOBL is a modified form of OBL so the first OBL is explained.

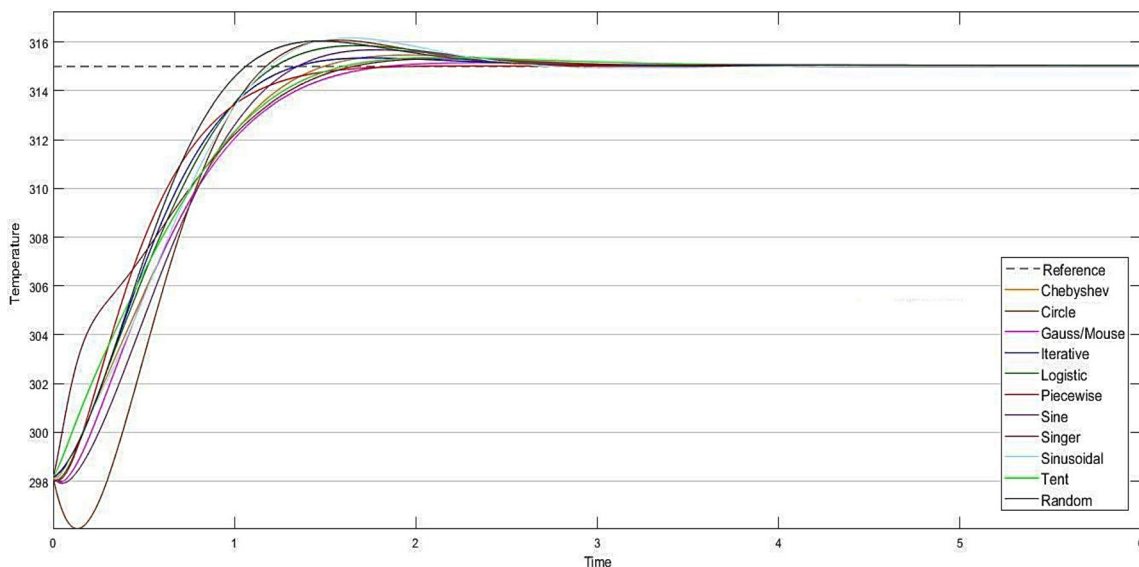


Figure 5. Simulation result for temperature control of CSTR with CSMSEOBL algorithm for different types of Chaotic Maps.

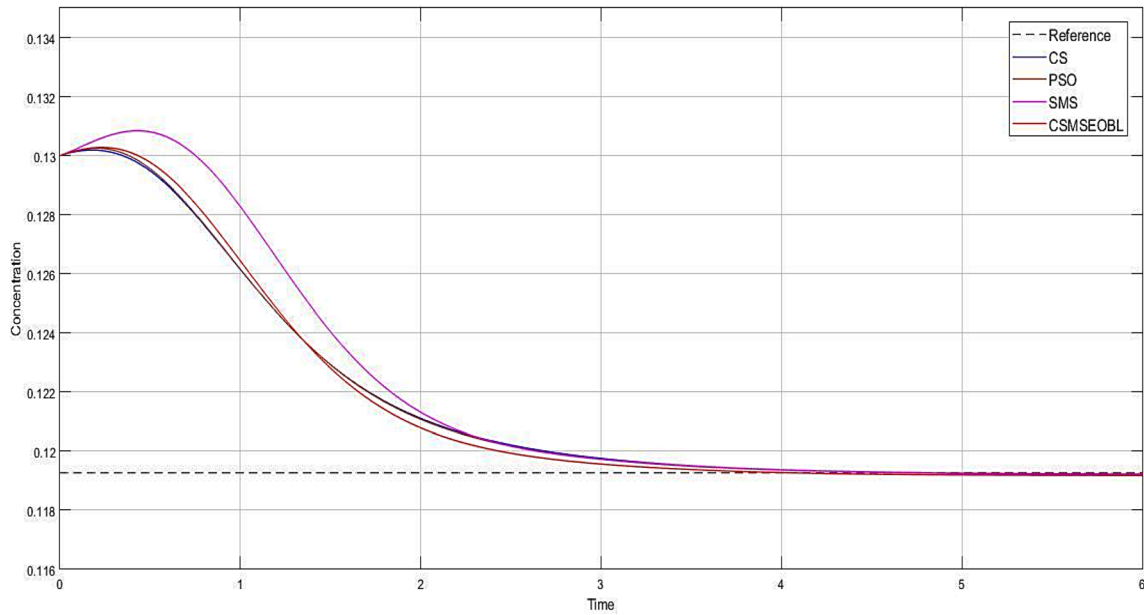


Figure 6. Comparison of concentration control of the CSTR system among different metaheuristic algorithms.

Let $\mathbf{x} = \{x_1, x_2, \dots, x_j\}$ is a point in the existing population and j is the dimension of search space, $x_j \in [a_j, b_j]$, where, $a_j = \min\{x_{ij}\}$ and $b_j = \max\{x_{ij}\}$. The opposite point of \mathbf{x} can be defined as follows:

$$\tilde{x}_j = a_j + b_j - x_j \quad (16)$$

The further elite individual in the current population is defined as $\mathbf{x}_e = \{x_{e1}, x_{e2}, \dots, x_{ej}\}$, and then the elite oppositional solution is given by

$$\tilde{x}_{i,j} = \rho * (da_j + db_j) - x_{e,j} \quad (17)$$

Where, $i = [1, 2, \dots, P]$, P is the size of the population, ρ is the generalized coefficient, $[da_j, db_j]$ are dynamic bounds and can be calculated as:

$$da_j = \min(x_{i,j}), \quad db_j = \max(x_{i,j}) \quad (18)$$

In EOBL, dynamic bounds are used instead of fixed bounds to secure the search space from shortening. If $\tilde{x}_{i,j}$

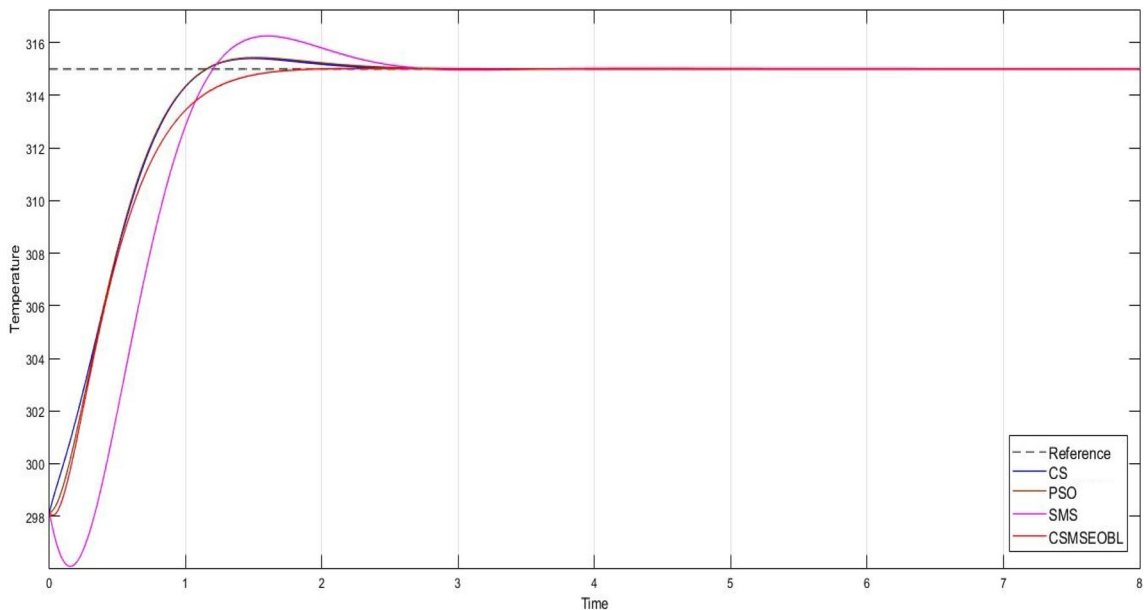


Figure 7. Comparison of Temperature control of the CSTR system among different metaheuristic algorithms.

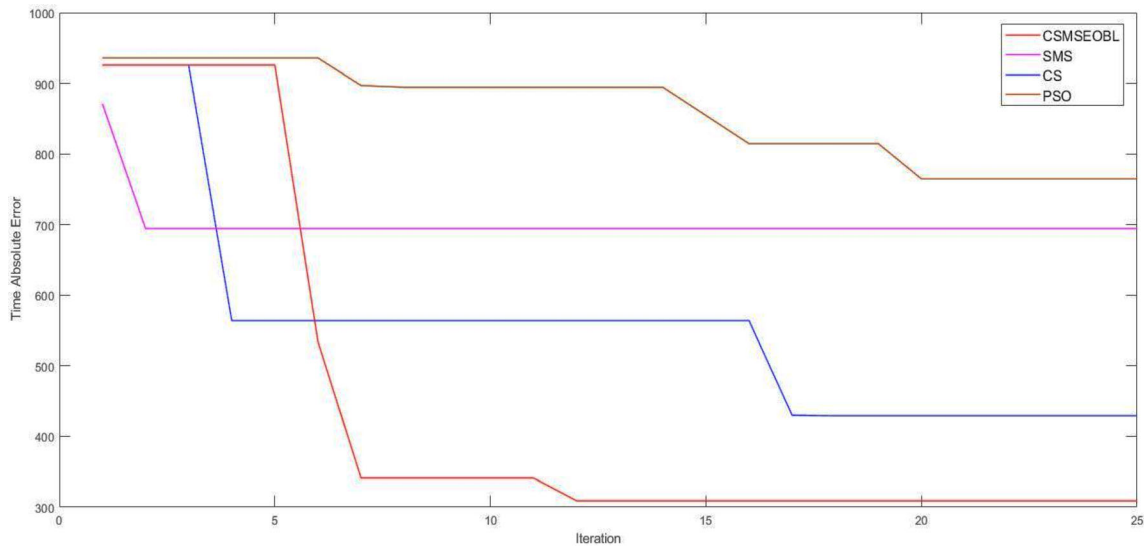


Figure 8. Variation of ITAE for different metaheuristic algorithms.

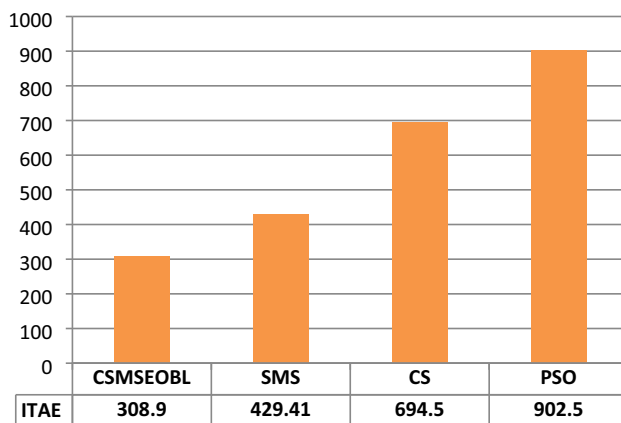


Figure 9. Comparison of objective function ITAE for different metaheuristic algorithms.

crosses its dynamic bound it can be reset by using the following equation:

$$\tilde{x}_{i,j} = \text{rand}(da_j, db_j) \quad (19)$$

The main benefit of EOBL is that the elite population and current population can be evaluated at the same time which further results in a diversified population and enhanced global search ability [48].

Pseudocode for CSMSEOBL

Input: Define fitness function $f(x)$, where, $X = (x_1, x_2, \dots, x_D)$

Output: The optimal solution x^*

- Step 1. Initialize the gas state parameter of the algorithm with the dynamic boundary of the search space.
- Step 2. While stop criterion is not satisfied do
- Step 3. The current population is updated by applying the EOBL strategy.
- Step 4. For each $x \in P$ do
- Step 5. All random variables are updated by employing chaotic maps

Table 4. Comparative analysis of controller parameters and time response specifications.

	FOPID Controller Parameter					Rise time	Peak time	Overshoot	Settling time
	K_p	K_I	K_D	γ	μ				
SMS	12.1	32.5	1	1.006	0.1000	1.22	1.53	.38	2.27
CS	21.7	50	0.2	1.002	0.7850	1.13	1.36	.12	1.86
CSMSEOBL	15.8	43.3	1.9	.9999	0.1386	2.04	1.68	0	1.43
PSO [50]	.2510	.0243	.499	.5968	.0706	3.65	4.76	7	14

- Step 6. For gas, the state calculates initial velocity and collision radius.
- Step 7. The new molecules are computed by using direction vector
- Step 8. By employing a collision operator, solve for the collision.
- Step 9. The new random position is generated by using collision operator
- Step 10. If the total no. of iterations completed $\leq 50\%$ of the total number of iterations
- Step 11. Go to the liquid state and repeat Steps 6, 7, 8, and 9.
Else
- Step 12. Check if the total no. of iterations completed $\leq 90\%$ of the total number of iterations
- Step 13. Go to solid-state and repeat Steps 6, 7, 8, 9
- Step 14. If 100% of the total iterations completed
- Step 15. Update x with x^*
End if
End for
End while

5. Simulation results

To confirm the practicality and viability of the proposed hybrid CSMSEOBL approach, a progression of comparative experiments has been performed on CSTR against the accompanying three states of the art metaheuristic optimization techniques: PSO, CS, SMS, and CSMS-EOBL. MATLAB 2018 is used for simulation and Intel (R) Core (TM) 2 Duo CPU T6400@ 2.00 GHz 1.20 GHz, 1.99 GB of RAM. The performance is verified for Control Temperature and Concentration of CSTR by running CSMSEOBL based FOPID, SMS based FOPID, CS-based FOPID, and PSO based FOPID controller, and results are compared. Parameter setting for all mentioned algorithms have been shown in Table 3.

For any optimization process convergence of metaheuristic algorithm towards the global optima of the tuned parameters of FOPID, the problem is defined with an objective function or fitness function. In this paper, to get the finest transient response as well as minimum steady-state error along with the least overshoot, Integral time absolute error (ITAE) is utilized as the objective functions. Since ITAE the most aggressive controller setting criteria that avoid peaks and give controllers with a greater load disturbance rejection and lessen the overshoot of the system while retaining the robustness of the system. ITAE is defined as

$$J_{ITAE} = \int_0^T t|e(t)|dt$$

The CSMSEOBL is used to optimize the parameters of FOPID for concentration and temperature control of CSTR and also to show the comparative study among Cuckoo Search, State of Matter Search and Particle Swarm Optimization algorithm is also implemented on CSTR. MATLAB Simulink environment is utilized for evaluating the results.

Further, different types of chaotic maps are used to enhance the randomness of the SMS algorithm. Figures 4 and 5 show the variation of concentration and temperature for different types of chaotic maps, respectively.

Further, to prove the superiority of the proposed algorithm, comparative result analysis with the best solution obtained from figures 4 and 5 are done with the existing algorithms i.e., SMS, Cuckoo Search (CS), and particle swarm optimization (PSO) are shown in figures 6 and 7, respectively. The setpoint for Concentration is taken as .119 (lb. mol/ft³) and the temperature is at 315K.

The fitness (or objective) function is optimized by using different metaheuristic algorithms, and the dynamic performance of the CSTR is strengthened by optimizing various performance indices like rise time, settling time, and overshoot using a mathematical formulation of the objective function ITAE (Integral Time Absolute Error). ITAE decreases not only the initial extent of error but also decreases the error which develops in later responses [49]. Variation of ITAE for different metaheuristic algorithms has been shown in figure 8. The comparative analysis of considered metaheuristic algorithms in terms of ITAE is shown in figure 9. The proposed CSMSEOBL algorithm outperformed the other metaheuristic algorithms and it has been shown in table 4.

From table 4 we could conclude that the proposed CSMSEOBL shows the promising approach for concentration and temperature control of CSTR because in process control problems main aim is to obtain the least settling time and minimum overshoot. Even though the rise time and the peak time are large for CSMSEOBL-FOPID as compared to SMS-FOPID, CS-FOPID, and PSO-FOPID but the cost function is minimized along with minimum overshoot and least settling time.

6. Conclusion

This paper fixes the limitations of the standard SMS Algorithm by hybridizing it with chaotic maps and Elite Opposition Based Learning. Further, this hybrid algorithm CSMSEOBL is used to find optimal parameters of the FOPID Controller for the temperature and concentration control of a Continuously Stirred Tank Reactor (CSTR). Major findings of the work are as follows:

- CSMSEOBL gives better exploration and exploitation capability.

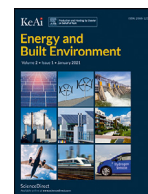
- The use of CSMSEOB on a non-linear control problem results in faster convergence.
- CSMSEOB shows promising results in terms of overshoot, settling time, and ITAE for optimizing the performance.
- The proposed controller is validated for concentration and temperature control of CSTR.

This study of hybrid metaheuristic can be further extended to reform the transient performance using multiple models, adaptive control strategy, and other latest metaheuristic algorithms.

References

- [1] Rajinikanth V and Latha K 2012 Controller parameter optimization for nonlinear systems using enhanced bacteria foraging algorithm. *Applied Computational Intelligence and Soft Computing* 2012: 1–12
- [2] Nasrabadi M S, Sharafi Y and Tayari M 2016 A parallel grey wolf optimizer combined with opposition-based learning. In: *1st Conference on Swarm Intelligence and Evolutionary Computation (CSIEC)*, 18–23
- [3] Sundaravadivu K, Arun B and Saravanan K 2011 Design of fractional order PID controller for liquid level control of the spherical tank. In: *IEEE International Conference on Control System, Computing and Engineering*, 291–295
- [4] Podlubny I 1994 Fractional-order systems and fractional-order controllers. Institute of Experimental Physics, Slovak Academy of Sciences. *Kosice* 12: 1–18
- [5] Nagarajan M and Asokan A 2014 Design and implementation of fractional order controller for CSTR process. *Int. J. Comput. Appl.* 975: 8887
- [6] Yadav P, Kumar R, Panda S K and Chang C S 2012 An intelligent tuned harmony search algorithm for optimization. *Inf. Sci.* 196: 47–72
- [7] Shayanfar H and Gharehchopogh F S 2018 Farmland fertility: a new metaheuristic algorithm for solving continuous optimization problems. *Appl. Soft Comput.* 71: 728–746
- [8] Catalbas M C and Gulten A 2018 Circular structures of pufferfish: a new metaheuristic optimization algorithm. In: *2018 Third International Conference on Electrical and Biomedical Engineering, Clean Energy and Green Computing (EBCEGC) IEEE*, 1–5
- [9] El-Henawy I and Ahmed N 2018 Meta-heuristics algorithms: a survey. *Int. J. Comput. Appl.* 179: 45–54
- [10] Gandomi A H, Yang X S, Talatahari S and Alavi A H 2013 Metaheuristic algorithms in modeling and optimization. *Metaheuristic Applications in Structures and Infrastructures*. 1–24
- [11] Banks A, Vincent J and Anyakoha C 2007 A review of particle swarm optimization. Part I: background and development. *Nat. Comput.* 6: 467–484
- [12] Kennedy J and Eberhart R 1995 Particle swarm optimization. In: *Proceedings of ICNN'95 - International Conference on Neural Networks*, pp. 1942–1948
- [13] Goyal R, Parmar G and Sikander A 2019 A new approach for simplification and control of linear time-invariant systems. *Microsyst. Technol.* 25: 599–607
- [14] Bhatt R, Parmar G, Gupta R and Sikander A 2019 Application of stochastic fractal search in approximation and control of LTI systems. *Microsyst. Technol.* 25: 105–114
- [15] Naidu K, Mokhlis H and Bakar A H A 2013 Application of Firefly Algorithm (FA) based optimization in load frequency control for interconnected reheat thermal power system. In: *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pp. 1–5. IEEE
- [16] Mirjalili S, Mirjalili S M and Lewis A 2014 Grey wolf optimizer. *Adv. Eng. Softw.* 69: 46–61
- [17] Ahmed A M, Rashid T A, and Saeed S A M 2020 Cat swarm optimization algorithm: a survey and performance evaluation. *Computational Intelligence and Neuroscience*
- [18] Rao R 2016 Review of applications of TLBO algorithm and a tutorial for beginners to solve the unconstrained and constrained optimization problems. *Decis. Sci. Lett.* 5: 1–30
- [19] Sahu B K, Pati S, Mohanty P K and Panda S 2015 Teaching–learning-based optimization algorithm based fuzzy-PID controller for automatic generation control of multi-area power system. *Appl. Soft Comput.* 27: 240–249
- [20] Cuevas E, Echavarría A and Ramírez-Ortegón M A 2014 An optimization algorithm inspired by the States of Matter that improves the balance between exploration and exploitation. *Appl. Intell.* 40: 256–272
- [21] Wang G G, Guo L, Gandomi A H, Hao G S and Wang H 2014 Chaotic krill herd algorithm. *Inf. Sci.* 274: 17–34
- [22] Fathollahi-Fard A M, Hajiaghahi-Keshteli M and Tavakkoli-Moghaddam R 2020 Red deer algorithm (RDA): a new nature-inspired meta-heuristic. *Soft Comput.* 1–29
- [23] Shadravan S, Naji H R and Bardsiri V K 2019 The Sailfish Optimizer: a novel nature-inspired metaheuristic algorithm for solving constrained engineering optimization problems. *Eng. Appl. Artif. Intell.* 80: 20–34
- [24] Liu G P, Daley S and Duan G R 2002 Application of optimal-tuning PID control to industrial hydraulic systems. *IFAC Proc.* 35: 179–184
- [25] Yesil E, Guzelkaya M and Eksin I 2003 Fuzzy PID controllers: an overview. In: *Third Triennial ETAI International Conference on Applied Automatic Systems, Skopje, Macedonia*, pp. 105–112
- [26] Ayomoh M K O and Ajala M T 2012 Neural network modeling of a tuned PID controller. *Eur. J. Sci. Res.* 71: 283–297
- [27] Khanduja N and Sharma S 2014 Performance analysis of CSTR using adaptive control. *Int. J. Soft Comput. Eng. (IJSCE)* 4: 80–84
- [28] Khanduja N and Bhushan B 2019 Control system design and performance analysis of PID and IMC controllers for continuous stirred tank reactor (CSTR). *J. Control Instrum.* 10: 16–22
- [29] Devadhas G and Pushpakumar S 2011 An intelligent design of PID controller for a continuous stirred tank reactor. *World Appl. Sci. J.* 14: 698–703
- [30] Huang K, Zhou Y, Wu X and Luo Q 2016 A cuckoo search algorithm with elite opposition-based strategy. *J. Intell. Syst.* 25: 567–593
- [31] Chaudhari Y 2013 Design and implementation of intelligent controller for a continuous stirred tank reactor system using genetic algorithm. *Int. J. Adv. Eng. Technol.* 6: 325

- [32] Goud H and Swarnkar P 2019 Investigations on metaheuristic algorithm for designing adaptive PID controller for continuous stirred tank reactor. *MAPAN* 34: 113–119
- [33] Khanduja N 2015 CSTR control by using model reference adaptive control and PSO. *Int. J. Mech. Mechatron. Eng.* 8: 2144–2149
- [34] Tejado I, Vinagre B M, Traver J E, Prieto-Arranz J and Nuevo-Gallardo C 2019 Back to basics: meaning of the parameters of fractional order PID controllers. *Mathematics* 7: 530
- [35] Podlubny I 1999 Fractional order systems and PI λ D μ -controllers. *IEEE Trans. Autom. Control* 44: 208–214
- [36] Zhang C, Peng T, Li C, Fu W, Xia X. and Xue X 2019 Multiobjective optimization of a fractional-order PID controller for pumped turbine governing system using an improved NSGA-III algorithm under multi-working conditions. *Complexity* 2019: 1–18
- [37] Poovarasan J, Kayalvizhi R and Pongiannan R K 2014 Design of fractional order PID controller for a CSTR process. *Int. Ref. J. Eng. Sci.* 3: 8–14
- [38] Agarwal J, Parmar G, Gupta R and Sikander A 2018 Analysis of grey wolf optimizer based fractional order PID controller in speed control of DC motor. *Microsyst. Technol.* 24: 4997–5006
- [39] Soukkou A, Belhour M C and Leulmi S 2016 Review, design, optimization and stability analysis of fractional-order PID controller. *Int. J. Intell. Syst. Appl.* 8: 73
- [40] Sengupta S, Basak S and Peters R A 2019 Particle swarm optimization: a survey of historical and recent developments with hybridization perspectives. *Mach. Learn. Knowl. Extract.* 1: 157–191
- [41] Sun L, Song X and Chen T 2019 An improved convergence particle swarm optimization algorithm with a random sampling of control parameters. *J. Control Sci. Eng.* 2019: 1–11
- [42] Khanduja N and Bhushan B 2019 CSTR control using IMC-PID, PSO-PID, and hybrid BBO-FF-PID Controller. In: *Applications of Artificial Intelligence techniques in Engineering*, pp. 519–526
- [43] Yang X S and Deb S 2010 Engineering optimization by cuckoo search. *Int. J. Math. Model. Numer. Optim.* 1: 330–343
- [44] Kumar S R and Ganapathy S 2013 Cuckoo search optimization algorithm based load frequency control of interconnected power systems with GDB nonlinearity and SMES units. *Int. J. Eng. Invent.* 2: 23–28
- [45] Tang R, Fong S and Dey N 2018 Metaheuristics and chaos theory. *Chaos Theory*. 182–196
- [46] Tizhoosh H R 2005 Opposition-based learning: a new scheme for machine intelligence. In: *International Conference on Computational Intelligence for Modeling, Control and Automation and International Conference on Intelligent Agents, Web Technologies, and Internet Commerce* 1: 695–701
- [47] Wu X, Zhou Y and Lu Y 2017 Elite opposition-based water wave optimization algorithm for global optimization. *Math. Probl. Eng.* 2017: 1–25
- [48] Ai B, Dong M G and Jang C X 2016 Simple PSO algorithm with opposition-based learning average elite strategy. *Int. J. Hybrid Inf. Technol.* 9: 187–196
- [49] Singh A and Sharma V 2013 Concentration control of CSTR through fractional order PID controller by using soft techniques. Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). 1–6.
- [50] Gupta R 2016 PSO based optimal design of fractional order controller for industrial application. *Int. J. Electr. Comput. Eng.* 9: 665–672



Performance analysis of solar driven combined recompression main compressor intercooling supercritical CO₂ cycle and organic Rankine cycle using low GWP fluids

Yunis Khan*, Radhey Shyam Mishra

Department of Mechanical Engineering, Delhi Technological University, Bawana Road, New Delhi 110042, India

ARTICLE INFO

Keywords:

Performance evaluation
Organic Rankine cycle
Recompression with main intercooler sCO₂ cycle
Low GWP working fluids
Solar power tower

ABSTRACT

Current study deals with performance evaluation of the solar power tower driven recompression with main compressor intercooling (RMCIC) supercritical CO₂ cycle incorporating the parallel double evaporator organic Rankine cycle (PDORC) as bottoming cycle using low global warming potential fluids to reduce the global warming and ozone depletion. Using the PDORC instead of the basic organic Rankine cycle, waste heat from the intercooler and cycle exhaust were recovered simultaneously to enhance performance of the standalone RMCIC cycle. Exergy, thermal efficiency, efficiency improvement and waste recovery ratio were considered as performance parameters. A computer program was made in engineering equation solver to simulate the model. It was concluded that by the incorporation of the PDORC thermal efficiency was improved by 7–8% at reference conditions. Maximum combined cycle's thermal and exergy efficiency were found 54.42% and 80.39% respectively of 0.95 kW/m² of solar irradiation based on R1243zf working fluid. Among the results it was also found that maximum waste heat was recovered by the R1243zf about 54.22 % at 0.95 effectiveness of low temperature recuperator.

1. Introduction

Due to growing concerns about global warming and energy shortages, concentrated solar power (CSP) is a potential alternative to the traditional fossil fuels [1]. The solar power tower (SPT) technology shows great competition among the various CSP technologies available owing to its improved potential for performance enhancement and cost savings correlated with SPT subsystem manufacturing [2]. At the current stage, however, compared to conventional power production technology, SPT technique is not yet cost-effective [3]. The increase in the maximum power cycle temperature contributes to an increase in the efficiency of the power cycle and can reduce the cost of power generation from the SPT [3]. Numerous studies have been conducted to generate high-temperature solar components through the high-temperature power cycle for power generation [4,5]. However, previous studies suggested that as maximum temperature of the cycle increased, the efficiency improvement of the traditional steam Rankine cycle was relatively negligible. In addition, the relative complex structure of the Rankine steam cycle entails a high cost of capital [6]. Therefore, to achieve low costs for SPT electricity production, Superior output power cycles are now being pursued with low capital expenditure [7].

The Brayton supercritical CO₂ (sCO₂) cycle has been considered in recent years to forward-looking energy cycle technologies for power

conversion systems in various energy sectors, particularly SPT plants [4,8]. Earlier studies has shown that with a maximum temperature of 450–800°C, the sCO₂ Brayton cycle demonstrates superior efficiency and Consequently, it satisfies the SPT technology framework criteria [9]. From the viewpoints of cycle enhancement, impact assessment of off-design and growth of control strategy, the viability of incorporating the S-CO₂ Brayton cycle into the SPT systems can be further discussed. In the context of the application of CSP, Dunham and Iverson[6] described different high-performance power cycles and suggested that, based on their comparison of device simulation outcomes, the sCO₂ recompression Brayton cycle has the maximum energy efficiency. Al-Sulaiman and Atif [10] carried out thermodynamic comparisons between the five sCO₂ Brayton cycles integrated in the SPT plant and carried out exergy and energy analyses at six different locations in Saudi Arabia [11] at SPT operated recompression cycles. Wang et al. [12,13] examined various cycle configurations of the sCO₂ Brayton cycle as a power block in the SPT plants from of the perspectives of cycle efficiency, basic work, and incorporation ability with thermal storage. The authors indicated that among different configuration choices, the recompression cycle with primary compression intercooling and partial cooling cycle layouts would be most prevalent in the context of large compressor inlet temperature. The impact on the SPT plant production, which was integrated under the sCO₂ Brayton cycle, was investigated by Osorio et al. [14] under various changes in the weather of multi-tank thermal storage, regenerative effectiveness, and solar receiver conductance. For an air cooled SPT system that uses the remaining heat to drive an absorption chiller as a cooler

* Corresponding author.

E-mail address: yuniskhan21@gmail.com (Y. Khan).

<https://doi.org/10.1016/j.enbenv.2021.05.004>

Received 10 January 2021; Received in revised form 11 May 2021; Accepted 12 May 2021

Available online xxx

2666-1233/Copyright © 2021 Southwest Jiatong University. Publishing services by Elsevier B.V. on behalf of KeAi Communication Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Please cite this article as: Y. Khan and R.S. Mishra, Performance analysis of solar driven combined recompression main compressor intercooling supercritical CO₂ cycle and organic Rankine cycle using low GWP fluids, Energy and Built Environment, <https://doi.org/10.1016/j.enbenv.2021.05.004>

after the cold end of the sCO₂ Brayton cycle. Ma et al. [15] suggested a novel power cycle concept; In the proposed cycle they registered greater thermodynamic and economic efficiency than in the stand-alone sCO₂ Brayton cycle. The design and partial load models for the sCO₂ cycles of Brayton were designed by Dyreby et al. [16] to investigate the inlet temperature effects. By introducing a comprehensive sub-model of a primary heat exchanger, Tse and Neises [17] further updated the models of Dyreby and, based on off-design modeling, studied the yearly efficiency of a sCO₂ Brayton cycle. Calle et al. [18] developed sCO₂ cycle design and off-design models and conducted an optimized cycle design for a concentrated solar power plant on the basis of cycle efficiency under different environmental conditions. In a solar thermal power plant, Singh et al. [19] have developed direct-heated sCO₂ Brayton dynamic mathematical model and examined the behavior of the dynamic cycle in various temperature inputs and environments.. Dynamic models were developed by Luu et al. [20] and the versatile operation of the sCO₂ Brayton cycle incorporated into the direct-heated SPT system was examined; There after two control methods were introduced and matched for temperature control of the turbine inlet. In addition, two control schemes were developed by the authors [21], namely inventory control and flexible recompressor control; and operational changing between the two schemes was proposed to ensure process consistency and efficiency. Iverson et al. [22] modeled the transient sCO₂ Brayton cycle process incorporated into SPT systems and matched the findings with experimental evidence, indicating that solar source disruption seems to be manageable, especially for short period.

Apart from this solar integrated combined cycle studies have been already performed such as Khan and Mishra [23] performed a solar parabolic trough collectors driven combined partial heating sCO₂ and organic Rankine cycle (ORC) for recovering waste heat. They considered six working fluids for bottoming cycle such as R1233zd(E), R1224yd(Z), R1234ze(Z), R1234yf, R1243zf and R1234ze(E). They investigated that after integration of the ORC the standalone partial heating sCO₂ cycle's thermal efficiency improved by 4.47% based on R1233zd(E). In another study Khan and Mishra [24] also performed a combined study of solar power tower driven pre-compression sCO₂ cycle and the ORC for recovering waste heat. They considered five working fluids for the analysis such as isopentane, R236fa, R245fa, isobutene and R227ea. They concluded that by integration of ORC as bottoming cycle thermal efficiency and maximum power output of the standalone cycle improved by 4.52 and 4.51% respectively based on the R227ea working fluids. Singh and Mishra [25] also carried out the a combined parabolic trough collectors driven study of the simple recuperated sCO₂ cycle and ORC for recovering the waste heat considering the R134a, R407C, R1234yf, R1234ze(E) and R245fa. Result of this study revealed that the highest exergy performance value of the R407c combined cycle is around 78.07 percent, followed by the R1234Zes, R1234YF and R245fa with 950 W/m² of solar radiation. In another study Considering R123, R290, R1234yf, R1234ze(E), Toluene, Cyclohexane, isobutane, and Isopentane as working fluids in the bottoming ORC for the recovery of waste heat, Singh and Mishra [26] conducted a combined solar parabolic trough collector combined sCO₂ recompression cycle and ORC as the bottoming cycle. They discovered that R-SCO₂-ORC based on R123 exhibits the highest thermal and energy efficiency: ~73.4 and 40.89 percent of solar irradiation at 0.5 kW/m².

If come across to the recompression with main compressor intercooling (RMCIC) cycle, it was investigated the integration of the main compressor intercooling (MCIC) to simple recompression cycle improved the thermal efficiency by 2.68% at reference conditions [27]. The major difference between IC and RC is that the main intercooling compression process is divided into two phases completed by the main compressor and the pre-compressor, and an intercooler is introduced between the two compressors. The highest efficiency is achieved by IC in combina-

Nomenclature

A_h	single heliostat area (m ²)
C_p	constant pressure specific heat (kJ/kg-K)
$\dot{E}D$	exergy destruction rate (kW)
\dot{E}_{solar}	solar exergy (kW)
\dot{m}	mass flow rate (kg/s)
f_{view}	receiver view factor
\dot{Q}_r	heat received by central receiver (kW)
T	temperature (°C)
G_b	solar irradiation (W/m ²)
s	specific entropy (kJ/kg-K)
h_{conv}	coefficient convective heat loss (W/ m ² -K)
h	specific enthalpy (kJ/kg)
N_h	heliostats number
\dot{E}	rate of exergy (kW)
\dot{Q}	heat rate in (kW)
η_{th}	thermal efficiency
\dot{Q}_h	actual solar heat received by heliostat field (kW)
η_{ex}	exergy efficiency
\dot{Q}_{solar}	solar heat received by heliostat field (kW)
$\dot{Q}_{loss,r}$	heat loss from the receiver (kW)
sCO ₂	supercritical carbon dioxide
T_R	surface temperature of receiver (K)
η_h	heliostat efficiency
\dot{W}	power (kW)
η_r	receiver thermal efficiency
COND	condenser
CR	concentration ratio
CFC	chlorofluorocarbon
GWP	global warming potential
HTR	high temperature recuperator
HEX1	heat exchanger 1
HEX2	heat exchanger 2
HFC	hydro fluoro carbon
HFO	hydro fluoro olefins
IC	intercooler
LTR	low temperature recuperator
MC1	main compressor-1
MC2	main compressor-2
OT1	ORC turbine-1
OT2	ORC turbine-2
ORC	Organic Rankine cycle
PDORC	parallel double evaporator organic Rankine cycle
P1	pump-1
P2	pump-2
RMCIC	recompression with main compressor intercooling
RC	recompressor
MCIC	main compressor intercooling
WHRR	waste heat recovery ratio
SPT	solar power tower

Greek letters

η	efficiency
α	solar absorbance
σ	Stephen Boltzmann constant (W/m ²)
ϵ	effectiveness
β	Sun's subtended cone half angle(rad)
ζ	thermal emittance
δ	change in property

Subscripts

e	exit
i	inlet

0	environmental conditions
r	receiver
h	heliostat
c	critical
b	boiling

tion with the molten salt system SPT [28]. This is the reason to choose the RMCIC for the analysis in the current study. Among other technologies, ORC is the latest technology for the recovery of waste heat from different topping cycles. The parallel double evaporator ORC (PDORC) is a newly discovered technology. PDORC is suitable for the recovery of more waste heat compared to basic ORC. PDORC has produced more work output than the basic ORC [29]. The reason for choosing PDORC in this study is that, due to the use of two evaporators, also waste heat from the intercooler has been recovered to improve the thermal performance of the standalone RMCIC cycle.

From the literature survey, it was found that recompression sCO_2 cycle's thermal performance was improved with MCIC. It was found that thermal performance of the standalone RMCIC cycle can be further enhanced by the incorporating the ORC as bottoming cycle for recovering waste heat. From the literature survey it was also investigated that in the RMCIC cycle waste heat also available in the intercooler. Therefore to recover the waste heat completely, PDORC system was used for recovering waste heat available before first main compressor and at the intercooler simultaneously in present study. Also this combined system was derived by solar power tower. This demonstrates the novelty of the present work. In present study parametric analysis of SPT driven combined RMCIC cycle and PDORC cycle has been carried out. First thermal efficiency of the standalone RMCIC cycle and combined cycle was compared. Later on parametric analysis of the combined cycle has been carried out. Thermal, exergy efficiency, thermal efficiency improvement and waste heat recovery ratio were considered as performance parameters. However, solar irradiations, split ratio, maximum cycle pressure and temperature, inlet temperature of the main compressors and effectiveness of the LTR were considered as independent variables. Design and analysis of the solar power tower is out of scope of the current study.

2. System description

The current model consist a solar power tower cycle which makes molten salt circuit and a recompression with main compression inter-cooling sCO_2 cycle and parallel double evaporator organic Rankine cycle as bottoming cycle for recovering waste heat from heat exchanger-2 (HEX2) and intercooler (IC) as displays in Fig. 1. The sCO_2 stream takes heat from the molten salt (heat transfer fluid) (HTF) through the heat exchanger-1 (HEX1) and expanded in the turbine (process 7–8). After that it passes through the high temperature recuperator (HTR) (process 8–9) where heat is recuperated by the stream of sCO_2 coming from the recompressor (RC). After this it goes to the low temperature recuperator (LTR) (process 9–10) where remaining heat is recuperated by low temperature stream coming from the main compressor 2(MC2). At the state 10 some fraction of the sCO_2 stream split to the recompressor where it was recompressed. Still low temperature heat is remaining, this remaining waste is utilized to drive the bottoming ORC through the heat exchanger 2 (HEX2) (process 10–1). The sCO_2 stream compressed through the main compressor 1(MC1) (process 1–2) and after first compression it intercooled through intercooler and again it compressed through the MC2 (process 3–4). Then it takes heat through the LTR (process 4–5).

Now come across the bottoming PDORC, after recovering the heat through the HEX2, the ORC working fluid expanded through the ORC turbine 1(OT1) (process 11–12) it mixes with the stream coming through the IC and then it again expanded through the ORC turbine 2(OT2) (process 13–14). Then through the condenser working fluid is condensed

Table 1

Input parameters for simulation of the proposed model.

Geometric and operating parameters for SPT	
Direct normal irradiation	0.4–0.95 kW/m^2 [25]
Sun temperature	4500 K [46]
Solar's multiple	2.8 [31]
Efficiency of heliostat	58.71 % [47]
Number of heliostat	141 [45]
Heliostat's total mirror area	9.04×7.89 [31]
Initial temperature difference	15 K [31]
Solar receiver's temperature approach	423.15 K [47]
Concentration ratio	900 [47]
Convective heat loss coefficient	$10 \text{ W/m}^2\text{-K}$ [47]
Tower height	74.62 m [45]
Convective heat loss factor	1 [47]
View factor	0.8 [47]
Absorptance	0.95 [47]
Thermal emittance	0.85 [47]
Input parameters for combined cycle	
Maximum cycle pressure	20 MPa [27]
Maximum cycle temperature	650 °C [27,31]
MC1 inlet pressure	6.25 [MPa] [27]
Main compressors inlet temperature	32–38 °C [30]
Turbine isentropic efficiency	0.9 [27]
Compressors isentropic efficiency	0.89 [27]
Heat exchanger effectiveness	0.95 [44]
HTR and LTR effectiveness	0.95 [27,44]
sCO_2 topping cycle mass flow rate	1.5 kg/s
Mass flow rate in bottoming ORC	0.67 kg/s
PDORC turbine inlet pressure	3 MPa [23,24]
PDORC turbine's isentropic efficiency	0.8 [23, 24]
PDORC pump's isentropic efficiency	0.7 [23,24]

(process 14–15). Then it splits two streams one goes to the intercooler through the pump 2 (P2) (process 15–16) to recover the waste heat from the intercooler (process 16–17). After this ORC working fluid mixes again with the stream coming from the OT1. Now another stream of ORC working fluid goes to the HEX2 through the pump 1 (P1) (process 15–18) to recover the waste heat through the HEX2 (process 18–11). This cycle repeats again and again. T-s diagram for the RMCIC cycle and PDORC are shown in Fig. 2a and 2b respectively.

3. Thermodynamic analysis

3.1. Assumptions

Considering the following assumptions to help the simulation, performance analysis of the SPT driven combined cycle was performed; (1) all system components are under conditions of steady state and equilibrium conditions. (2) In each element, pressure and friction loss are ignored. (3) All processes involving thermodynamics are polytropic. (4) Energy is ignored due to the height and velocity of each component (5) Heliostat and the receiver parameters have remained constant and the input data assumed to support the mathematical modeling are listed Table 1. (6) The temperature of the molten salt inlet at HEX1 was 700 °C [30]. (7) Due to thermal losses, the turbine's inlet temperature is 50 °C lower than the molten salt temperature of the turbine's inlets HEX1. (8) Identical inlet temperatures of the two main compressors are assumed to be same [27].

3.2. Thermal modeling of SPT

Thermal modeling equations of the proposed system were developed in this part based on the conservation of exergy and energy equations, taking into consideration of assumptions those are made in above section. Also each component has been treated as control volume.

Direct solar heat incidence upon heliostat field is defined as [24,31];

$$\dot{Q}_{\text{solar}} = G_b \cdot A_h \cdot N_h \quad (1)$$

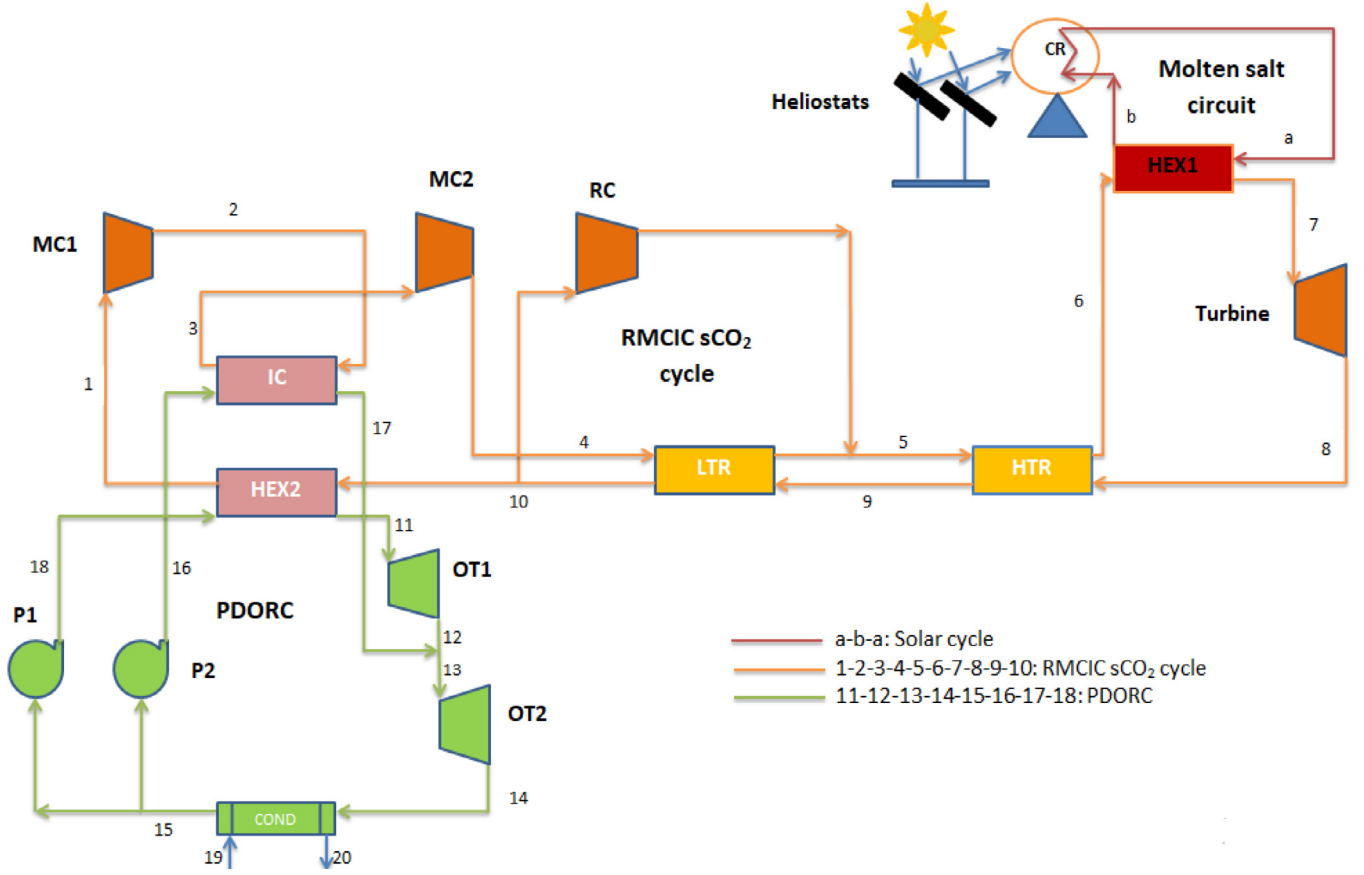


Fig. 1. Schematic diagram of combined recompression with main compressor intercooling and PDORC.

Where, G_b is the solar irradiation per unit area also said to be direct normal irradiation (DNI), A_h is single heliostat area (m^2) and N_h is the heliostats number. However, due to heliostat efficiency, some of that heat is lost in the surroundings. The amount of actual heat obtained through the heliostat field is therefore specified as [24,31];

$$\dot{Q}_h = \dot{Q}_{solar} \cdot \eta_h \quad (2)$$

Where, η_h is the efficiency of heliostat. This amount of heat is directed to the solar receiver where the heat transfer fluid flows. But a part of heat is lost in the atmosphere. The heat available at the solar center receiver is therefore determined as [24,31];

$$\dot{Q}_r = \dot{Q}_h \cdot \eta_r \quad (3)$$

Where, η_r is the receiver thermal efficiency, is defined as [24, 31];

$$\eta_r = \alpha - \frac{\zeta \cdot f_{view} \cdot \sigma \cdot T_R^4 + h_{conv} \cdot f_{conv} \cdot (T_R - T_{air})}{G_b \cdot \eta_h \cdot CR} \quad (4)$$

Where, T_R is the surface temperature of solar receiver and CR is concentrated ratio. ζ is the solar emittance. To calculate heat loss, this can be approximated as [24,31];

$$T_R = T_1 + \delta T_R \quad (5)$$

Where, T_1 is the turbine's inlet temperature and δT_R is approach temperature of solar receiver.

The operating and geometric parameters of the solar receiver and the heliostat field are listed in Table 1.

Furthermore, exergy of the any system can be explained as maximum work obtainable from the system when system is brought to its dead conditions. Control volume exergy balance equation can be determined as [32];

$$\sum \left(1 - \frac{T_0}{T_Q} \right) \dot{Q}_j - \dot{W}_{c.v} - \sum (\dot{m}_i E_i) - \sum (\dot{m}_e E_e) - \dot{E}D = 0 \quad (6)$$

Where, $\dot{E}D$ is the exergy destruction rate and subscript j refers to thermal property at particular state. Solar exergy inlet to the combined system is determined as [24,31];

$$E_{solar} = \left(\frac{\dot{Q}_r}{\eta_r \cdot \eta_r} \right) \cdot E_s \quad (7)$$

Where, E_s is the dimensionless maximum useful work obtained from the solar irradiation. E_s is expressed as [24,31];

$$E_s = 1 + \frac{1}{3} \left(\frac{T_0}{T_{su}} \right)^4 - \frac{4}{3} \left(\frac{T_0}{T_{su}} \right) (1 - \cos \beta)^{1/4} \quad (8)$$

Where, T_{su} and T_0 are the sun and reference temperature respectively. β is the sun's disc subtended half cone angle. Its value has been taken 0.005 rad on solar energy limiting efficiency [33]. Further, in the receiver, useful exergy obtained by the molten salt is defined as

$$\dot{E}_r = \dot{m}_{ms} \cdot C_{p_{ms}} \cdot \left[(T_b - T_a) - \left(T_0 \cdot \ln \frac{T_b}{T_a} \right) \right] \quad (9)$$

Further chemical exergy of the system is constant throughout. After neglecting energy due to velocity and height, specific physical exergy at j^{th} point is defined as [31,34];

$$E_j = (h_j - h_0) - T_0(h_j - s_0) \quad (10)$$

3.3. Thermal modeling for combined cycle

In this section main modeling equations are discussed while detailed modeling equations already discussed in the previous literature such as Ref. [26,27].

Effectiveness approach is considered for calculating heat transfer from the all heat exchanger. Heat received by the combined cycle from the SPT field is given by the heat balance equation in the HEX1 [24];

$$\dot{Q}_r = \dot{Q}_{HEX1} = \dot{m}_{ms} \cdot C_{p_{ms}} \cdot (h_b - h_a) = \dot{m}_{sCO2} \cdot (h_7 - h_6) \quad (11)$$

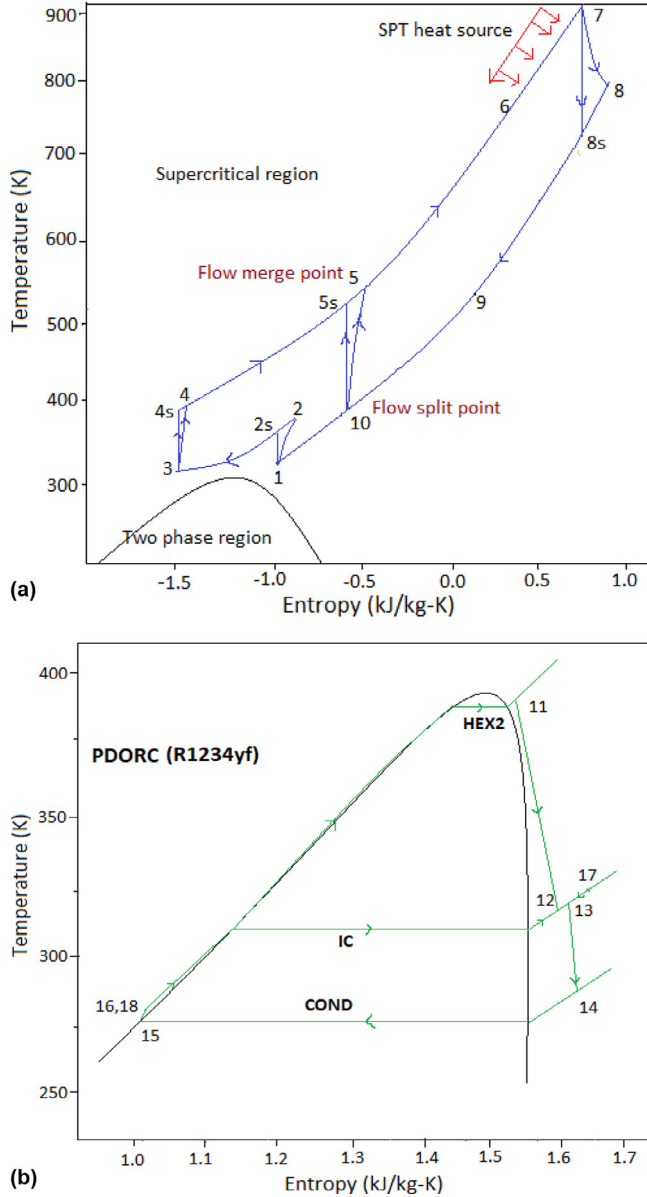


Fig. 2. a. T-s (temperature- entropy) diagram for the RMCIC sCO₂ cycle. 2b.T-s (temperature-entropy) diagram for the PDORC.

Heat transfer in the HTR is determined by effectiveness formula and energy balance equation [24];

$$\dot{Q}_{HTR} = \dot{m}_{sCO_2} \cdot (h_8 - h_9) = \dot{m}_{sCO_2} \cdot (h_6 - h_5) \quad (12)$$

Where, \dot{m}_{sCO_2} is total sCO₂ mass flow rate in topping cycle. Effectiveness of the HTR having same mass flow rate in hot side and cold side of HTR is defined as [24];

$$\varepsilon_{HTR} = \frac{T_8 - T_9}{T_8 - T_5} \quad (13)$$

Heat transfer in LTR is also defined by applying the heat balance and effectiveness formula;

$$\dot{Q}_{LTR} = (1 - x) \cdot \dot{m}_{sCO_2} \cdot (h_5 - h_4) = \dot{m}_{sCO_2} \cdot (h_8 - h_7) \quad (14)$$

Where, 'x' is the fraction of \dot{m}_{sCO_2} goes to main compressors

Effectiveness of the LTR having different mass flow rate in both hot and cold sides of LTR is defined as [24];

$$\varepsilon_{LTR} = \frac{C_{sCO_2} \cdot (T_4 - T_5)}{C_{min} \cdot (T_9 - T_4)} \quad (15)$$

Where, C_{sCO_2} is the heat capacity of sCO₂. C_{min} is the minimum of both hot and cold sides .

Heat absorbed by the PDORC through the HEX2 can be given by energy balance and the effectiveness formula;

$$\dot{Q}_{HEX2} = \dot{m}_{sCO_2} \cdot (h_{10} - h_1) = \dot{m}_{PDORC} \cdot (h_{11} - h_{18}) \quad (16)$$

Where, \dot{m}_{PDORC} is the PDORC's working fluid mass flow rate

Effectiveness of the HEX2 is defined as [24];

$$\varepsilon_{HEX2} = \frac{C_{PDORC} \cdot (T_{11} - T_{18})}{C_{min} \cdot (T_{10} - T_{18})} = \frac{C_{sCO_2} \cdot (T_{10} - T_1)}{C_{min} \cdot (T_{10} - T_{18})} \quad (17)$$

Where, C_{PDORC} and C_{sCO_2} are the heat capacity of PDORC working fluid and sCO₂ respectively. C_{min} is the minimum of these two values.

Similarly, Heat absorbed by the PDORC through the intercooler (IC) as heat exchanger can be given as [24];

$$\dot{Q}_{IC} = x \cdot \dot{m}_{sCO_2} \cdot (h_2 - h_3) = \dot{m}_{PDORC} \cdot (h_{17} - h_{16}) \quad (18)$$

Effectiveness of the IC is defined as;

$$\varepsilon_{IC} = \frac{C_{PDORC} \cdot (T_{17} - T_{16})}{C_{min} \cdot (T_{10} - T_{18})} = \frac{C_{sCO_2} \cdot (T_2 - T_3)}{C_{min} \cdot (T_{10} - T_{18})} \quad (19)$$

Net power output obtained from standalone RMCIC cycle is defined as;

$$\dot{W}_{net \text{ RMCIC}} = \dot{W}_{Turbine} - \dot{W}_{MC1} - \dot{W}_{MC2} - \dot{W}_{RC} \quad (20)$$

Net power obtained by the bottoming PDORC cycle defined as;

$$\dot{W}_{net \text{ PDORC}} = \dot{W}_{OT1} + \dot{W}_{OT2} - \dot{W}_{P1} - \dot{W}_{P2} \quad (21)$$

Therefore, net power obtained by the combined cycle is defined as;

$$\dot{W}_{net \text{ combined}} = \dot{W}_{net \text{ RMCIC}} + \dot{W}_{net \text{ PDORC}} \quad (22)$$

Solar powered standalone RMCIC cycle and combined cycle's thermal efficiency are determined respectively as;

$$\eta_{th \text{ RMCIC}} = \frac{\dot{W}_{net \text{ RMCIC}}}{\dot{Q}_{solar}} \quad (23)$$

$$\eta_{th \text{ combined}} = \frac{\dot{W}_{net \text{ combined}}}{\dot{Q}_{solar}} \quad (24)$$

In addition, the combined system exergy analysis must also be addressed in this section. The destruction of exergy in each component is calculated by applying the Eq. (6) for exergy balance for each component after assuming no loss of heat in the component [32].

After calculating the exergy destruction rate for each component, total exergy destruction rate for the combined cycle is calculated as;

$$\begin{aligned} \dot{E}D_{RMCIC} = & \dot{E}D_{HEX1} + \dot{E}D_{Turbine} + \dot{E}D_{HTR} + \dot{E}D_{LTR} + \dot{E}D_{MC1} \\ & + \dot{E}D_{MC2} + \dot{E}D_{RC} + \dot{E}D_{HEX2} + \dot{E}D_{intercooler} \end{aligned} \quad (25)$$

$$\dot{E}D_{combined} = \dot{E}D_{RMCIC} + \dot{E}D_{OT1} + \dot{E}D_{OT2} + \dot{E}D_{P1} + \dot{E}D_{P2} + \dot{E}D_{COND} \quad (26)$$

On the basis of the thermal modeling, numerous mathematical relations are used in the thermodynamic analysis of the SPT driven combined cycle have been discussed below;

Standalone RMCIC and combined cycle exergy efficiency are determined as [32,34];

$$\eta_{ex \text{ RMCIC}} = 1 - \frac{\dot{E}D_{RMCIC}}{\dot{E}_{solar}} \quad (27)$$

$$\eta_{ex \text{ combined}} = 1 - \frac{\dot{E}D_{combined}}{\dot{E}_{solar}} \quad (28)$$

The combined cycle's thermal efficiency can also be defined by the relation between thermal and exergy efficiency of the combined cycle [32];

$$\eta_{th} = \eta_{ex} \cdot \eta_{Carnot} \quad (29)$$

Table 2

Thermo-physical properties of molten salt (magnesium dichloride + potassium chloride) [42].

Parameters	Values
Density	1593 (kg/m ³)
Specific heat capacity	1.028 (kJ/kg-K)
Thermal conductivity	0.39 (W/m-K)
Solidification temperature	699 K
Stability limit	1691 K

Efficiency (thermal) improvement by incorporating the PDORC as the bottoming cycle can be defined as;

$$\eta_{\text{improvement}} = \frac{\eta_{\text{th, combined}} - \eta_{\text{th, RMCIC}}}{\eta_{\text{th, RMCIC}}} \times 100 \quad (30)$$

It can be also written as;

$$\eta_{\text{improvement}} = \left(\frac{\eta_{\text{th, combined}}}{\eta_{\text{th, RMCIC}}} - 1 \right) \times 100 \quad (31)$$

At last waste heat recovery ratio (WHRR) to be defined which represents the capacity of PDORC for recovering waste heat from the topping cycle. WHRR is defined as the ratio of the net power output (net power output of PDORC) to the maximum available waste heat to be recovered from waste heat source [35]. WHRR for the bottoming cycle is defined as;

$$WHRR = \frac{\dot{W}_{\text{net, ORC}}}{x \cdot \dot{m}_{\text{sCO}_2} \cdot ((h_{10} - h_0) + (h_2 - h_0))} \quad (32)$$

Where, h_0 , h_{10} and h_2 are the enthalpy of waste heat of the topping cycle at environmental temperature and at the inlet of HEX2 and IC respectively. \dot{m}_{sCO_2} is mass flow rate of the sCO₂ flowing in topping cycle. While x is the split ratio, it is fraction of the total sCO₂ mass flow rate going to compress in main compressors. Modeling equations of the SPT powered combined cycle were solved in engineering equation solver (EES) [36].

3.4. Selection of working fluids

Care must be taken when selecting the working fluid for the any thermodynamic cycle because it affects the cycle performance, economic feasibility and environmental aspects [38]. A mixture of magnesium dichloride (MgCl₂) and potassium chloride has been used as molten salt HTF in the receiver with mass fraction of 32% and 68% respectively [31]. Reason behind choosing this HTF is that this is the cheapest option for the heliostat driven sCO₂ cycle as compared to the solar salt and liquid sodium (Na) [37]. Table 2 listed the thermo-physical properties of this molten salt. The choice of working fluid for the ORC is difficult since it destroys its chemical stability beyond its maximum temperature, but at optimum pressure and temperature, it obtains optimum thermo-physical properties [39]. Various parameters, critical point, including global warming potential (GWP), thermal stability and ozone depleting potential (ODP), were analyzed to select suitable fluids for the study. High GWP fluids have been omitted from the analysis, such as hydro fluorocarbons (HFCs) and high ozone depleting potential (ODP) fluids, such as chlorofluorocarbons (CFCs). The ODP was restricted to less than 1. The GWP was restricted to less than 150, as constrained by regulations such as that of the European Union [40]. For the ORC system, working fluids are known as dry, isentropic, and wet fluid. Due to high-quality vapour at the expander outlet, dry and isentropic work is better appropriate than the other form of fluid [23,24]. The waste heat supply also has a low temperature in the current analysis. In current study, considering the above criteria and low temperature applications in present study, ultra-low GWP eight HFO working fluids such as R1234ze(Z), R1224yd(Z), R1225ye(Z), R1233zd(E), R1234yf, R1243zf, R1234ze(E), and R1336mzz(Z) were therefore considered for the PDORC analysis.

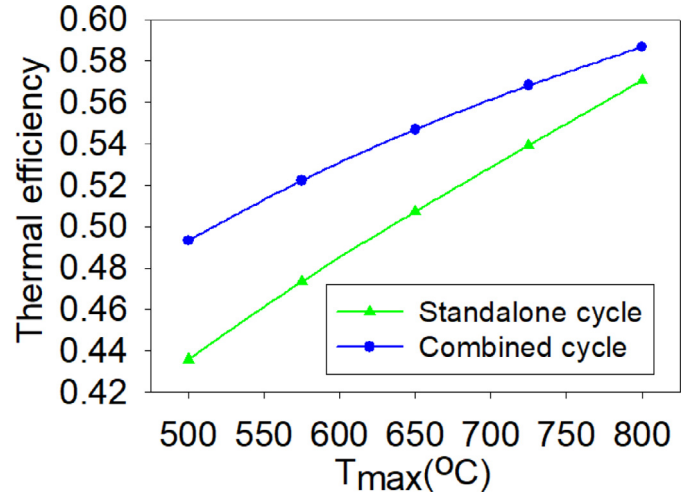


Fig. 3. Thermal efficiency comparison and variation with the maximum cycle temperature.

Table 3 includes the thermal properties, protection and environmental data of these working fluids. For each refrigerant, the protection category designation has two or three numeric values (e.g., B1 or A2L). The first character indicates toxicity and the numeral indicates flammability, with or without a suffix letter. For toxicity, there are two classes: lower toxicity (Class A) and higher toxicity (Class B). There are four flammability classes: 1, 2L, 2 or 3 [41].

3.5. Validation of the proposed model

Mathematical model for the standalone RMCIC and PDORC was validated independently. RMCIC and PDORC models were validated with the previous literature [27] and Ref. [29], respectively, at same operating conditions corresponding to respective literature. The calculated thermal efficiency of the RMCIC and PDORC were given in the Tables 4 and 5, respectively. Calculated thermal efficiencies for the both the cycles were found very close to the literature at the same operating conditions.

4. Results and discussion

All other variables, such as 20 MPa and 650°C maximum pressure and temperature respectively, were kept constant as described in Table 1 during the investigation of the effect of one variable. Thermo-dynamic properties of main stations are given in Table 6 and calculated by EES software.

4.1. Effects of bottoming cycle on the standalone RMCIC cycle

It was found that the thermal efficiency of the combined cycle was improved by 7.8% based on the R1234yf working fluid at the assumed parameters as described in Table 1, by integrating the PDORC into the existing previous study recompression with the main compressor inter-cooling sCO₂ cycle [27]. This also demonstrates the main contribution of the present study as compared to the previous study Ma et al. [27]. With the highest cycle temperature, the thermal efficiency of the standalone cycle and the combined cycle grew. The rate of change in the thermal efficiency of the stand-alone cycle, as shown in Fig. 3, is greater than that of the combined cycle. The thermal efficiency of the standalone cycle and the combined cycle improved by 30.92 and 18.95%, respectively, as the maximum cycle temperature increased from 500 to 800°C.

Table 3

Properties of organic working fluids [23,43,51].

Working substance	P _c (MPa)	T _c (°C)	T _b * (°C)	Weight (Kg/Kmole)	Type	ODP	GWP	Lifetime (years)	Security group
R1234ze(Z)	3.53	150.1	9.8	114.04	Isentropic	0	<10	-	-
R1224yd(Z)	3.33	155.5	14	148.5	Isentropic	0.00023	0.88	-	A1
R1225ye(Z)	3.335	106.5	-20	130.5	Isentropic	0.00012	0.87	-	-
R1233zd(E)	3.57	165.5	18.32	130.5	Isentropic	0.00024	1	-	A1
R1234yf	4.597	94.7	-30	114.04	Isentropic	0	<1	-	A2L
R1243zf	3.518	104.44	-25.41	96.05	Dry	0	<1	-	A2
R1234ze(E)	3.64	109.4	-19.0	114.043	Dry	0	6	0.025	A2L
R1336mzz(Z)	2.903	171.3	33.4	164	Dry	0	8.9	0.0602	A1

Table 4Validation of topping RMCIC sCO₂ cycle.

Operating conditions	Thermal efficiency	Estimated error
Maximum pressure = 20MPa	Ma et al.[27]	Current model
Maximum temperature = 650°C		1.35%
MC1 inlet pressure = 6.25 (MPa)		
MC1 inlet temperature = 35°C		
$\eta_{\text{compressors}} = 0.89$		
$\eta_{\text{turbine}} = 0.9$		
Effectiveness of HTR and LTR = 0.95	50.05%	50.73%

Table 5

Validation of bottoming PDORC.

Operating conditions	Thermal efficiency	Estimated error
Heat source temperature = 110 °C $\eta_{\text{OT}} = 0.82$ $\eta_{\text{pump}} = 0.72$ Working fluid = R245fa	Dai et al.[29]	Current model
	6.37%	6.41%

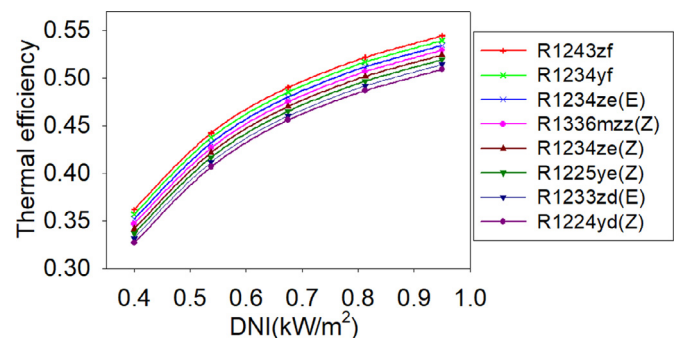
Table 6

Thermodynamic properties of main stations [36].

Main stations	Working fluid	Pressure (MPa)	Temperature (°C)	Enthalpy (kJ/kg)	Entropy(kJ/kg-K)
Main compressor-1 inlet (1)	sCO ₂	6.25	35	-70.01	-0.946
Main compressor -1outlet (2)	sCO ₂	16.82	117.7	-23.11	-0.9327
Intercooler outlet (3)	sCO ₂	16.82	35	-237.2	-1.551
Main compressor-2 outlet (4)	sCO ₂	20	138.55	-3.036	-0.906
HTR inlet (5)	sCO ₂	20	222.1	121	-0.6306
Heat exchanger-1 inlet (6)	sCO ₂	20	488.2	452	-0.09441
Turbine inlet (7)	sCO ₂	20	650	653.3	0.1453
Turbine outlet (8)	sCO ₂	6.25	505	485.8	0.1772
HTR outlet (9)	sCO ₂	6.25	205.7	145.6	-0.378
LTR outlet (10)	sCO ₂	6.25	103	29.62	-0.651
Organic turbine-1 inlet (11)	R1234yf	3	97.98	418.3	1.645
Organic turbine-2 outlet (12)	R1234yf	1	51.76	401	1.652
Organic turbine-2 inlet (13)	R1234yf	1	66.83	417.6	1.702
Organic turbine-2 outlet (14)	R1234yf	0.5	46.76	404.6	1.707
Condenser outlet (15)	R1234yf	0.5	14.33	218	1.066
Pump-2 outlet (16)	R1234yf	1	14.87	220	1.068
Intercooler outlet(ORC side) (17)	R1234yf	1	81.98	434.1	1.749
Pump-1outlet (18)	R1234yf	3	15.55	221.1	-0.09464
Condenser inlet (water side) (19)	water	0.43	5.02	21.54	0.07655
Condenser outlet (water side)(20)	water	0.43	24.04	101.1	0.3534

4.2. Performance evaluation with solar irradiation

Performance of the current model was affected with solar irradiation. Thermal and exergy efficiency of the combined system increased continuously with the solar irradiation keeping constant all other variables as listed in Table 1. As solar irradiation increases the central receiver utilized the solar energy effectively leads to more exergy at the inlet of the combined cycle corresponding less exergy destruction this further leads to the improvement of the exergy as well as thermal efficiency of the combined system. It was found that maximum thermal and exergy efficiency were obtained by R1243zf working fluid among all other selected HFO fluids. However minimum thermal and exergy efficiency were obtained by R1224yd(Z) as shown in Figs. 4 and 5, respectively. As solar irradiation increased from 0.4 to 0.95 kW/m², maximum thermal and

**Fig. 4.** Thermal efficiency variation with the solar irradiation.

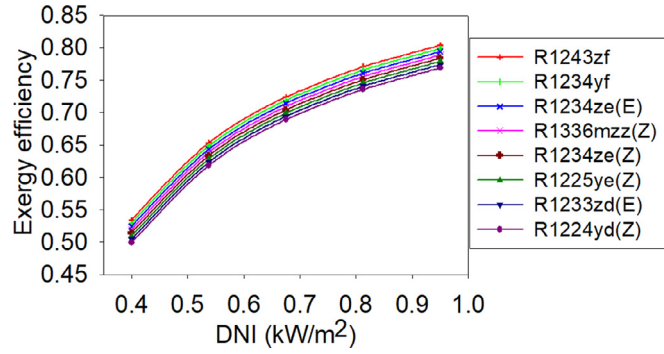


Fig. 5. Exergy efficiency variation with the solar irradiation.

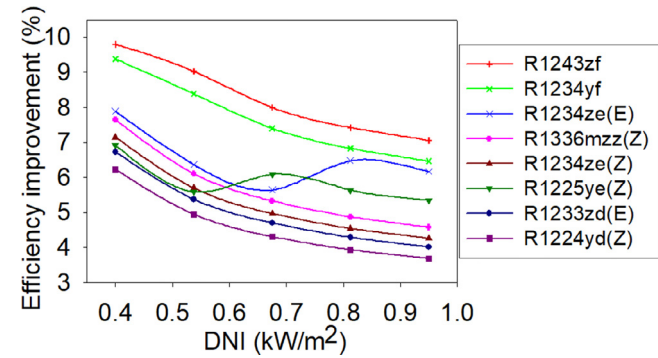


Fig. 6. Efficiency improvement variation with the solar irradiation.

exergy efficiency increased from 36.17 to 54.42% and 53.42 to 80.39%, respectively, based on R1243zf working fluid.

Improvement in thermal efficiency decreases with solar irradiation. As solar radiation increases, the thermal efficiency of the standalone RCMIC increases faster than the combined cycle. As a result, the improvement in thermal efficiency decreased from Eq. (22). Maximum and minimum thermal efficiency improvements were achieved by R1243zf and R1224yd(Z) respectively. Maximum efficiency improvement was achieved by 9.8% at 0.4 kW/m² and decreased to 7.075% at 0.95 kW/m² of solar irradiation as shown in Fig. 6. It can be seen from the Figure 6, the fluids R1234ze(E) and R1225ye(Z), show a different trend compared to other fluids, these fluids show a specific variation in the specific heat [49,50]. As the DNI increases the maximum temperature of the cycle, which affects the specific heat of the working fluids. As a result, thermal performance of the combined cycle has been affected.

4.3. Performance evaluation with the maximum temperature of cycle

Thermal and exergy efficiency of the combined cycle were improved with the maximum temperature of the cycle. Maximum thermal and exergy efficiency of the combined cycle were obtained by the R1243zf working fluid among the other selected HFO working fluids while R1224yd(Z) gave minimum value. Thermal and exergy efficiency for all other working fluids lies between these working fluids. Maximum thermal and exergy efficiency were obtained 61% at the 800 °C based on the R1243zf working fluid as illustrated in the Figs. 7 and 8, respectively. As maximum temperature of the cycle increased from the 500 to 800 °C, thermal and exergy efficiency of the combined cycle increased from 47.78 to 61% and from 70.57 to 90.1%, respectively, based on R1243zf working fluid. It can also be seen from the figure that efficiency variation with different fluids is very close to each other where some working fluids curve overlap as displays in the Figs. 7 and 8. This is due to the close variation of the thermophysical properties of the HFO fluids.

Improvement in the thermal efficiency also depends on the maximum cycle temperature. Improvement in the thermal efficiency de-

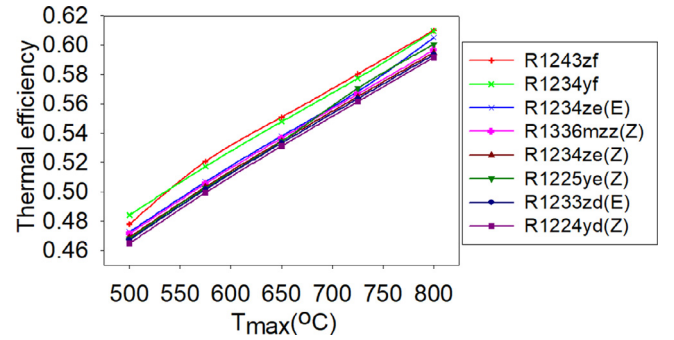


Fig. 7. Thermal efficiency variation with the maximum temperature of cycle.

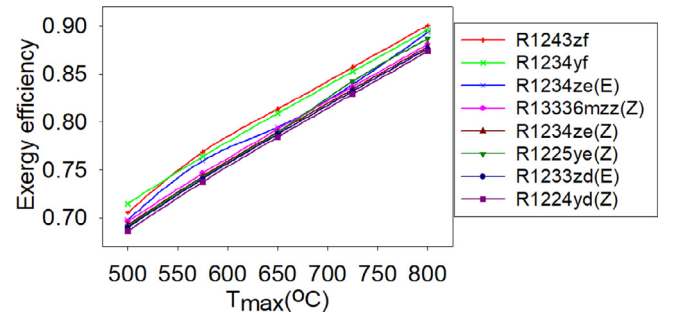


Fig. 8. Exergy efficiency variation with the maximum temperature of cycle.

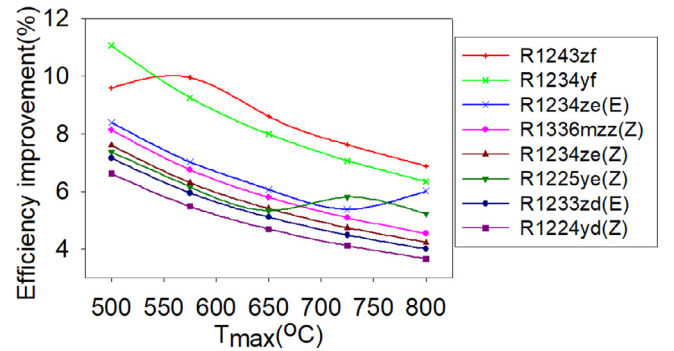


Fig. 9. Efficiency improvement variation with maximum cycle temperature.

creases with the maximum cycle temperature for all working fluids. Because of rate of thermal efficiency improvement with maximum cycle temperature for standalone cycle is greater than the combined cycle. Therefore, efficiency improvement decreased according to the Eq. (31). R1234yf displayed maximum efficiency improvement only for temperature range from 500 to 548°C. However beyond the 548°C, R1243zf showed the maximum thermal efficiency improvement for all temperature range as illustrated in Fig. 9. Efficiency improvement decreased from 9.59 to 6.88% as temperature increased from 500 to 800°C based on the R1243zf. While minimum improvement in the thermal efficiency was obtained with the R1224yd(Z) HFO working fluid.

4.4. Performance evaluation with the maximum cycle pressure

Figs. 10 and 11 displayed that thermal and exergy efficiency were improved with the maximum cycle pressure. This is due to the as pressure increases the enthalpy difference across the turbine increases this leads to the improvement in the turbine work. Consequently, thermal efficiency increased. It was found that among the all selected HFO working fluids R1243zf gave the highest thermal and exergy efficiency while R1224yd(Z) gave lowest. Performance for all other working lie in be-

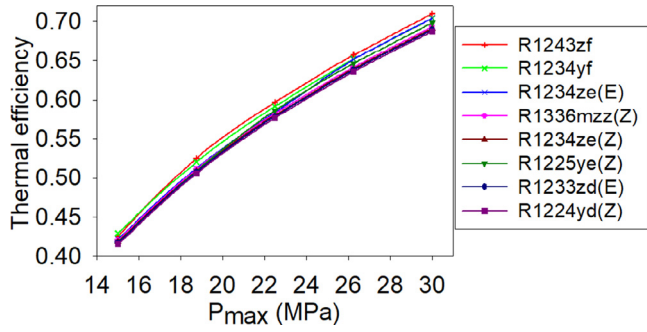


Fig. 10. Thermal efficiency variation with maximum cycle pressure.

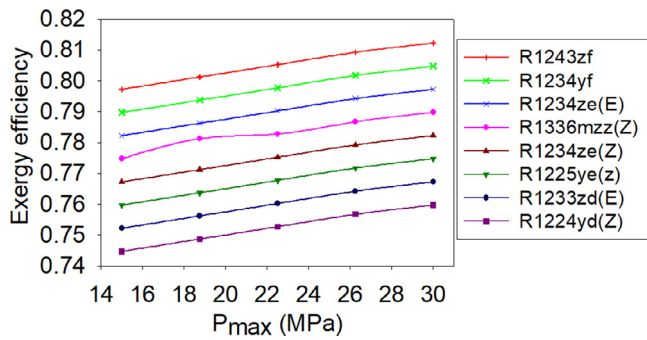


Fig. 11. Exergy efficiency variation with maximum cycle pressure.

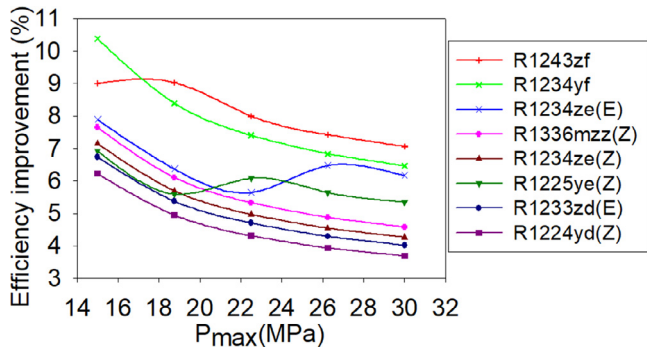


Fig. 12. Efficiency improvement variation with maximum cycle pressure.

tween two fluids. As pressure increased from the 15 to 30 MPa, highest thermal and exergy efficiency improved by 42.6 to 71% and 79.72 to 81.22% respectively. Improvement in the thermal efficiency with the maximum pressure is greater than the exergy efficiency of the combined cycle. This is because of the exergy destruction rate is faster with the maximum pressure of the cycle.

Also thermal efficiency improvement decreases with the maximum cycle pressure. It is known that the thermal efficiency of standalone cycle increases faster than the combined cycle, therefore from the Eq. (31) thermal efficiency improvement decreases. Maximum thermal efficiency improvement was shown by the R1243zf working fluid. Maximum thermal efficiency improvement decreases from the 9 to 7.05% based on the R1243zf. While minimum improvement in thermal efficiency decreases 6.22 to 3.68% based on the R1224yd(Z). Improvement in thermal efficiency for all other fluids lies between these fluids. While the overall maximum improvement in thermal efficiency was found 10.38% for the R1234yf at 15 MPa of maximum cycle pressure as shown in the Fig. 12.

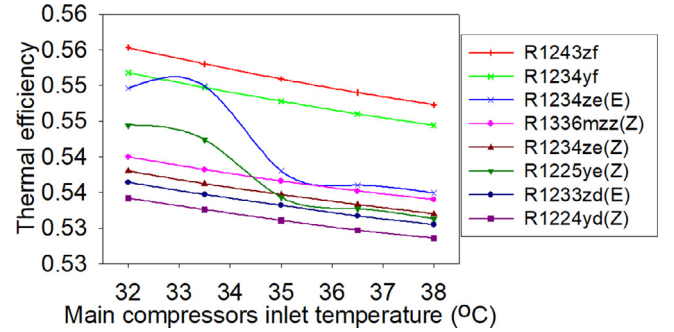


Fig. 13. Thermal efficiency variation with the main compressors inlet temperature.

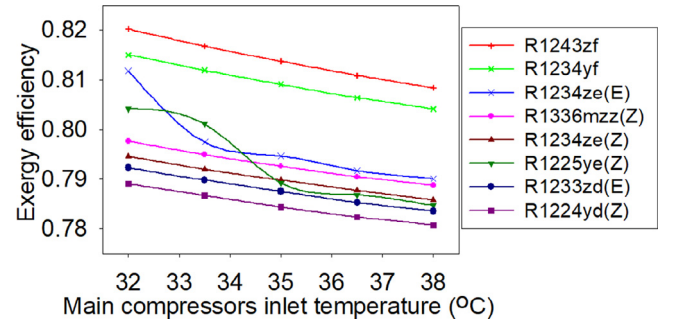


Fig. 14. Exergy efficiency variation with the main compressors inlet temperature.

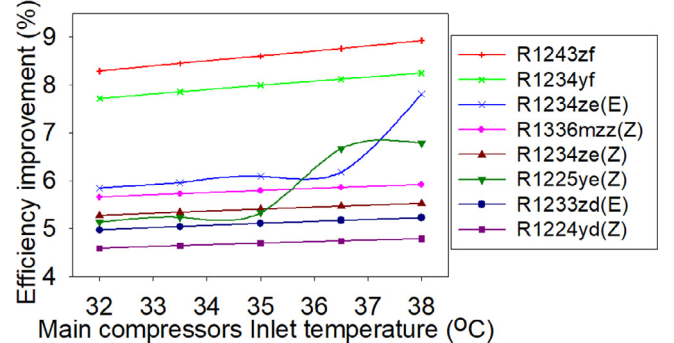


Fig. 15. Thermal efficiency improvement variation with main compressors inlet temperature.

4.5. Performance evaluation with the main compressors inlet temperature

From the Fig. 13 it is observed that thermal and the exergy efficiency of the combined cycle decreases with the main compressor inlet temperature keeping constant all other simulated data as listed in the Table 1. Maximum thermal and exergy efficiency were found for R1243zf fluid and decreased from 55.53 to 54.73% and 82.02 to 80.84% respectively as compressors inlet temperature increased 32–38°C as shown in the Figs. 13 and 14, respectively. However R1224yd(Z) showed the lowest both the efficiencies for all other working fluids thermal and exergy efficiency varies between these two fluids.

Further, it was observed that the thermal efficiency improvement increased slightly with main compressors inlet temperature as shown in the Fig. 15. Reason behind this is that as known than compressors inlet temperature increased thermal efficiency of the standalone recompression with main compressors intercooling cycle decreased sharply. Since performance of the bottoming PDORC varied slightly with the compressors inlet temperature, therefore, combined cycle's thermal efficiency decreased slower than the standalone cycle, therefore, according to the

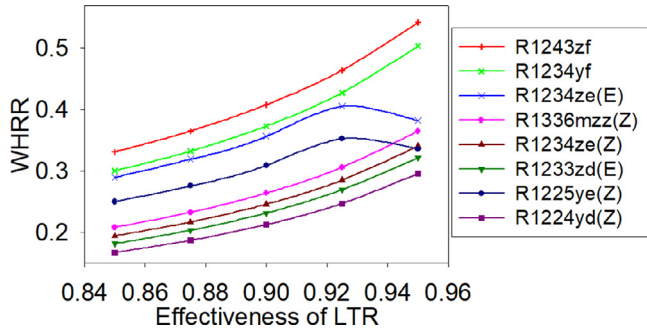


Fig. 16. WHRR variation with the effectiveness of the LTR.

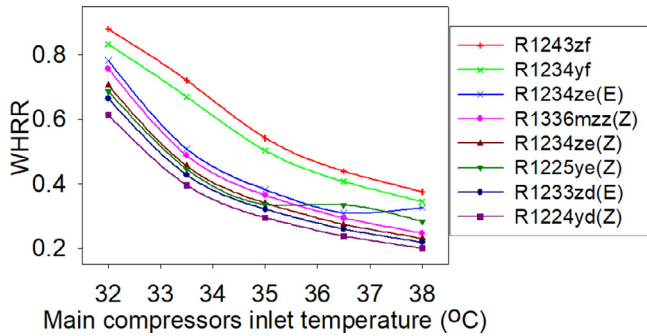


Fig. 17. WHRR variation with main compressors inlet temperature.

Eq. (31) thermal efficiency improvement increased with the main compressors inlet temperature.

4.6. Waste heat recovery ratio variation

This is one of the important parameters to be discussed. Maximum and minimum waste heat recovery ratio (WHR) was obtained by the R1243zf and R1224yd(Z) respectively. WHRR was increased with the LTR effectiveness as shown in the Fig. 16. The reason behind this increased WHRR is that when the effectiveness of the LTR is increased, more heat is recovered by the cold stream of the sCO_2 . This leads to a lower sCO_2 temperature at inlet to HEX2. In other words, the low heat at the inlet to the HEX2 can be said. It reduces the inlet temperature of the ORC turbine. As a result, the lower inlet temperature of the ORC turbine increases the output power of the ORC turbine in case of organic working fluids [48]. Maximum WHRR were found 0.5422, 0.5036, 0.3824, 0.3651, 0.3406, 0.3359, 0.3218 and 0.2958 at 0.95 effectiveness of LTR by the working fluids R1243zf, R1234yf, R1234ze(E), R1336mzz(Z), R1234ze(Z), R1233zd(E), R1225ye(Z) and R1224yd(Z) respectively. Maximum WHRR was found 0.5422 with R1243zf, it means by incorporating the PDORC as bottoming cycle to the standalone cycle RCMIC cycle, 54.22% of total waste heat recovered by the R1243zf. While lowest 29.58% of total waste heat was recovered by the R1224yd(Z). It can be said that R1243zf has been chosen as best fluid working fluid among the all considered fluids.

Apart from this WHRR also depends on the main compressors inlet temperature. WHRR decreased with the main compressors inlet temperature as shown in Fig. 17. Reason behind is that as compressors inlet temperature increased, outlet temperature of the compressors increased, maximum waste heat available increased but rate of output power obtained in the ORC cycle is less than that of the increased in the maximum waste heat available. Therefore from the Eq. (32) WHRR decreased with the compressor inlet temperature. Here also, R1243zf recovered more waste heat than all other considered working fluids. At the given input variables as listed Table 1, Maximum WHRR were found at 32°C of main compressors inlet temperature 0.88, 0.833, 0.7828, 0.7571,

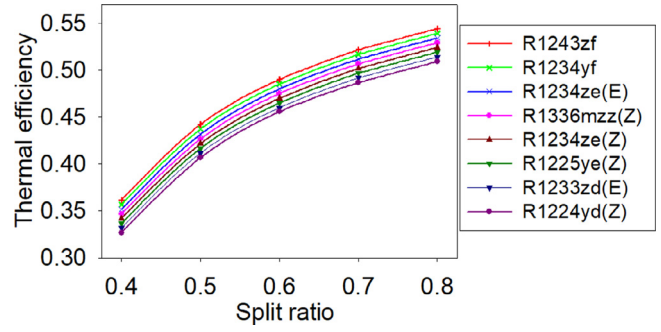


Fig. 18. Thermal efficiency variation with the split ratio.

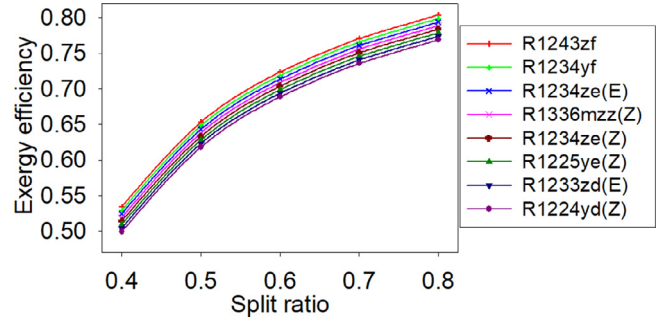


Fig. 19. Exergy efficiency variation with the split ratio.

0.7056, 0.6872, 0.6651 and 0.6141 by R1243zf, R1234yf, R1234ze(E), R1336mzz(Z), R1234ze(Z), R1233zd(E), R1225ye(Z) and R1224yd(Z) respectively.

4.7. Performance evaluation with the split ratio

As mentioned in the modeling section split ratio is the fraction of total mass of sCO_2 which goes to the main compressors. In this section performance of the system were evaluated with the split ratio at given operating conditions such as maximum pressure and temperature of the cycle is 20 MPa and 650 °C, while minimum corresponding values 6.25 MPa and 35 °C, respectively. Thermal and exergy efficiency of the combined system improves with the split ratio as shown in Figs. 18 and 19, respectively. It can be explained as the split ratio increases total heat at the inlet of compressors increased. More heat energy to be converted in to work by PDORC. Therefore, performance of the combined system improved with the split ratio. It was also seen from the Figs. 18 and 19. Slope of the curve is decreasing; it means if further split ratio increases, temperature of the cold stream of the LTR is increased. While the work output from the PDORC increases. Combined effect shows that rate of improvement in both efficiencies with respect to split ratio decreases. Highest thermal and exergy efficiency were obtained 53.99% and 79.91% respectively at the 0.8 split ratio based on R1243zf fluid.

Furthermore in this section, efficiency improvement decreases with the split ratio as can be observed in Fig. 20. It was already explained in previous section that rate of the increment in standalone cycle efficiency is greater than the combined cycle efficiency. Therefore, improvement in efficiency decreases. Highest efficiency improvement decreases from 9.66% to 6.95% as split ratio increases from 0.4 to 0.8 based on the R1243zf. While R1224yd(Z) revealed lowest efficiency improvement among other considered fluids.

5. Conclusions

In the present research, performance analysis of the SPT driven combined recompression with main compression intercooling sCO_2 and organic Rankine cycle has been carried out considering the eight low GWP

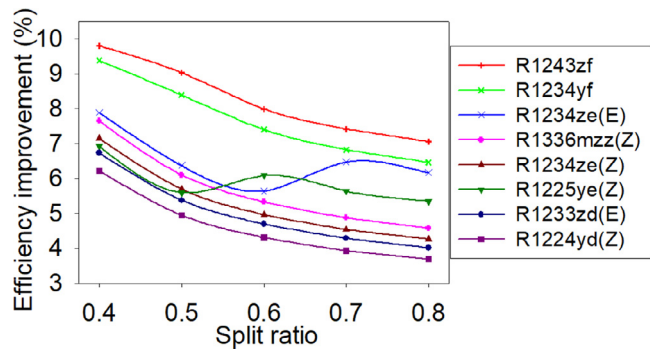


Fig. 20. Efficiency improvement variation with split ratio.

fluids. Following conclusions have been made from the results and discussion section;

- The thermal efficiency of the RMCIC sCO₂ cycle was enhanced by 7–8% by incorporating the PDORC as bottoming cycle.
- The thermal and exergy efficiency of combined cycle were improved with solar irradiation, split ratio, maximum cycle pressure and temperature, however decreased with compressors inlet temperature.
- Maximum exergy and thermal efficiency of the combined cycle were obtained 54.42% and 80.39% respectively at solar irradiation of 0.95 kW/m² of based on R1243zf fluid.
- Cycle thermal efficiency improvement decreased with the solar irradiation, maximum cycle temperature and pressure while increased with main compressors inlet temperature.
- Maximum efficiency improvement was found 9% at 15 MPa of maximum cycle pressure based on R1243zf fluid.
- Maximum WHRR was found to be 0.5422, 0.5036, 0.3824, 0.3651, 0.3406, 0.3359, 0.3218 and 0.2958 at 0.95, R1243zf, R1234yf, R1234ze(E), R1336mzz(Z), R1234ze(Z), R1233zd(E), R1225ye(Z) and R1224yd(Z) respectively.

Acknowledgment

The author (Yunis Khan) acknowledges the kind support of supervisor Prof. R.S. Mishra and Department of Mechanical, Delhi Technological University, New Delhi, India.

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

Yunis Khan: Formal analysis, Validation, Visualization, Writing – original draft, Writing – review & editing. **Radhey Shyam Mishra :** Supervision, Conceptualization, Investigation, Project administration, Resources, Software.

References

- [1] C. Philibert, Technology Roadmap: Concentrating Solar Power, OECD/IEA, 2010.
- [2] O Behar, A Khellaf, K. Mohammadi, A review of studies on central receiver solar thermal power plants, *Renew. Sustain. Energy Rev.* 23 (2013) 12–39.
- [3] M Mehos, C Turchi, J Vidal, M Wagner, Z Ma, C Ho, et al., Concentrating Solar Power Gen3 Demonstration Roadmap, National Renewable Energy Laboratory (NREL, Golden (CO), USA, 2017.
- [4] J Coventry, C Andracka, J Pye, M Blanco, J. Fisher, A review of sodium receiver technologies for central receiver solar power plants, *Sol. Energy* 122 (2015) 749–762.
- [5] CK Ho, BD. Iverson, Review of high-temperature central receiver designs for concentrating solar power, *Renew. Sustain. Energy Rev.* 29 (2014) 835–846.
- [6] MT Dunham, BD. Iverson, High-efficiency thermodynamic power cycles for concentrated solar power systems, *Renew. Sustain. Energy Rev.* 30 (2014) 758–770.
- [7] M Persichilli, A Kludis, E Zdzankiewicz, T. Held, Supercritical CO₂ Power Cycle Developments and Commercialization: Why sCO₂ Can Displace Steam, Power-Gen India & Central Asia, 2012.
- [8] Y Ma, M Liu, J Yan, J. Liu, Thermodynamic study of main compression intercooling effects on supercritical CO₂ recompression Brayton cycle, *Energy* 140 (2017) 746–756.
- [9] SA Wright, TM Conboy, GE. Rochau, in: Overview of Supercritical CO₂ Power Cycle Development at Sandia National Laboratories, University Turbine Systems Research Workshop, Columbus, Ohio, 2011, pp. 25–27. October.
- [10] FA Al-Sulaiman, M. Atif, Performance comparison of different supercritical carbon dioxide Brayton cycles integrated with a solar power tower, *Energy* 82 (2015) 61–71.
- [11] M Atif, FA. Al-Sulaiman, Energy and exergy analyses of solar tower power plant driven supercritical carbon dioxide recompression cycles for six different locations, *Renew. Sustain. Energy Rev.* 68 (2017) 153–167.
- [12] K Wang, M Li, J Guo, P Li, Z. Liu, A systematic comparison of different S-CO₂ Brayton cycle layouts based on multi-objective optimization for applications in solar power tower plants, *Appl. Energy* 212 (2018) 109–121.
- [13] K Wang, Y He, H. Zhu, Integration between supercritical CO₂ Brayton cycles and molten salt solar power towers: a review and a comprehensive comparison of different cycle layouts, *Appl. Energy* 195 (2017) 819–836.
- [14] JD Osorio, R Hovsapien, JC. Ordóñez, Effect of multi-tank thermal energy storage, recuperator effectiveness, and solar receiver conductance on the performance of a concentrated solar supercritical CO₂-based power plant operating under different seasonal conditions, *Energy* 115 (2016) 353–368.
- [15] Y Ma, X Zhang, M Liu, J Yan, J. Liu, Proposal and assessment of a novel supercritical CO₂ Brayton cycle integrated with LiBr absorption chiller for concentrated solar power applications, *Energy* 148 (2018) 839–854.
- [16] JJ Dyreby, SA Klein, GF Nellis, DT. Reindl, Modeling off-design and part-load performance of supercritical carbon dioxide power cycles, in: Proceedings of the ASME Turbo Expo 2013: Turbine Technical Conference and Exposition, American Society of Mechanical Engineers, 2013 V008T34A014–V008T34A014.
- [17] AT Louis, T. Neises, Analysis and optimization for off-design performance of the recompression s-CO₂ cycles for high temperature CSP applications, in: Proceedings of the 5th International Symposium-Supercritical CO₂ Power Cycles, 2016.
- [18] A de la Calle, A Bayon, YC Soo Too, Impact of ambient temperature on supercritical CO₂ recompression Brayton cycle in arid locations: finding the optimal design conditions, *Energy* 153 (2018) 1016–1027.
- [19] R Singh, SA Miller, AS Rowlands, PA. Jacobs, Dynamic characteristics of a direct-heated supercritical carbon-dioxide Brayton cycle in a solar thermal power plant, *Energy* 50 (2013) 194–204.
- [20] MT Luu, D Milani, R McNaughton, A. Abbas, Analysis for flexible operation of supercritical CO₂ Brayton cycle integrated with solar thermal systems, *Energy* 124 (2017) 752–771.
- [21] MT Luu, D Milani, R McNaughton, A. Abbas, Advanced control strategies for dynamic operation of a solar-assisted recompression supercritical CO₂ Brayton power cycle, *Appl. Therm. Eng.* 136 (2018) 682–700.
- [22] BD Iverson, TM Conboy, JJ Pasch, AM. Kruizenga, Supercritical CO₂ Brayton cycles for solar-thermal energy, *Appl. Energy* 111 (2013) 957–970.
- [23] Y. Khan, R.S. Mishra, Parametric (exergy-energy) analysis of parabolic trough solar collector-driven combined partial heating supercritical CO₂ cycle and organic Rankine cycle, *Energy Sources Part A* (2020), doi:10.1080/15567036.2020.1788676.
- [24] Y. Khan, R.S. Mishra, Performance evaluation of solar-based combined pre-compression supercritical CO₂ cycle and organic Rankine cycle, *Int. J. Green Energy* (2020), doi:10.1080/15435075.2020.1847115.
- [25] H. Singh, R.S. Mishra, Performance analysis of solar parabolic trough collectors driven combined supercritical CO₂ and organic Rankine cycle, *Eng. Sci. Technol. Int. J.* 21 (2018) 451–464.
- [26] H. Singh, R.S. Mishra, Energy- and exergy-based performance evaluation of solar powered combined cycle (recompression supercritical carbon dioxide cycle/organic Rankine cycle), *Clean Energy* (2018), doi:10.1093/ce/zky011.
- [27] Y. Ma, M. Liu, J. Yan, J. Liu, Thermodynamic study of main compression intercooling effects on supercritical CO₂ recompression Brayton cycle, *Energy* 140 (2017) 746–756.
- [28] K Wang, Y-L He, H-H. Zhu, Integration between supercritical CO₂ Brayton cycles and molten salt solar power towers: a review and a comprehensive comparison of different cycle layouts, *Appl. Energy* 195 (2017) 819–836.
- [29] Y. Dai, D. Hu, Y. Wu, Y. Gao, Y. Cao, Comparison of a basic organic Rankine cycle and a parallel double-evaporator organic Rankine cycle, *Heat Transf. Eng.* 38 (2017) 990–999, doi:10.1080/01457632.2016.1216938.
- [30] M.A. Reyes-Belmonte, A. Sebastián, M. Romero, Optimization of a recompression supercritical carbon dioxide cycle for an innovative central receiver solar power plant, *Energy* 112 (2016) 17–27.
- [31] S. Khatoun, M. Kim, Performance analysis of carbon dioxide based combined power cycle for concentrating solar power, *Energy Convers. Manag.* (2020) 20, doi:10.1016/j.enconman.2019.112416.
- [32] Y.A. Cengel, M.A. Boles, Thermodynamics An Engineering Approach, 5th edition, McGraw-Hill publication, New York, USA, 2004.
- [33] J.E. Parrott, Theoretical upper limit to the conversion efficiency of solar energy, *Sol. Energy* 21 (1978) 227–229.
- [34] F.A. Al-Sulaiman, Exergy analysis of parabolic trough solar collectors integrated with combined steam and organic Rankine cycles, *Energy Convers. Manag.* 77 (2014) 441–449.

- [35] Y.M. Kim, D.G. Shin, C.G. Kim, G.B. Cho, Single-loop organic Rankine cycles for engine waste heat recovery using both low- and high-temperature heat sources, *Energy* 96 (2016) 482–494.
- [36] Klein, S.A., 2020. Engineering Equation Solver (EES), academic commercial V7.714. F-chart software, www.fChart.com.
- [37] S. Polimeni, M. Binotti, L. Moretti, G. Manzolini, Comparison of sodium and KCl-MgCl₂ as heat transfer fluids in CSP solar tower with sCO₂ power cycles, *Sol. Energy* 162 (2018) 51024.
- [38] X. Wang, E.K. Levy, C. Pan, C. Wang, E. Romero, A. Banarjee, C. Rubio-Maya, L. Pan, Working fluid selection for organic Rankine cycle power generation using hot produced supercritical CO₂ from a geothermal reservoir, *Appl. Therm. Eng.* 149 (2019) 1287–1304.
- [39] Y. Koc, H. Yaglı, A. Koc, Exergy analysis and performance improvement of a subcritical/supercritical Organic Rankine Cycle (ORC) for exhaust gas waste heat recovery in a biogas fuelled Combined Heat and Power (CHP) 520 engine through the use of regeneration, *Energies* 12 (4) (2019) 575, doi:10.3390/en12040575.
- [40] F. Moloney, E. Almatrafi, D.Y. Goswami, Working fluids parametric analysis for the regenerative supercritical organic Rankine cycle for medium geothermal reservoir temperatures, *Energy Proc.* 129 (2017) 599–606.
- [41] J.M. Calm, Refrigerant safety, *ASHRAE J.* 36 (7) (1994) 17–26.
- [42] X. Xu, X. Wang, P. Li, Y. Li, Q. Hao, B. Xiao, Experimental test of properties of KCl–MgCl₂ eutectic molten salt for heat transfer and thermal storage fluid in concentrated solar power systems, *J. Sol. Energy Eng.* 20 (5) (2018) 051011 140.
- [43] N.E. Joaquín, F. Molés, B. Peris, M.B. Adriá, K. Kontomaris, Experimental study of an Organic Rankine Cycle with HFO-1336mzz-Z as a low global warming potential working fluid for micro-scale low temperature applications, *Energy* (2017), doi:10.1016/j.energy.2017.05.092.
- [44] S.M. Besarati, D.Y. Goswami, Analysis of advanced supercritical carbon dioxide power cycles with a bottoming cycle for concentrating solar power applications, *J. Sol. Energy Eng.* 136 (2014) 010904-1-7, doi:10.1115/1.4025700.
- [45] X. Wang, Q. Liu, J. Lei, W. Han, H. Jin, Investigation of thermodynamic performances for two-stage recompression supercritical CO₂ Brayton cycle with high temperature thermal energy storage system, *Energy Convers. Manag.* 165 (2018) 477–487.
- [46] A. Bejan, D.W. Kearney, F. Kreith, Second law analysis and synthesis of solar collector systems, *J. Sol. Energy Eng. Trans. ASME* 103 (1981) 23–28.
- [47] C.K. Ho, B.D. Iverson, Review of high-temperature central receiver designs for concentrating solar power, *Renew. Sustain. Energy Rev.* 29 (2014) 835–846.
- [48] Y. Dai, J. Wang, L. Gao, Parametric optimization and comparative study of organic Rankine cycle (ORC) for low grade waste heat recovery, *Energy Convers. Manag.* 50 (2009) 576–582, doi:10.1016/j.enconman.2008.10.018.
- [49] S. Fukuda, C. Kondou, N. Takata, S. Koyam, Low GWP refrigerants R1234ze(E) and R1234ze(Z) for high temperature heat pumps, *Int. J. Refrig.* (2013), doi:10.1016/j.ijrefrig.2013.10.014.
- [50] L. Fedele, G.D. Nicola, J.S. Brown, L. Colla, S. Bobbo, Saturated pressure measurements of cis-pentafluoroprop-1-ene (r1225ze(z)), *Int. J. Refrig.* (2015) <http://dx.doi.org/doi:10.1016/j.ijrefrig.2015.10.012>.
- [51] Y. Khan, R.S. Mishra, Thermo-economic analysis of the combined solar based pre-compression supercritical CO₂ cycle and organic Rankine cycle using ultra low GWP fluids, *Therm. Sci. Eng. Prog.* 23 (2021) 100925 doi.org/10.1016/j.tsep.2021.100925.

This article was downloaded by: [INFLIBNET India Order]

On: 25 May 2009

Access details: Access Details: [subscription number 909277354]

Publisher Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Polymer-Plastics Technology and Engineering

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title-content=t713925971>

Physicochemical Studies on Interaction Behavior of Potato Starch Filled Low Density Polyethylene Grafted Maleic Anhydride and Low Density Polyethylene Biodegradable Composite Sheets

A. P. Gupta ^a; Vijai Kumar ^b; Manjari Sharma ^a; S. K. Shukla ^a

^a Department of Polymer Science and Applied Chemistry, Delhi College of Engineering, University of Delhi, Delhi, India ^b Central Institute of Plastics Engineering and Technology, Lucknow, Uttar Pradesh, India

Online Publication Date: 01 June 2009

To cite this Article Gupta, A. P., Kumar, Vijai, Sharma, Manjari and Shukla, S. K. (2009) 'Physicochemical Studies on Interaction Behavior of Potato Starch Filled Low Density Polyethylene Grafted Maleic Anhydride and Low Density Polyethylene Biodegradable Composite Sheets', Polymer-Plastics Technology and Engineering, 48:6, 587 — 594

To link to this Article: DOI: 10.1080/03602550902824416

URL: <http://dx.doi.org/10.1080/03602550902824416>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Physicochemical Studies on Interaction Behavior of Potato Starch Filled Low Density Polyethylene Grafted Maleic Anhydride and Low Density Polyethylene Biodegradable Composite Sheets

A. P. Gupta¹, Vijai Kumar², Manjari Sharma¹, and S. K. Shukla¹

¹Department of Polymer Science and Applied Chemistry, Delhi College of Engineering, University of Delhi, Delhi, India

²Central Institute of Plastics Engineering and Technology, Lucknow, Uttar Pradesh, India

Study presents compatibility behavior of polar group potato starch and non-polar LDPE using 50%, 0.5% maleated LDPE. The aim was to improve intermolecular interaction between two different types of moieties (LDPE and Potato Starch). Samples were prepared by mixing potato starch (upto 30% by weight) with LDPE and LDPE-grafted maleic anhydride in a single step twin screw extruder having vent zone. XRD and DSC results suggested that maleic-anhydride group of LDPE helped the interaction with starch and brought two chemical moieties closer to each other. FTIR results also strongly supported new bond formation between two chemical moieties.

Keywords Crystallinity; Dis-integrable polymer; FTIR; Full Width Half Maxima (FWHM); LDPE-g-mA; Low Density Polyethylene-Grafted-Maleic Anhydride; LDPE: Low Density Polyethylene; Scherer's equation; SEM; Thermal analysis; XRD

INTRODUCTION

Starch is inexpensive, renewable, fully biodegradable natural material^[1] and available in abundance in agricultural resources rich country like India. The continuous and exponential increase in the demand of polymer products in daily use and thereafter the difficulties in their disposal in environmentally sound manner, have led to immense interest of research scientists, which are busy in developing polysaccharide-filled polymers for last 20 years.

Polysaccharide-filled polymers have potential to provide a solution to: a range of environmental concerns like decreasing landfill space, declining use of petrochemical resources^[2–8]. The mixing of olefins with organic filler such as starch has been done to develop dis-integrable polymer material with desired combination of properties. The starch

such as corn^[9], wheat^[10], rice^[11] and maize have been successfully added into LDPE.

The dry starch has processing difficulties, because it occurs in the form of discrete and partially crystalline microscopic granules that are held together by an extended micellar network of associated molecules^[12], which makes it difficult to melt or process. But in the presence of plasticizers, such as glycerol^[13–15], glycol^[16] and water etc., the glass transition temperature and melting temperature of the starch are lowered and under high temperature and shear conditions, a deformable thermoplastic material can be achieved. Thus starch is suitable for thermoplastic processing to become an essentially homogeneous material^[17,18]. The compatibilization of incompatible polymer compositions is a major area of research and development. The degree of compatibility is generally related to the level of adhesion between the phases and the ability to transmit stress across the interface.

Starch and polyethylene have less compatibility due to the polar character of hydrophilic starch^[19] and non-polar character of hydrophobic polyethylene, thereby polyethylene and starch are immiscible at molecular level. To bring compatibility in between these two moieties, polyethylene should contain polar functional group that can interact with hydroxyl group of starch.

It is reported that olefins based polymers having functional group such as carboxylic acid, anhydride, epoxy can react with the hydroxyl group of starch to form miscible polymer^[20]. Maleic anhydride (mA) is one of the most widely used vinyl monomers^[21] for graft modification of poly-olefins because highly polar maleic anhydride functional group is more compatible with the polar moieties like starch.

Till date maleic anhydride grafted LDPE, was used as compatibilizer with maximum amount up to 10% by several researchers^[4,22,23]. In the present study maleic

Address correspondence to A. P. Gupta, Department of Polymer Science and Applied Chemistry, Delhi College of Engineering, Bawana, Delhi 110042, India. E-mail: drap_gupta@yahoo.co.in

anhydride grafted LDPE have been used up to 50% in the prepared samples for better compatibility and to overcome processing difficulties of starch, it was plasticized with the help of plasticizer Glycerol. This way, we could manage physical and chemical bond formation between non-polar LDPE and polar potato starch at molecular level.

EXPERIMENTAL

Materials

The native potato starch (10% moisture) was procured from S. D. Fine Chemicals Limited, Mumbai, India. LDPE-g-mA OPTIM E-126, 0.5% grafted was procured from M/s Pluss Polymers, Delhi, India. The low Density Polyethylene, Film Grade was procured from M/s Reliance Industries, India.

Sample Preparation

Potato starch and glycerol were mixed in the ratio of 70:30. Then the suspension was left overnight to allow the swelling action. Afterward, the suspension was rotated in a high speed mixer at 3000 rpm and converted into powder form. The LDPE and LDPE-g-mA was taken in a 1:1 ratio. The prepared composition was used as base material and mixed with already prepared starch suspension in different ratios. Prepared mixtures were manually fed into an industrial standard twin screw extruder (JSW, made in Japan) having screw diameter $d=30$ mm and length to diameter ratio $l/d=36$. The extrusion conditions were as follows.

The temperature profile along the extruder barrel was kept at 100–115–120–125–130–130–130–140°C (from feed zone to die) and the screw speed was 337 rpm. The die was a round shape with six millimeter diameter. The material was oven dried before feeding in the hopper to remove moisture content.

Compression Moulding

The extruded samples were compression moulded into one millimeter thick sheets using a 12×12 cm window frame model in an ELCCN hydraulic pressure machine. The plates of press were heated to $120 \pm 5^\circ\text{C}$. The window was placed between glazed sheets already sprayed with silicon mould release agent on the contact surface and filled with material. The assembly was then placed in a hydraulic press of capacity up to 1600 kg/cm^2 and initially heated for three minutes without applying pressure to ensure uniform heat flow through the material. The temperature was maintained at $120 \pm 5^\circ\text{C}$ for all samples for 15 minutes at a pressure of 1600 kg/cm^2 . The sheet was removed after cooling of the press through a water cooling system. The compositions of prepared samples are given in Table 1.

TABLE 1
Sample compositions

Sample		Raw material		
Number	%	LDPE, %	LDPE-g-mA, %	Potato starch, %
1	0	50	50	0
2	5	47.5	47.5	5
3	10	45	45	10
4	15	42.5	42.5	15
5	20	40	40	20
6	25	37.5	37.5	25
7	30	35	35	30

CHARACTERIZATION OF POLYMER COMPOSITION

Infrared Spectroscopy

IR spectroscopy is a valuable tool for detecting specific structural group in organic and polymeric materials. Transitions between vibrational or rotational states of molecule can be detected by infrared spectroscopy. Fourier Transform IR spectrums of samples using Attenuated Total Reflectance (ATR) technique were obtained in Perkin Elmer spectrometer (Model No. RX-1) in the spectral region in between 4000 and 1500 cm^{-1} .

Thermal Properties (DSC)

Polymer melts over a temperature range due to the difference in size and regularity of the individual crystallites. The melting point of a polymer is generally reported as a single temperature, where the melting of the polymer is complete. Crystallinity is a state of molecular structure referring to a long periodic geometric pattern of atomic spacing. In semi-crystalline polymer such as polyethylene the degree of crystallinity influences the degree of stiffness, hardness and heat resistance. LDPE is a semi-crystalline thermoplastic polymer, which upon the application of heat undergoes a process of fusion or melting, where the crystalline character of the polymer is destroyed. The percent crystallinity was calculated on the assumption that heat of fusion of 100% crystalline LDPE is $276 \text{ J/g}^{[24]}$.

The thermograms of LDPE-g-mA, LDPE and Potato Starch were obtained using Differential Scanning Calorimeter (Model No. Pyris-6, Perkin Elmer Corp., UK). The 5.0 to 8.0 milligram of samples encapsulated in hermetically sealed aluminium pane were prepared for each sample. Samples were heated at $10^\circ\text{C min}^{-1}$ and cooled at 5°C min^{-1} . The thermal transition; such as fusion, was scanned from 25 to 300°C . The thermal transition was calculated from the second heating cycle. The same temperature profile was applied to all samples. Each run was performed under Nitrogen atmosphere. The melting

temperature of the samples was obtained from the maximum peak and area under the peak, respectively.

The degree of crystallinity was calculated via the total enthalpy method, according to Eq. 1^[25].

$$X_c = \frac{\Delta H_m}{\Delta H_m^+} \quad (1)$$

where X_c is the degree of crystallinity, ΔH_m^+ is the specific melting enthalpy for 100% crystalline LDPE. We have taken the value of ΔH_m^+ for LDPE as 276 J/g^[24].

X-Ray Diffraction (XRD)

The X-ray diffraction pattern of LDPE, LDPE-g-mA and Potato Starch compositions were obtained with (Model No. Rigaku RUB-200, Japan) diffractometer operating at 50KV and 60mA using Cu/K α wavelength of 1.542 Å. Scattered radiations were detected at ambient temperature in the angular region of 2θ of 10 to 60° at the rate of 2° per minute.

The degree of crystallinity was estimated from XRD data recorded on the area detector as the ratio of scattering from the crystalline region I_{cr} (sample no. 1, considering it as 100% crystalline component) to the total sample scattering $I_{cr} + I_{am}$ (Amorphous and crystalline regions of different samples) using a simple peak area method. The equation used for the analysis (Eq. 2) is:

$$\alpha_x = \frac{I_{cr}}{I_{cr} + I_{am}} \quad (2)$$

The crystal size has also been determined from the XRD recorded data, by applying the Scherer's equation (Eq. 3) with FWHM as the β is Full Width Half Maxima (FWHM) of the reflection.

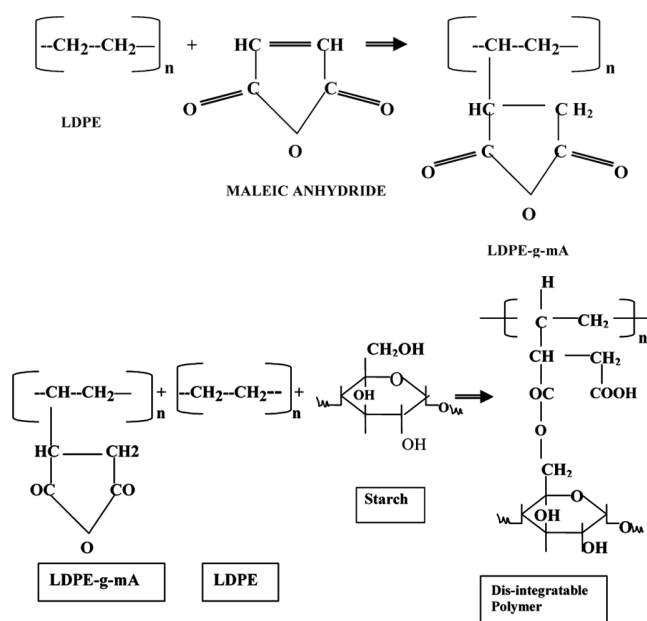
$$D = \frac{0.9 \lambda}{\beta \cos \theta} \quad (3)$$

where D is crystal size in Å, λ is wavelength in Å of X-ray, β is Full Width Half Maxima (FWHM) and θ is diffraction angle.

RESULTS AND DISCUSSION

Infrared Spectroscopy (FTIR:ATR)

FTIR results show that chemical reaction between LDPE-g-mA and starch could take place resulting in ester group formation, which has improved the compatibility in between non-polar LDPE and Polar Starch. This has helped in improving the dispersion of starch and interfacial adhesion in the resultant polymer composition. The steps of plausible scheme for ester formation between the polar group of LDPE-g-mA and hydroxyl group of starch are given below.

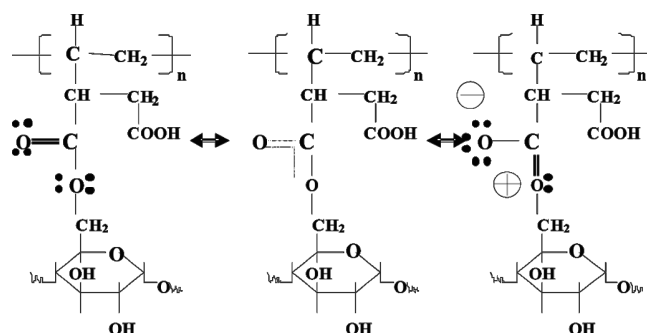


Step-I shows the conversion of LDPE into LDPE-grafted maleic anhydride and Step-II shows the formation of bio-disintegrable polymer with the help of LDPE-g-mA, LDPE and Starch.

Figure 1 depicts FTIR spectrum of LDPE-g-mA, LDPE and Potato Starch compositions, wherein important peaks are summarized in Table 2, along with specific characteristic reasoning.

Spectrum analysis shows a peak around 1720 to 1730 cm^{-1} region. However it is known that the saturated ester shows C=O stretching bend near 1740 cm^{-1} but carbonyl group in conjugation with double bond, such as $\text{CH}=\text{CH}-\text{CO}-\text{O}$ the frequency is lowered near 1720 cm^{-1} ^[26]. The peak observed in this region, in analyzed samples supported the formation of ester group due to the reaction between grafted maleic anhydride and hydroxyl group of starch.

A plausible structure of cumulative double bond formation is given below:



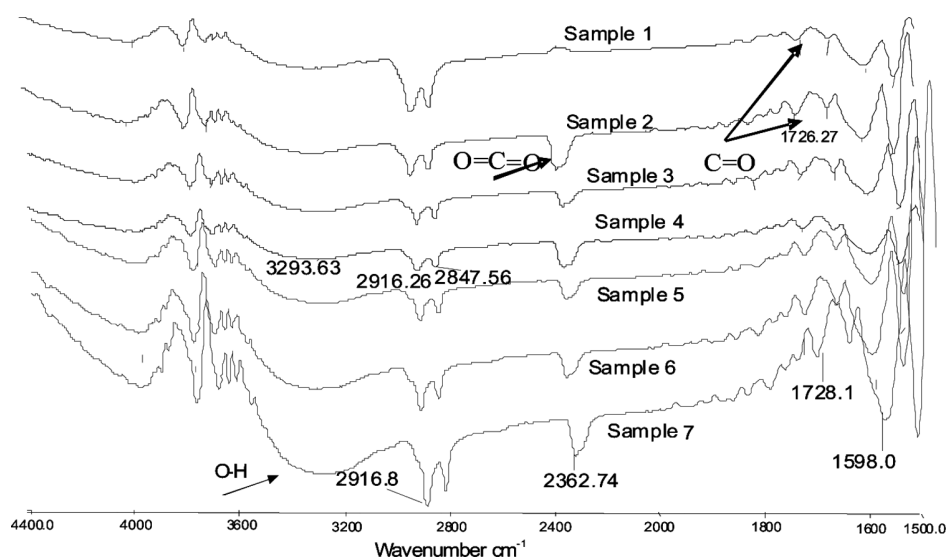


FIG. 1. FTIR Analysis of Samples (1) LDPE-g-mA + LDPE, (2) LDPE-g-mA + LDPE + 5% Starch, (3) LDPE-g-mA + LDPE + 10% Starch, (4) LDPE-g-mA + LDPE + 15% Starch, (5) LDPE-g-mA + LDPE + 20% Starch, (6) LDPE-g-mA + LDPE + 25 Starch and (7) LDPE-g-mA + LDPE + 30% Starch.

A peak is observed around the 2350 cm^{-1} region: The presence of cumulative double bond $\text{O}=\text{C}=\text{O}$ gives strong absorption spectrum peak at 2350 cm^{-1} [26]. Results show that after mixing of starch, cumulative double bond structures are formed, supporting chemical compatibility

between starch and grafted maleic anhydride. Both peaks have strongly supported the new bond formation, which proves that the starch and maleic anhydride of LDPE had chemically reacted. A peak was observed in between 3000 to 2800 cm^{-1} region. The FTIR spectrum

TABLE 2
Major IR absorptions and assignments for LDPE-g-mA/LDPE/potato starch composites

Major IR bands obtained from prepared samples		
Wave number cm^{-1}	Assignments	Remarks
2916(s)	C-H stretching	LDPE Characteristics All samples are showing the C-H characteristics band within this region.
1598(s)	C=C stretching	LDPE Characteristics Alkene characteristics peak present in all samples.
1726 (s)	C=O stretching	Keto group If the carbonyl group in conjugation of double bond $\begin{array}{c} \text{O} \\ \\ \text{CH}=\text{CH}-\text{C}-\text{O} \end{array}$ the frequency is lowered to near 1720 cm^{-1} All the samples show peak in 1722 to 1726 cm^{-1} region which is strong evidence of anhydride group presence in LDPE backbone as a pendant chain.
2350 (s)	O=C=O stretching	Cumulative double bonds are formed after the mixing of starch in LDPE-g-mA which shows that starch has chemically reacted with anhydride group.
3600–3000 strong peak (hump like shape)	O-H Stretching	Hydrogen bonding is getting stronger and stronger as the ratio of starch is increasing in samples.

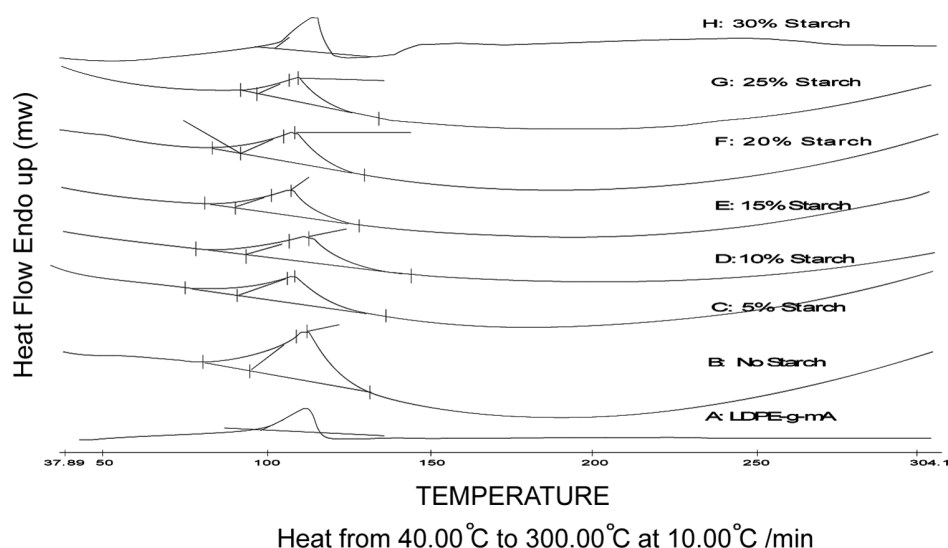
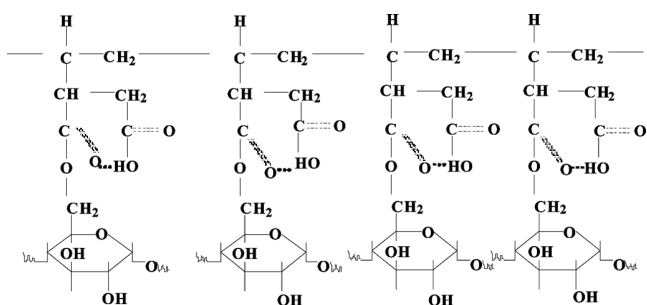


FIG. 2. DSC Curves of (A) LDPE-g-mA, (B) 0% Starch, (C) 5% Starch, (D) 10% Starch, (E) 15% Starch, (F) 20% Starch, (G) 25% Starch, and (H) 30% Starch.

of LDPE/LDPE-g-mA and potato starch in different compositions have a characterized C–H stretching bend.

A peak was observed in between 3600 to 3000 cm^{-1} region. The observed peak is due to the Hydrogen bonding, which is consistently broadening on increment of starch contents in prepared samples. Hydrogen bonding characteristically involves a bond between hydrogen (proton donor) and another group (proton acceptor or electron donor). The polymers exhibiting ability of formation of hydrogen bonds are usually more miscible with wide range of polymers. The broadening of peak observed is due to the formation of strong hydrogen bond, which supports the molecule-molecule interaction between two polar groups.

A plausible structure of hydrogen bond formation is given below:



As a result, it can be said that the compatibility in between LDPE/LDPE-g-mA and Potato Starch has been improved significantly at the molecular level.

Thermal Analysis (DSC)

The DSC results are shown in Table 3. It is clear from the table that, as Potato Starch contents increase, the heat of fusion and crystallinity decrease among prepared samples because heat of fusion is directly proportional to the amount of crystalline LDPE. An apparent decrease in the heat of fusion is due to the decrease in the weight fraction of LDPE in the copolymer because of incorporation of starch and bulky group of maleic anhydride.

Percent crystallinity of LDPE has significantly decreased with the increase of Potato Starch contents among samples. This decrease was attributed to interaction of polar group of the LDPE-g-mA with hydroxyl group of Potato Starch during extrusion process. This interaction has reduced the interfacial tension between Potato Starch and LDPE, so that the nucleus has migrated to the interface and LDPE

TABLE 3

T_m , ΔH , crystallinity analysis of (1) 0% Starch (2) 5% Starch (3) 10% Starch (4) 15% Starch (5) 20% Starch (6) 25% Starch (7) 30% Starch, Composite samples

Sample No.	T_m °C	ΔH J/g	% Crystallinity
1	112.14	79.635	28.64
2	111.50	73.895	26.53
3	112.0	68.280	24.56
4	112.0	53.713	19.50
5	111.50	54.232	20.04
6	111.89	42.649	15.34
7	108.65	41.887	15.00

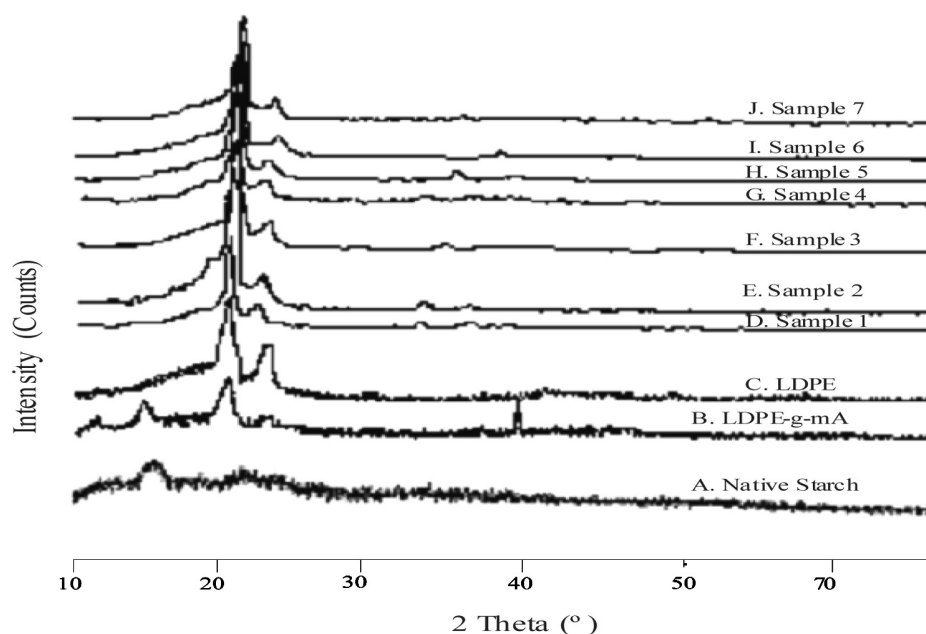


FIG. 3. XRD Curves of (A) Native Starch, (B) LDPE-g-mA, (C) LDPE, (D) 0% Starch, (E) 5% Starch, (F) 10% Starch, (G) 15% Starch, (H) 20% Starch, (I) 25% Starch, and (J) 30% Starch.

crystals have grown on the compatible interface between LDPE and LDPE-g-mA.

X-Ray Diffraction (XRD)

X-ray diffraction of various compositions are presented in Figure 3 (patterns A-J). Native starch being an amorphous material, A-type of crystalline pattern was observed starting with first reflection (2θ) at 16.95° and other broadened reflections (2θ) at 22.18° , 24.06° , 26.11° , 35.57° , pattern is reflected in trace A. The diffraction profiles of raw LDPE-g-mA and LDPE samples exhibited well defined peak predominant (2θ) at 21.6° (trace B and C)^[27]. The typical A-type crystallinity pattern of native starch could not be retained in extruded samples obtained by experimental conditions employed in this work. Traces of B-type crystalline pattern characterized by peaks with angular locations at $2\theta = 21.0^\circ$, 23.34° are detected in different diffractograms (trace D, E, F, G, H, I and J) with increasing starch contents up to 30%. It is worth noting that the former diffraction peaks are coincident with LDPE characteristic reflection. It shows the conversion of native starch leads to loss of natural organization of starch molecules. Therefore amylose and amylopectin molecules crystallize into B-type crystalline structures.

The development of crystalline phase of LDPE in samples is evidenced by the appearance of its characteristics peaks located at $2\theta = 21.0^\circ$, 23.3° approximately. No additional reflection is observed in the diffractograms of the samples. This fact suggests that the LDPE structure do not change appreciably in the presence of starch and

LDPE-g-mA. The microstructural parameters such as d-spacings, crystal size, D, and Degree of crystallinity has been calculated employing X-ray Diffraction patterns which are shown in Table 4.

The results indicate that on account of this incorporation of Potato Starch in LDPE-g-mA and LDPE compositions:

1. The crystal size has been increased due to development of secondary bonds (H-bonds/van der Waal forces) between the anhydride group of LDPE and hydroxyl group of starch.
2. The d-spacing has been observed in XRD showing the reduction, which might be due to the interaction

TABLE 4

d-Spacing, Crystal size, D, and Degree of Crystallinity of LDPE and LDPE grafted maleic anhydride with different potato starch contents in wt% (1) 0 (2) 5 (3) 10 (4) 15 (5) 20 (6) 25 (7) 30, using XRD

Sample No.	d spacing in Å	Crystal size D	Degree of crystallinity
1	4.2445	2.199	1.00
2	4.2219	2.231	0.86525
3	4.2262	2.491	0.79600
4	4.2270	2.474	0.73800
5	4.1939	2.585	0.61649
6	4.1969	2.717	0.62560
7	4.2023	2.926	0.52810

between the hydroxyl group of starch and anhydride group of LDPE.

3. The intermolecular interaction has brought two chemical moieties closer to each other. The results of DSC also supported of secondary bond formation.

Scanning Electron Microscopy (SEM)

SEM micrographs of the prepared samples displayed homogeneous phase between Potato Starch, LDPE-g-MA and LDPE. The SEM micrographs have been presented in earlier paper of the author^[28]. The SEMs have indicated good interfacial adhesion between Potato Starch, LDPE-g-MA and LDPE. This is due to the hydrogen bonding interaction between anhydride group and hydroxyl group of starch. These interactions lead to lowering the interfacial tension between Potato Starch and LDPE phase, leading to better compatibility and miscibility.

CONCLUSIONS

The mixing of 50% LDPE-g-MA in the prepared composition has shown better compatibility as well as miscibility in between LDPE and Potato Starch. This is a better approach to improve the miscibility and compatibility between hydrophobic and non-polar LDPE and hydrophilic and polar Potato Starch. The high end technologies like DSC, FTIR, SEM and XRD available as on date, for the determination of compatibility in between these components, were used and have supported the fact.

LDPE containing a reactive polar group, has significantly improved the interaction at molecular level. The LDPE-g-MA have chemically reacted with hydroxyl group of starch and resulted into ester formation, which developed a strong primary bond in between two moieties and enhanced the compatibility. This reactive group has also generated a strong interfacial secondary bond such as hydrogen bond, because of which two chemical moieties have come closer to each other. Therefore, it is expected that the oxygen of anhydride group, which has been introduced in the LDPE shall also help in degradation of LDPE together with starch.

REFERENCES

1. Rodriguez-Gonzalez, F.J.; Ramsay, B.A.; Favis, B.D. Rheological and thermal properties of thermoplastic starch with high glycerol content. *Carbohydr. Polym.* **2004**, *58*, 139–147.
2. Ahmed, N.T.; Singhal, R.S.; Kulkarni, P.R.; Kale, D.D.; Pal, M. Studies on Chenopodium quinoa and Amaranthus paniculatas starch as biodegradable fillers in LDPE films. *Carbohydr. Polym.* **1996**, *3*, 157–160.
3. Bikiaris, D.; Prinos, J.; Perrier, C.; Panayiotou, C. Thermoanalytical study of the effect of EAA and starch on the thermo-oxidative degradation of LDPE. *Polym. Degrad. Stabil.* **1997**, *57*, 313.
4. Bikiaris, D.; Panayiotou, C. LDPE/Starch blends with PE-g-MA copolymers. *J. Appl. Polym. Sci.* **1998**, *70*, 1503–1520.
5. Chandra, R.; Rustgi, R. Biodegradation of maleated linear low-density polyethylene and starch blends. *Polym. Degrad. Stabil.* **1997**, *56*, 185–202.
6. Psomiadou, E.; Arvanitoyannis, I.; Billiaderis, C.G.; Ogawa, H.; Kawasaki, N. Biodegradable films made from low density polyethylene (LDPE), wheat starch and soluble starch for food packaging applications. Part 2. *Carbohydrate Polymer* **1997**, *33*, 227–242.
7. Utpal, R.V.; Mrinal, B.; Zhang, D. Effect of processing conditions on the dynamic mechanical properties of starch and anhydride functional polymer blends. *Polymer* **1995**, *36*, 1179–1188.
8. Zhihong, Y.; Mrinal, B.; Utpal, R.V. Properties of ternary blends of starch and maleated polymers of styrene and ethylene propylene rubber. *Polymer* **1996**, *37*, 2137–2150.
9. Goheen, S.M.; Wool, R.P. Degradation of polyethylene-starch blends in soil. *J. Appl. Polym. Sci.* **1991**, *42*, 2691–2701.
10. Mota, R.V.; da Lajolo, F.M.; Ciacco, C.; Cordenunsi, B.R. Composition and functional properties of banana flour from different varieties. *Starch/Stärke* **2000**, *52*, 63–68.
11. Arvanitiyannis, I.; Biliaderis, C.G.; Ogawa, H.; Kawasaki, N. Biodegradable films made from low-density polyethylene (LDPE), rice starch and potato starch for food packaging applications: Part 1. *Carbohydr. Polym.* **1998**, *36*, 89–104.
12. Dufresne, A.; Vingnon, M.R. Improvement of starch film performances using cellulose microfibrils. *Macromolecules* **1998**, *31*, 2693–2763.
13. Forssell, P.; Mikkila, J.M.; Moates, G.K.; Parker, R. Phase and glass transition behavior of concentrated barley starch–glycerol–water mixtures, a model for thermoplastic starch. *Carbohydr. Polym.* **1998**, *34*, 275–282.
14. Fishman, M.; Coffin, D.R.; Konstance, R.P.; Onwulata, C.I. Extrusion of pectin/starch blends plasticized with glycerol. *Carbohydr. Polym.* **2000**, *41*, 317–325.
15. Liu, Z.; Yi, X.S.; Feng, Y. Effects of glycerin and glycerol monostearate on performance of thermoplastic starch. *J. Mater. Sci.* **2001**, *36*, 1809–1815.
16. Yu, J.; Gao, J.; Lin, T. Biodegradable thermoplastic starch. *J. Appl. Polym. Sci.* **1996**, *62*, 1491–1494.
17. Suvorova, A.I.; Tyukova, I.S.; Trufanova, E.I. Biodegradable starch-based polymeric materials. *Russ. Chem. Rev.* **2000**, *69*, 451–459.
18. Stepto, R.F.T. The processing of starch as a thermoplastic. *Macromolecules Symposium* 2003, (pp. 201–203).
19. Wang, S.; Jingao, Y.; Jingalin, Y. Preparation and characterization of compatible thermoplastic starch/polyethylene blends. *Polym. Degrad. Stabil.* **2005**, *87*, 395–401.
20. Otey, F.H.; Westhoff, R.P.; Doane, W.M. Starch-based blown films 2. *Ind. Eng. Chem. Prod. Res. Dev.* **1987**, *26* (8), 1659–1663.
21. Villarreal, N.; Pastor, J.M.; Perara, R.; Rosales, S.; Merino, J.C. Use of the Raman-active longitudinal acoustic mode in the characterization of reactively extruded poly-ethylenes. *Macromol. Chem. Phys.* **2002**, *203*, 238–244.
22. Bikiaris, D.; Prinos, J.; Koutsopoulos, K.; Pavlidou, E.; Frangis, N.; Panayiotou, C. LDPE/starch blends containing PE-g-MA copolymer as compatibilizer. *Polym. Degrad. Stabil.* **1998**, *59*, 287–291.
23. Usarat, R.; Duangdao, A.-O. (2006). Preparation and characterization of low-density polyethylene/banana starch films containing compatibilizer and photosensitizer. *J. Appl. Polym. Sci.* **2006**, *100*, 2717–2724.
24. Thakore, I.M.; Desai, S.; Sarawade, B.D.; Devi, S. Studies on biodegradability, morphology and thermo-mechanical properties of LDPE/modified starch blends. *Euro. Polym. J.* **2001**, *37*, 151–160.
25. Khonakdar, H.A.; Jafari, S.H.; Taheri, M.; Wagenknecht, U.; Jehnichen, D.; Häusseler, L. Thermal and wide angle X-ray analysis of chemically and radiation-crosslinked low and high density polyethylenes. *J. Appl. Polym. Sci.* **2006**, *100*, 3264–3271.

26. Kemp, W. *Infrared Spectroscopy; Organic Spectroscopy*, 3rd Ed., New York, Palgrave Publisher Limited: New York, pp. 72–75, 1991.
27. Singh, S.K.; Tambe, S.P.; Samui, A.B.; Kumar, D. Maleic acid grafting on low density polyethylene. *J. Appl. Polym. Sci.* **2004**, *93*, 2802–2807.
28. Gupta, A.P.; Sharma, M.; Kumar, V. Preparation and characterization of potato starch based low density polyethylene/low density polyethylene grafted maleic anhydride biodegradable polymer composite. *Polym. Plastics Technol. Eng.* **2008**, *47* (9), 953–959.



Plasmon assisted tunnelling through silver nanodisk dimer-optical properties and quantum effects

Venus Dillu¹ · Preeti Rani¹ · Yogita Kalra² · Ravindra Kumar Sinha²

Received: 21 August 2020 / Accepted: 11 April 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Tunnelling is a quantum mechanical effect which becomes significant in plasmonic systems with nanogap regions. In a system of closely spaced metal nanoparticles plasmon tunnelling plays an important role in the transfer of energy and hence governs the optical properties of the system. Plasmon assisted tunnelling through a system depends on the skin depth of the material in consideration, which in turn is controlled by the wavelength of incident light. Here, we present, ‘gradient potential dependent skin-depth theory (GPST)’ explaining resonant plasmons assisted tunnelling through metal nanoparticles for the operating wavelength of 1.1 μm . For a system of silver nanodisk dimer with sub-nanometer interparticle distance, the nanogap region between adjacent nanodisks give rise to gradient potential forming the tunnelling zone and is verified by finite difference time domain computational method. The energy eigenvalues and corresponding eigen frequencies are obtained for the dimer system. The proposed GPST can predict the behaviour of plasmon tunnel diode, plasmonic Josephson junction assisted superconductivity, plasmon tunnelled field-effect transistors etc. significantly improving the performance of integrated circuits.

Keywords Gradient potential · Resonant plasmon tunnelling · Skin depth · Nanodisk dimer and energy eigen value

1 Introduction

Plasmonics is fast becoming an imperative technology being used in the development of nanoscale devices (Maier et al. 2001; Huang et al. 2009, 2010; Haes and Duyne 2002). With advancement in nanotechnology, metal nanostructures such as metal-insulator-metal geometry or metal nanoparticles (MNPs) of different size and shapes can be patterned as required (Wang et al. 2011; Stellacci et al. 2002). These metallic nanostructures offer a myriad of permutations for plasmonic devices with remarkable properties (Schaadt et al.

✉ Preeti Rani
preeti1703.soni@yahoo.in

¹ Sharda University, Knowledge Park III, Greater Noida, UP 201310, India

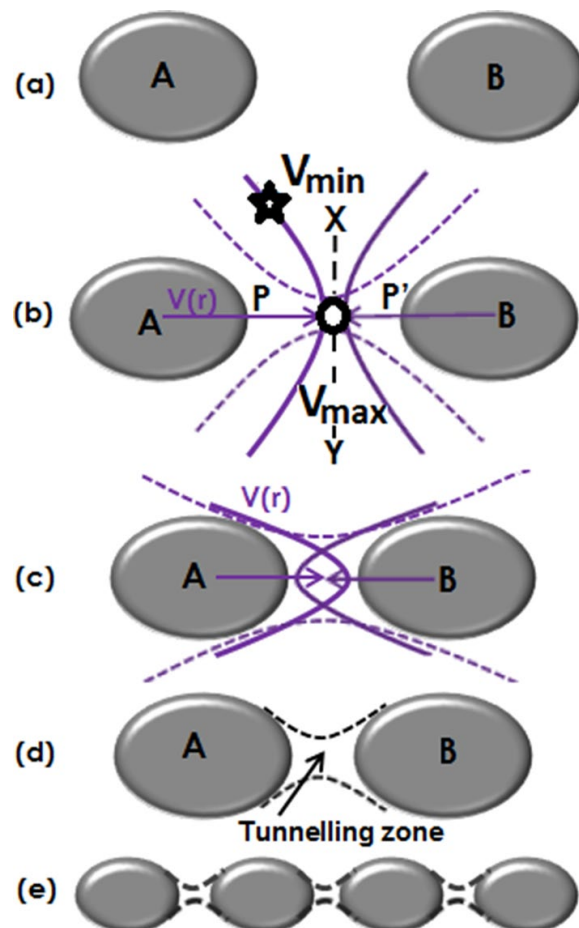
² Department of Applied Physics, TIFAC- Centre of Relevance and Excellence in Fiber Optics and Optical Communication, Delhi Technological University (Formerly Delhi College of Engineering), Bawana Road, Delhi 110042, India

2005; Nagel et al. 2013; Stolz et al. 2014; Chen et al. 2016). Plasmonic systems consisting of metal nanoparticles exhibit transmission characteristics that can be tuned by varying the size, shape, interparticle distance and arrangement of metal nanostructures and it also depends on the plasmon resonance frequency specific to the system. It is observed that resonant plasmon tunnelling occurs in closely spaced metal-insulator-metal systems, consequently contributing to the transmission spectra. Sidorenko and Martin studied resonant tunnelling of surface plasmon polariton across interruption in the metallic film and observed high tunnelling efficiency with greater amplitude transmission (Sidorenko and Martin 2007). Metal-insulator-metal geometry also offers perfect tunnelling junction yielding highly efficient surface plasmon polariton excitation as well as provide control over outcoupling mechanism (Makarenko et al. 2020). Resonant plasmon tunnelling transmits energy and hence causes enhanced forward transmission (Xiao and Mortensen 2008; Park et al. 2008) and it is observed that in periodically corrugated thin metal film it leads to extraordinary transmission even without the presence of holes (Avrutsky et al. 2000). Furthermore, the structures with subwavelength aperture also exhibit enhancement in transmission several orders of magnitude compared to the predicted values and were explained as plasmon enhanced light tunnelling through subwavelength holes (Popov et al. 2000; Martin-Moreno et al. 2001; Liu and Tsai 2002). Besides, the surface plasmon (SP) induced resonant tunnelling can be manipulated using an external static magnetic field through thin structured semiconductor film (Lan et al. 2007) and monitored during the excitation of SP modes on the cavity embedded metal films (Lan et al. 2009) to be used as resonator or waveguide. In addition, plasmons in tunnel-coupled graphene layers structures can act as quantum cascade gain media (Svintsov et al. 2016), and ultrathin plasmonic devices can be possible due to the electrical excitation of plasmons in graphene monolayers sandwiched structures (Vega and Abajo 2017; Abajo et al. 2005). Thus, it is apparent that resonant plasmon tunnelling through plasmonic devices is opening avenues for unique applications such as tunnelling microscopy, two dimensional semiconductors, superconducting circuits etc. (Garg and Kern 2020; Ramazani et al. 2020; Smith et al. 2020). To name a few, Uskov et al. (2016), whereas Liu, Wolf and Kumagai reported plasmon assisted resonant electron tunnelling from a silver/gold tip to field emission resonances of Ag (111) surface induced by a continuous laser excitation of a scanning tunnelling microscope junction at visible wavelength (Liu et al. 2018). Another interesting work was reported by Pshenichnyuk et al. (2019) where they demonstrated edge-plasmon assisted electro-optical modulator. Hot carriers open another exciting research topic in chemical analysis, optoelectronic processes and these hot carriers can be excited with tunnelling electrons. Researchers have designed 10^{11} tunnel junctions per square centimetre containing an array of electrically driven plasmonic nanorods which demonstrate hot electron activation of oxidation and reduction reactions in the junctions (Wang et al. 2018). Generation of hot electrons can make the nanoscale tunnel junctions highly reactive and can hence activate strongly confined chemical reactions which in turn can modulate the tunnelling processes. Xu, Li and Jin reported about negative differential resistance (NDR), a novel phenomenon based on planar plasmonic tunnel junction, which results from plasmon assisted long range electron tunnelling and electron caching effect of Au@SiO₂ nanoparticles and have demonstrated programmable organic-free memristor based on plasmonic tunnel junction (Xu et al. 2020). Hence it is needed that various structures of metal nanostructures should be studied and also the underlying mechanism to set a platform for plasmon assisted tunnelled devices such as plasmon tunnel diode, plasmonic Josephson junction assisted superconductivity, plasmon tunnelled field-effect transistors etc.

2 Gradient potential dependent skin-depth theory (GPST)

In this paper we propose ‘gradient potential dependent skin-depth theory (GPST)’ a novel way to investigate resonant plasmon tunnelling through the dimer system with sub-nanometer interparticle distance. The region between adjacent nanodisks gives rise to gradient potential due to the property of its geometry leading to the formation of tunnelling zone and is substantiated by finite difference time domain (FDTD) computational method. For spatial resolution, the grid size of $0.020\ \mu\text{m}$ in x , y and z directions were taken with a perfectly matching layer for the FDTD simulation with pulse of the type Gaussian excitation. The material silver was chosen from the materials library for the nanodisk dimer. We further obtain the energy and frequency eigenvalues for the dimer system. If the interparticle distance between two metal nanodisks is just few nanometre (see Fig. 1a), the force and hence the potential on bisector XY is maximum at radial centre represented by a circle (Fig. 1b), from both the disks lying at the shortest distance from the boundary points P and P' of the two particles. Therefore, the potential decreases as the distance between the boundary of the particle and that of the bisector increases, with decreasing potential represented by a star symbol in Fig. 1b. So, the geometry of the nanodisks instigate region of varying potential (Fig. 1b) with V_{max} and V_{min} represented as circle and star respectively. As the nanodisks come closer, this region of varying potential between both the particles overlaps (Fig. 1c). This overlapping gives rise to a region of gradient potential between two consecutive nanodisks as shown in Fig. 1d and is marked as a tunnelling zone. The common overlapped domain acts as a tunnelling zone for plasmons, facilitating strong

Fig. 1 Schematic representation of closely spaced silver nanodisks revealing formation of tunnelling zone between the nanodisks. **a** Closely spaced silver nanodisks with subnanometer interparticle distance. **b** Variation of potential between the nanodisks. **c** and **d** The region of high potential overlaps as the disks come closer forming a tunnelling zone which assists resonant plasmon tunnelling. **e** Array of closely spaced nanodisks with tunnelling zone forming a periodic arrangement



plasmon-plasmon interaction for adjacent metal nanodisks in an array of nanoparticles forming a periodic system (Fig. 1e). This periodic system along with the region of gradient potential between the particles acts as a corrugated metal-insulator system and thus skin depth comes into effect.

2.1 Skin depth

Skin effect is an important phenomenon which cannot be ignored when we expect a region of high potential between nanoparticles with sub-nanometre interparticle spacing. It is possible via plasmonics that optical field can be localized tightly if the size of metal nanoparticles is smaller than the skin depth (Stockman 2011). Skin depth which is a wavelength dependent as well as material dependent function plays a pivotal role in resonant tunnelling mechanism since it can control the losses due to retardation effects, transmission, radiation etc. As skin-depth is a function of incident wavelength, we calculate and observe that larger incident wavelength results in increased skin depth (see Fig. 2) and hence increases the total tunnelled intensity at the output end of the system. Figure 2 reveals that as the incident wavelength increases it has a deeper skin effect while propagating through the particles. Therefore, gradient potential present between the silver nanodisks along with wavelength dependent skin depth effect assists resonant plasmon tunnelling through the nanoparticles system.

2.2 Structure

The system under study comprises a chain of silver nanodisks with radius 10 nm and interparticle spacing of 5 nm or less. An important point to be considered in case of the metal nanoparticles is that there is local field enhancement at each individual nanoparticle. With interparticle distance less than a few nanometres these local fields interact, and the resultant output is the synergy of the enhanced local field and the skin depth effect. Also, the size of the nanoparticle plays an important role in tunnelling mechanism, that is, if the particle size is smaller than the skin depth for a particular incident wavelength then the tunnelling efficiency is more and if the particle size is larger than the skin depth then the system exhibit less tunnelling efficiency. Furthermore, the output is expected to be the resultant of in-phase interaction among the tunnelled plasmons (bonding mode) or out-of-phase interaction (antibonding mode).

3 Result and discussion

It is realized from the plot (Fig. 2a) that as the incident wavelength increases from 0.1 to 3.5 μm , the skin depth increases respectively and is between 3 and 4 nm for infrared regime. To comprehend the effect of skin depth on a nanoparticle we take a silver nanodisk of radius 10 nm and study the transmission spectra for the same over a range of input wavelengths from 0.1 to 3.5 μm with transverse magnetic (TM) polarization. The normalized transmission spectra are shown in Fig. 2b and the system exhibits a distinctive configuration. It is perceived that there are few allowed energy bands that are particular to this system and the spectra is illustrated with respect to wavelength in the inset of Fig. 2b. Plot in the inset reveals that the incident wavelength of 1.1 μm has gained the most due to skin depth effect. The threshold value of the input energy can be tailored and optimized by

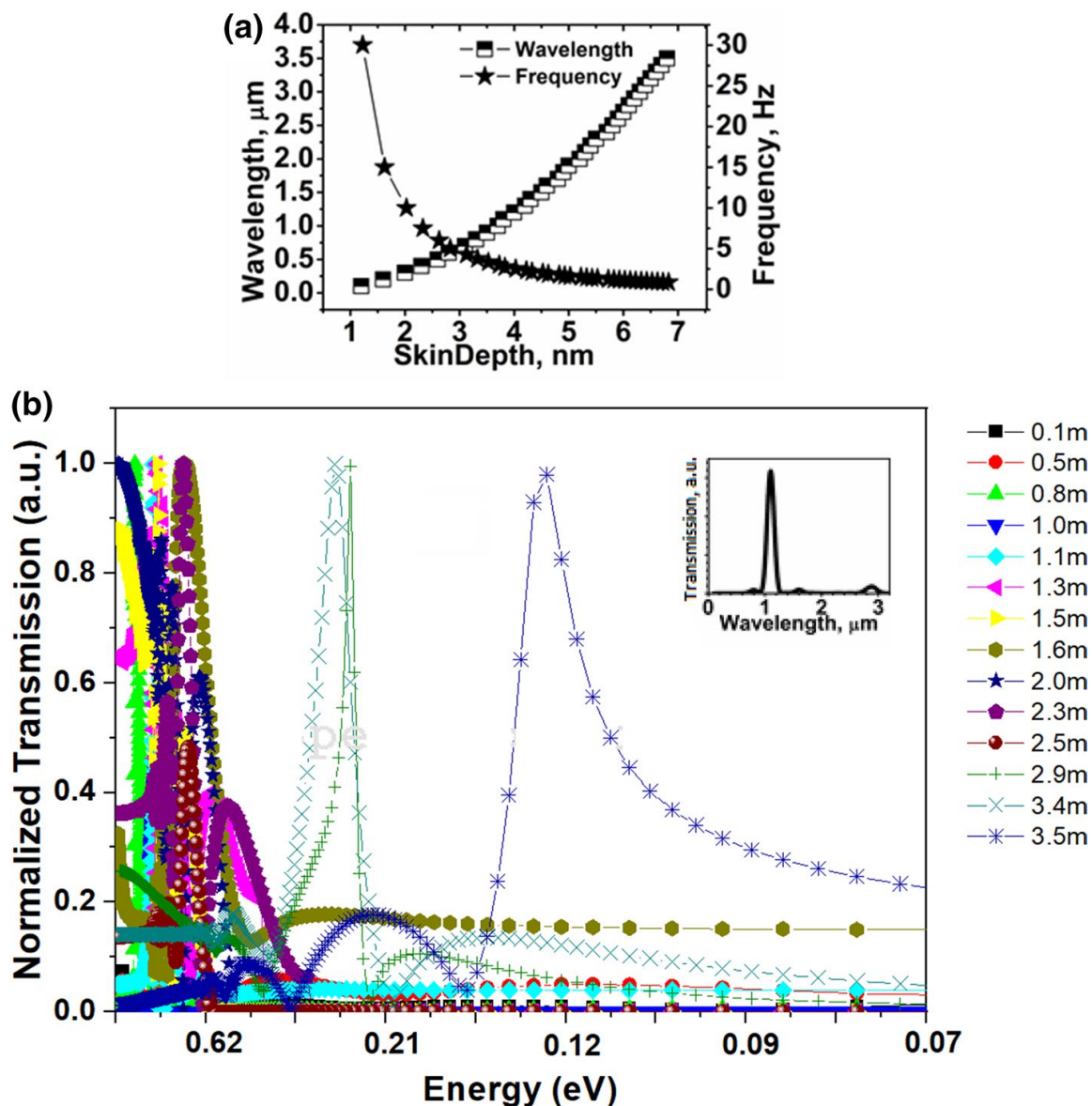


Fig. 2 **a** Plot showing skin depth dependence on incident wavelength and the corresponding frequency. **b** Normalized transmission spectra for a silver nanodisk revealing allowed energy bands for the system. The inset highlights the wavelength of 1.1 μm gains the most due to skin depth effect

changing the device parameters such as size, shape etc. and varies for system to system. The peak value of these bands are the energy eigenvalues and corresponding frequency for the nanodisks system. Further, these eigenvalues correspond to the quantum tunnelling phenomena where the plasmons assist the electron tunnelling through the dimer system and form the solutions of the system. This is further confirmed by the data recorded on different monitors (M_1 , M_2 and M_3) that are placed at different positions as shown in Fig. 3 and is described later in this section. Hence, the eigenvalues entails more tunnelling efficiency and the effect is maximum. We set 1.1 μm as the operating wavelength from launch L and proceed with a system of two silver nanodisks shown in Fig. 3a and vary the interparticle distance with monitors M_1 , M_2 and M_3 to record the transmission. It is observed that when the interparticle distance between the two nanodisks is less than 1.5 nm, there is a significant overlap of the gradient potential of the two resulting in the formation of tunnelling zone and it is also verified through electric field profile obtained via finite difference time domain method. Electric field E_y , profile obtained for the system (see Fig. 3b) confirms

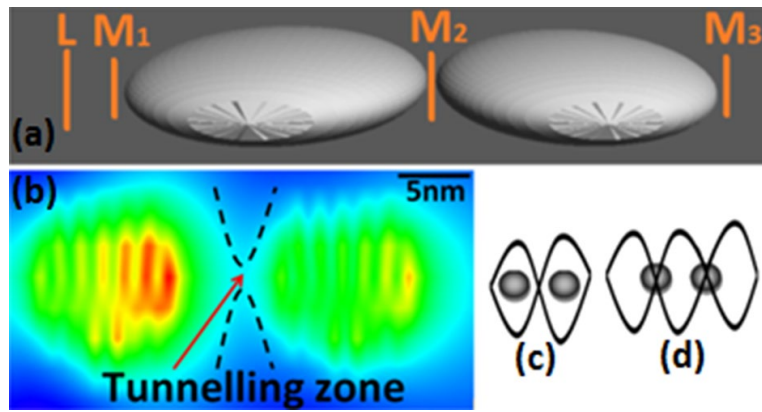


Fig. 3 **a** System of silver nanodisk dimer with subnanometer interparticle spacing with ‘L’ as the launch and monitors M_1 , M_2 and M_3 to record the transmission spectra. **b** Electric field profile for the nanodisk dimer with significant field intensity between the disks edifying the tunnelling zone. **c** and **d** Illustration depicting formation of node and antinode between the nanodisk consequently contributing to the total tunnelled efficiency

tunnelling zone which assists resonant plasmonic tunnelling. Besides, depending upon the interparticle distance there can be a node (Fig. 3c) or antinode (Fig. 3d) at the position of the monitors M_2 and M_3 contributing to the overall tunnelling efficiency.

The incident wavelength of $1.1\ \mu\text{m}$ is launched from L, and the monitors M_1 , M_2 and M_3 measure the transmission spectra with respect to the position of the monitor as shown in Fig. 4a–c and hence revealing the tunnelling efficiency as the distance between the two

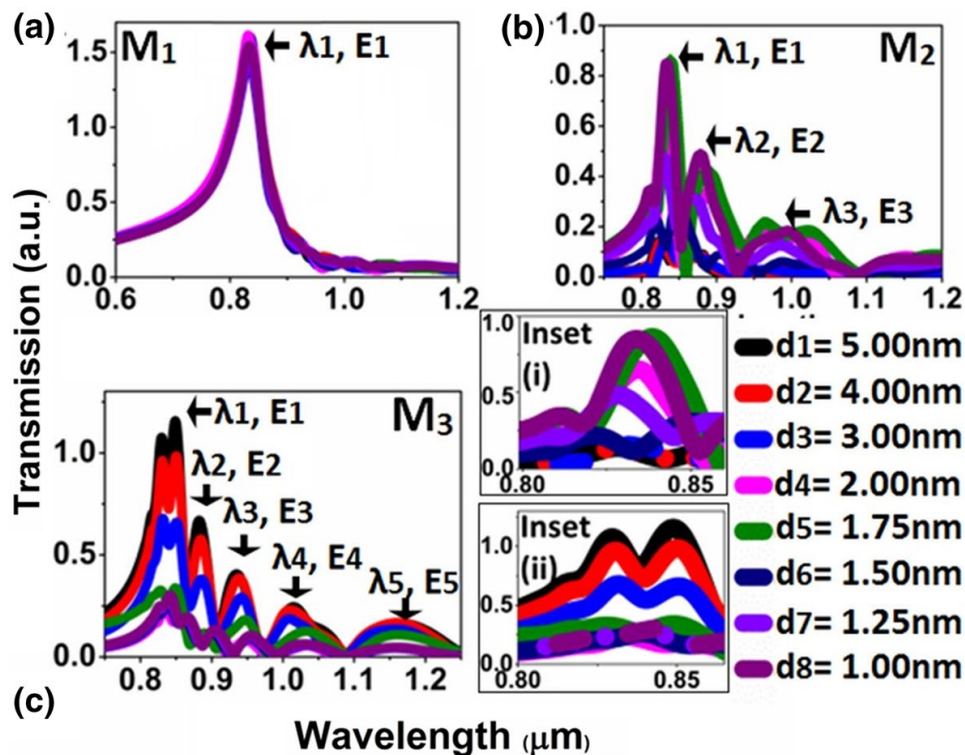


Fig. 4 **a–c** Show transmission spectra obtained on monitors M_1 , M_2 and M_3 respectively with wavelength and energy eigenvalues in (b) and (c) after tunnelling through the first and second nanodisk. Inset (i) and (ii) reveal zoomed in transmission recorded at M_2 and M_3 respectively

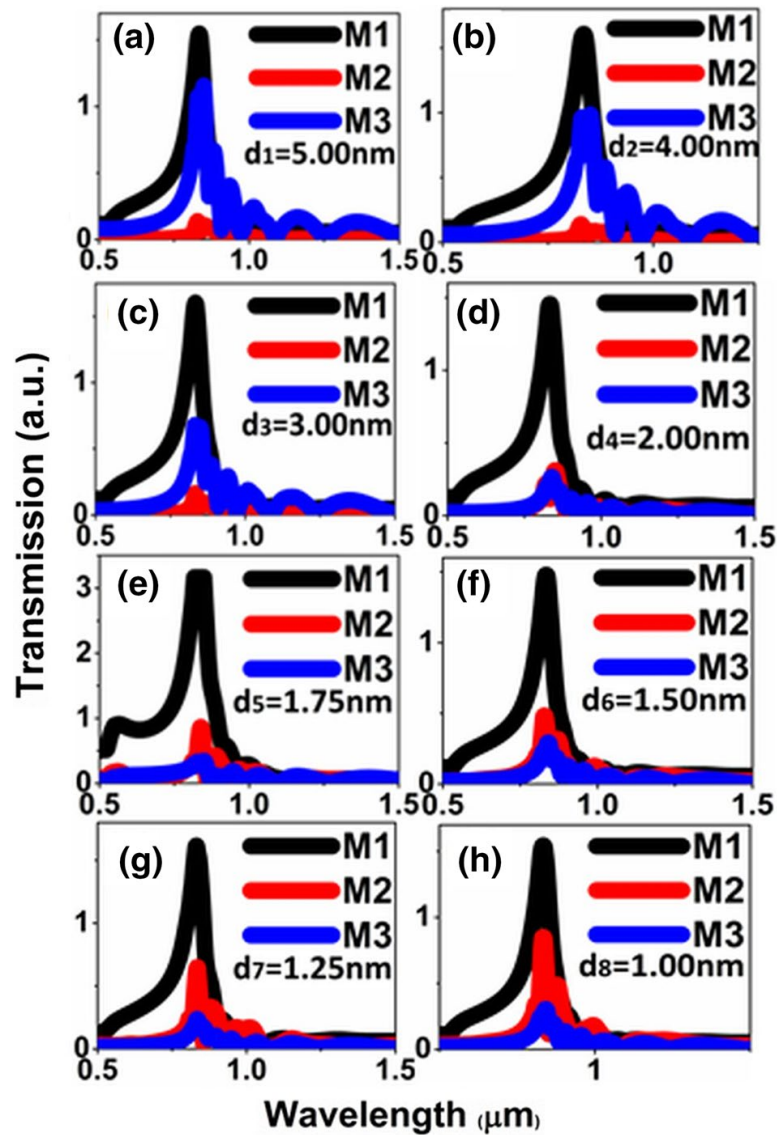
disks decreases from d_1 to d_8 as 5.00 nm, 4.00 nm, 3.00 nm, 2.00 nm, 1.75 nm, 1.50 nm, 1.25 nm and 1.00 nm. The reason for analysing them separately through these plots is exciting as the response received on each monitor is different and the reason behind it is the position where these monitors are placed. M_1 which is exactly in front of the launch shows symmetric response unlike M_2 where tunneling effect through one disk is observed and M_3 where tunneling through the dimer is observed. It is observed that for the input monitor M_1 all the curves obtained are symmetrical and exhibit single peak centred at $\lambda_1 = 0.84 \mu\text{m}$ (Fig. 4a) although there is a blue shift (or high energy shift) with respect to the input wavelength which happens due to the interaction of the enhanced local field of the first particle with the incident beam.

But at second monitor M_2 we notice some remarkable results. After the resonant plasmons tunnel through the first nanodisk the transmission characteristics obtained on monitor M_2 reveal two additional peaks λ_2 at $0.88 \mu\text{m}$ ($\pm 0.01 \mu\text{m}$) and λ_3 at $1.0 \mu\text{m}$ ($\pm 0.01 \mu\text{m}$) along with the central peak λ_1 (Fig. 4b). These subsequent multiple peaks at frequent intervals represent the allowed wavelengths corresponding to the allowed energy eigenvalues E_1 , E_2 and E_3 for the quantum system. Also, another important observation is that as the distance between the nanodisks decreases from d_1 to d_3 it is realised that there is steady increase in transmission (see inset (i)) but as the distance decreases further from d_4 to d_8 transmission due to plasmon assisted electron tunnelling increases significantly; with d_4 and d_5 showing maximum which is attributed to the in-phase (bonding mode) interaction of tunnelled plasmons at M_2 . This happens because when the distance is 1.5 nm or less, the gradient potential of the two nanodisks form a tunnelling zone supporting resonant plasmons assisting electron tunnelling and hence increase in transmission.

Further, as we record the transmission on the third monitor M_3 we observe that there are four additional peaks λ_2 at $0.88 \mu\text{m}$ ($\pm 0.01 \mu\text{m}$), λ_3 at $0.94 \mu\text{m}$ ($\pm 0.01 \mu\text{m}$), λ_4 at $1.02 \mu\text{m}$ ($\pm 0.01 \mu\text{m}$) and λ_5 at $1.16 \mu\text{m}$ ($\pm 0.01 \mu\text{m}$) along with the central peak λ_1 as shown in Fig. 4c. These wavelengths correspond to the energy eigenvalues E_1 to E_5 (Fig. 4c) for the system. The intensity of transmission on the monitor placed at M_3 position decreases (inset (ii)) with decreasing distance from d_1 to d_8 because as the nanodisks get closer, the dimer acts as a bigger particle offering larger resistance to the plasmons assisted electron tunnelling as well as out-of-phase interaction (antibonding mode) resulting in decreased intensity. Moreover, observing these transmission spectra in Fig. 4b, c, it is realised that there is an increase in the number of eigenvalues on M_3 as compared to that on M_2 . This is attributed to the redistribution of energy owing to the quantum effects when the plasmons assisted tunnelled electrons pass through second nanodisk (in case of M_3) and the system behaves as a coupled oscillating system due to the formation of tunnelling zone between the two nanodisk unlike in case of M_2 . Hence the proposed theory of gradient potential with skin depth dependence helps in understanding the output of the quantum system of silver nanodisks dimer.

In Fig. 5 we see the transmission spectra with respect to the distance having curves obtained on the three monitors for a fixed value of distance. Figure 5a–h demonstrates the transmission spectra obtained on M_1 , M_2 and M_3 when the distance between the nanodisks varies from d_1 to d_8 respectively. Another important observation is that there is notable intensity of red (M_2) and blue (M_3) curves which imply transmission due to significant plasmon assisted tunnelling with decreasing distance between the nanodisks (Fig. 5a–h). But, there is no regular increasing or decreasing pattern when distance between the nanodisks decreases from d_1 to d_8 . This can be understood from Fig. 3c, f since the position of the monitors M_2 and M_3 varies when distance changes from d_1 to d_8 and there can be formation of node or antinode at the position of M_2 and

Fig. 5 a–h Transmission spectra obtained on monitors M_1 , M_2 and M_3 with distance varying from d_1 to d_8 between the nanodisks showing variation of tunnelling efficiency with varying interparticle distance



M_3 resulting in less or high intensity peaks. Hence, it is inferred from Fig. 3c, d that when the distance is more than what is required for tunnelling zone, then the factors which control the tunnelling are node or antinode formation and in-phase or out-of-phase interaction at the monitor position yielding overall resonating plasmons assisted tunnelling efficiency. But when the disks are appreciably close resulting in a tunnelling zone then it is the gradient potential and the skin depth effect which control the total tunnelled efficiency.

We also obtained the electric field profile for the silver nanodisks quantum system to understand the field behaviour. Figure 6a–h display the electric field profile of the silver nanodisks quantum system demonstrating formation of a tunnelling zone when the distance between the nanodisks decreases from d_1 to d_8 . When observed carefully these results confirm that there is a region of high potential when the disks get closer leading to the formation of tunnelling zone thereby encouraging resonant plasmons assisted tunnelling and if the input wavelength is sufficient enough to result in higher skin depth then tunnelling efficiency is indeed high (Fig. 6e–h).

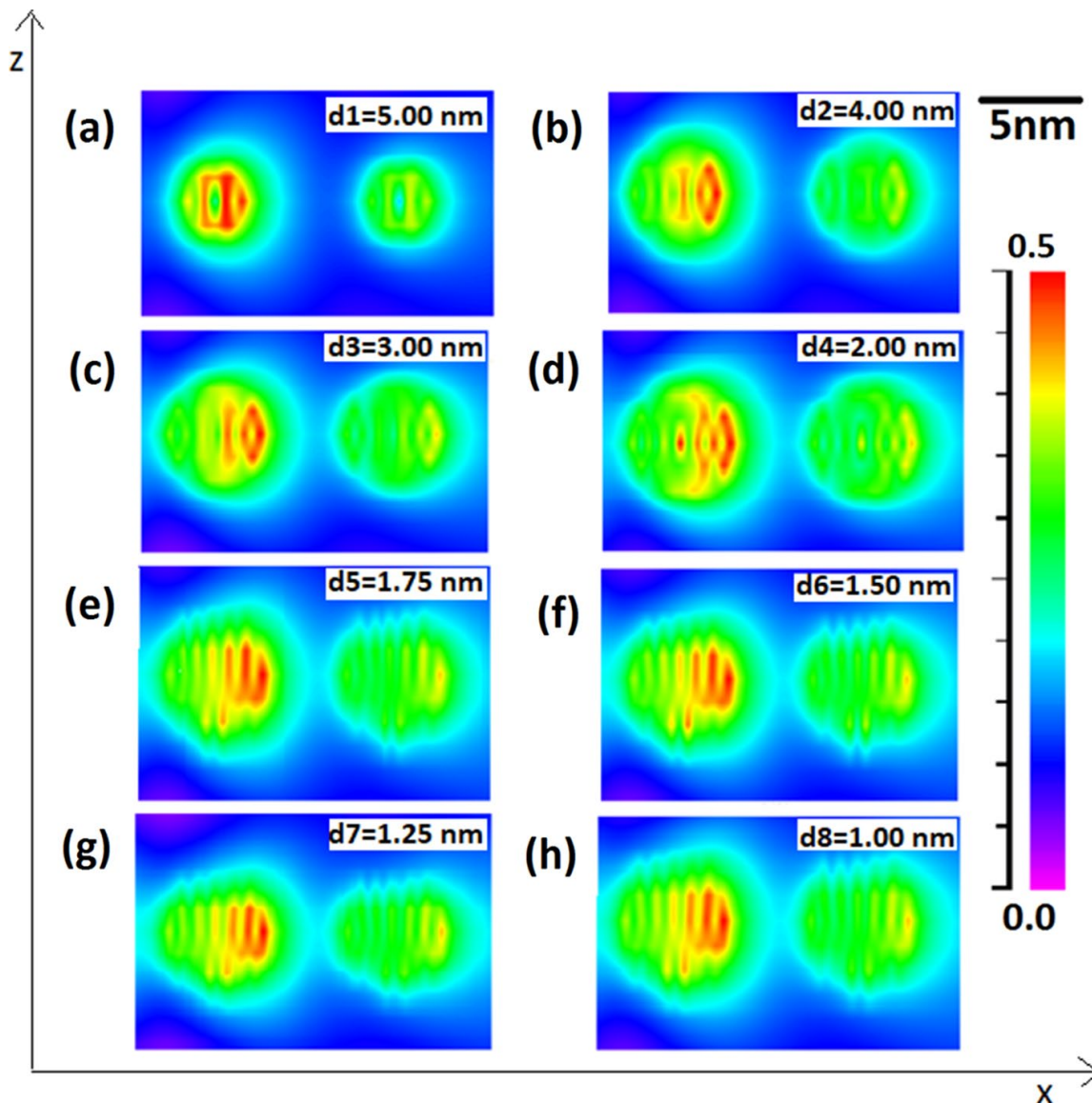


Fig. 6 a–h Electric field profile of the silver nanodisks quantum system with varying distance demonstrating formation of tunnelling zone when the distance between the nanodisks decreases from d_1 to d_8

4 Conclusion

Proposed gradient potential dependent skin depth theory efficiently describes the quantum behaviour of the metal nanoparticles system and helps in envisaging the transmission characteristics of the system precisely. It is observed that when the interparticle distance is less than 1.5 nm, the system of silver nanodisks dimer exhibits tunnelling zone and skin depth plays an important role in resonant plasmon tunnelling mechanism. Appreciably the energy eigenvalues and corresponding wavelengths are obtained for the dimer system with significant validation with the proposed theory. Hence it is the overall effect of skin depth, gradient potential along with in-phase or out-of-phase interaction, node or antinode formation and the role of distance between the nanodisks that comes as the final output of the plasmons assisted tunnelling quantum effect through the silver nanodisks dimer system. As the advent of nanotechnology brings down the experimental regime to nanoscales, the quantum effects become significant in plasmonic nanostructures. Quantum models for plasmonic systems help in appreciating the unique properties exhibited by them. Proposed

GPST provides an insight to explore the quantum plasmonic system, to apprehend the results hence providing the tool to predict the behaviour of the system under study. It can help to comprehend various quantum systems such as plasmon tunnel diode, plasmonic Josephson junction assisted superconductivity, plasmon tunnelled field-effect transistors etc. significantly improving the performance of integrated circuits.

Acknowledgements Authors gratefully acknowledge the support of “TIFAC-Center of Relevance and Excellence in Fiber Optics and Optical Communication” at Delhi College of Engineering, Delhi, through Mission Reach Program of Technology Vision 2020, Government of India and Sharda University for providing various resources.”

References

- Avrutsky, I., Zhao, Y., Kochergin, V.: Surface-plasmon-assisted resonant tunnelling of light through a periodically corrugated thin metal film. *Opt. Lett.* **25**(9), 595–597 (2000)
- Chen, P.Y., Hajizadegan, M., Sakhdari, M., Alu, A.: Giant photoresponsivity of midinfrared hyperbolic metamaterials in the photon-assisted-tunnelling regime. *Phys. Rev. Appl.* **5**(4), 041001 (2016)
- de Abajo, F.J.G., Gomez-Santos, G., Blanco, L.A., Borisov, A.G., Shabanov, S.V.: Tunneling mechanism of light transmission through metallic films. *Phys. Rev. Lett.* **95**(6), 067403 (2005)
- de Vega, S., de Abajo, F.J.G.: Plasmon generation through electron tunnelling in graphene. *ACS Photon.* **4**(9), 2367–2375 (2017)
- Garg, M., Kern, K.: Attosecond coherent manipulation of electrons in tunnelling microscopy. *Science* **367**(6476), 411–415 (2020)
- Haes, A.J., Van Duyne, R.P.: A nanoscale optical biosensor: sensitivity and selectivity of an approach based on the localized surface plasmon resonance spectroscopy of triangular silver nanoparticles. *J. Am. Chem. Soc.* **124**(35), 10596–10604 (2002)
- Huang, L., Maerkl, S.J., Martin, O.J.: Integration of plasmonic trapping in a microfluidic environment. *Opt. Express* **17**(8), 6018–6024 (2009)
- Huang, J.S., Callegari, V., Geisler, P., Brünig, C., Kern, J., Prangsma, J.C., Wu, X., Feichtner, T., Ziegler, J., Weinmann, P., Kamp, M.: Atomically flat single-crystalline gold nanostructures for plasmonic nanocircuitry. *Nat. Commun.* **1**(1), 1–8 (2010)
- Lan, Y.C., Chang, Y.C., Lee, P.H.: Manipulation of tunnelling frequencies using magnetic fields for resonant tunnelling effects of surface plasmons. *Appl. Phys. Lett.* **90**(17), 171114 (2007)
- Lan, Y.C., Chang, C.J., Lee, P.H.: Resonant tunnelling effects on cavity-embedded metal film caused by surface-plasmon excitation. *Opt. Lett.* **34**(1), 25–27 (2009)
- Liu, W.C., Tsai, D.P.: Optical tunnelling effect of surface plasmon polaritons and localized surface plasmon resonance. *Phys. Rev. B* **65**(15), 155423 (2002)
- Liu, S., Wolf, M., Kumagai, T.: Plasmon-assisted resonant electron tunneling in a scanning tunneling microscope junction. *Phys. Rev. Lett.* **121**, 226802–226801 (2018)
- Maier, S.A., Brongersma, M.L., Kik, P.G., Meltzer, S., Requicha, A.A.G., Atwater, H.A.: Plasmonics—a route to nanoscale optical devices. *Adv. Mater.* **13**(9), 1501–1505 (2001)
- Makarenko, K.S., Hoang, T.X., Duffin, T.J., Radulescu, A., Kalathing, V., Lezec, H.J., Chu, H.S., Nijhuis, C.A.: Efficient surface plasmon polariton excitation and control over outcoupling mechanisms in metal–insulator–metal tunneling junctions. *Adv. Sci.* **7**(8), 1900291 (2020)
- Martin-Moreno, L., Garcia-Vidal, F.J., Lezec, H.J., Pellerin, K.M., Thio, T., Pendry, J.B., Ebbesen, T.W.: Theory of extraordinary optical transmission through subwavelength hole arrays. *Phys. Rev. Lett.* **86**(6), 1114–1117 (2001)
- Nagel, P.M., Robinson, J.S., Harteneck, B.D., Pfeifer, T., Abel, M.J., Prell, J.S., Neumark, D.M., Kaindl, R.A., Leone, S.R.: Surface plasmon assisted electron acceleration in photoemission from gold nanoparticles. *Chem. Phys.* **414**, 106–111 (2013)
- Park, J., Kim, H., Lee, I.M., Kim, S., Jung, J., Lee, B.: Resonant tunnelling of surface plasmon polariton in the plasmonic nano-cavity. *Opt. Express* **16**(21), 16903–16915 (2008)
- Popov, E., Nevier, M., Enoch, S., Reinisch, R.: Theory of light transmission through subwavelength periodic hole arrays. *Phys. Rev. B* **62**(23), 16100–16108 (2000)
- Pshenichnyuk, I.A., Nazarikov, G.I., Kosolobov, S.S., Maimistov, A.I., Drachev, V.P.: Edge-plasmon assisted electro-optical modulator. *Phys. Rev. B* **100**(19), 195434 (2019)

- Ramazani, A., Shayeganfar, F., Jalilian, J., Fang, N.X.: Exciton-plasmon polariton coupling and hot carrier generation in two-dimensional SiB semiconductors: a first-principles study. *Nanophotonics* **9**(2), 337–349 (2020)
- Schaadt, D.M., Feng, B., Yu, E.T.: Enhanced semiconductor optical absorption via surface plasmon excitation in metal nanoparticles. *Appl. Phys. Lett.* **86**(6), 063106 (2005)
- Sidorenko, S., Martin, O.J.F.: Resonant tunnelling of surface plasmon-polaritons. *Opt. Express* **15**(10), 6380–6388 (2007)
- Smith, W.C., Kou, A., Xiao, X., Vool, U., Devoret, M.H.: Superconducting circuit protected by two-Cooper-pair tunnelling. *NPJ Quant. Inf.* **6**(1), 1–9 (2020)
- Stellacci, F., Bauer, C.A., Meyer-Friedrichsen, T., Wenseleers, W., Alain, V., Kuebler, S.M., Pond, S.J.K., Zhang, Y., Marder, S.R., Perry, J.W.: Laser and electron-beam induced growth of nanoparticles for 2D and 3D metal patterning. *Adv. Mater.* **14**(3), 194–198 (2002)
- Stockman, M.I.: Nanoplasmonics: the physics behind the applications. *Phys. Today* **64**(2), 39–44 (2011)
- Stolz, A., Berthelot, J., Mennemanteuil, M.M., Colas des Francs, G., Markey, L., Meunier, V., Bouhelier, A.: Nonlinear photon-assisted tunnelling transport in optical gap antennas. *Nano Lett.* **14**(5), 2330–2338 (2014)
- Svintsov, D., Devizorova, Z., Otsuji, T., Ryzhii, V.: Plasmons in tunnel-coupled graphene layers: backward waves with quantum cascade gain. *Phys. Rev. B* **94**(11), 115301 (2016)
- Uskov, A.V., Khurgin, J.B., Protsenko, I.E., Smetanin, I.V., Bouhelier, A.: Excitation of plasmonic nanoantennas by nonresonant and resonant electron tunnelling. *Nanoscale* **8**(30), 14573–14579 (2016)
- Wang, G., Lu, H., Liu, X., Mao, D., Duan, L.: Tunable multi-channel wavelength demultiplexer based on MIM plasmonic nanodisk resonators at telecommunication regime. *Opt. Express* **19**(4), 3513–3518 (2011)
- Wang, P., Krasavin, A.V., Nasir, M.E., Dickson, W., Zayats, A.V.: Reactive tunnel junctions in electrically driven plasmonic nanorod metamaterials. *Nat. Nanotechnol.* **13**(2), 159–164 (2018)
- Xiao, S., Mortensen, N.A.: Resonant-tunnelling-assisted crossing for subwavelength plasmonic slot waveguides. *Opt. Express* **16**(19), 14997–15005 (2008)
- Xu, C., Li, C., Jin, Y.: Programmable organic-free negative differential resistance memristor based on plasmonic tunnel junction. *Small* **16**(34), 2002727 (2020)

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

Point-of-Care PCR Assays for COVID-19 Detection

Niharika Gupta ¹, Shine Augustine ¹ , Tarun Narayan ² , Alan O'Riordan ² , Asmita Das ¹, D. Kumar ³, John H. T. Luong ^{4,*}  and Bansi D. Malhotra ^{1,*} 

¹ Department of Biotechnology, Delhi Technological University, Shahbad Daulatpur, Delhi 110042, India; niharika.gupta990@gmail.com (N.G.); shine2089@gmail.com (S.A.); asmita1710@gmail.com (A.D.)

² Nanotechnology Group, Tyndall National Institute, University College Cork, T12 K8AF Cork, Ireland; tarun.narayan@tyndall.ie (T.N.); alan.oriordan@tyndall.ie (A.O.)

³ Department of Applied Chemistry, Delhi Technological University, Shahbad Daulatpur, New Delhi 110042, India; dkumar@dce.ac.in

⁴ School of Chemistry, University College Cork, T12 K8AF Cork, Ireland

* Correspondence: j.luong@ucc.ie (J.H.T.L.); bansi.malhotra@gmail.com (B.D.M.)

Abstract: Molecular diagnostics has been the front runner in the world's response to the COVID-19 pandemic. Particularly, reverse transcriptase-polymerase chain reaction (RT-PCR) and the quantitative variant (qRT-PCR) have been the gold standard for COVID-19 diagnosis. However, faster antigen tests and other point-of-care (POC) devices have also played a significant role in containing the spread of SARS-CoV-2 by facilitating mass screening and delivering results in less time. Thus, despite the higher sensitivity and specificity of the RT-PCR assays, the impact of POC tests cannot be ignored. As a consequence, there has been an increased interest in the development of miniaturized, high-throughput, and automated PCR systems, many of which can be used at point-of-care. This review summarizes the recent advances in the development of miniaturized PCR systems with an emphasis on COVID-19 detection. The distinct features of digital PCR and electrochemical PCR are detailed along with the challenges. The potential of CRISPR/Cas technology for POC diagnostics is also highlighted. Commercial RT-PCR POC systems approved by various agencies for COVID-19 detection are discussed.

Keywords: polymerase chain reaction; COVID-19; electrochemical; digital PCR; point-of-care



Citation: Gupta, N.; Augustine, S.; Narayan, T.; O'Riordan, A.; Das, A.; Kumar, D.; Luong, J.H.T.; Malhotra, B.D. Point-of-Care PCR Assays for COVID-19 Detection. *Biosensors* **2021**, *11*, 141. <https://doi.org/10.3390/bios11050141>

Received: 16 March 2021

Accepted: 28 April 2021

Published: 1 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The coronavirus disease 2019 (COVID-19) outbreak crisis has changed the shape of our world since its first report in December 2019. While some countries seem to be recovering from the crisis and are reporting fewer cases, others are still witnessing an increasing number of cases [1]. Clinical diagnosis has been the forerunner in controlling the COVID-19 pandemic. Molecular nucleic acid amplification tests (NAATs) were the first to be developed for detecting SARS-CoV-2 RNA in patient samples. Particularly, reverse transcriptase-polymerase chain reaction (RT-PCR) and its quantitative variant (qRT-PCR) have been the keystone for diagnosis of SARS-CoV-2 with the capacity to detect target nucleic acids (<100 copies/mL) with remarkable sensitivity [2]. However, the analysis proved time-intensive, requiring up to a few hours, and could only be performed in a centralized laboratory. The high false-negative rates with some RT-PCR assays also raised concern. Thus, attention shifted to faster, cheaper, and equally sensitive (if not more) point-of-care (POC) biosensing devices that could be deployed for mass screening.

Therein began a major shift in the clinical diagnostic industry, with point-of-care testing (POCT) becoming the focus of attention almost overnight. Lateral flow assays (LFAs), chemiluminescence, and nanoparticle-based colorimetric detection were developed for detecting SARS-CoV-2-related antigens and antibodies produced in response to its infection [3–8]. Faster, miniaturized isothermal amplification tests emerged that could detect the virus within a few minutes and with sensitivity at par with RT-PCR assays [5,9,10].

Although different types of POCT devices have been authorized in various countries for emergency use, many novel biosensing strategies and designs still seek validation and are currently subject to academic inquiry.

These devices have shorter response times and have cost-effectively enabled population-wide mass screening. However, evidence suggests that the analytic performance (sensitivity, specificity, positive and negative predictive values, etc.) of current antigen diagnostic tests is not at par with that of RT-PCR and other NAATs [11]. Thus, while rapid antigen tests and other POCT are being widely used for COVID-19 screening, it is still uncertain whether such tests will be regularized and used in routine diagnostic procedures. In attempt to synergize the sensitivity of NAATs and the ease of use of POCT assays, miniaturized NAAT-based POCT devices and assays were devised for faster screening and diagnosis of COVID-19. One of the first such devices was the Abbott ID Now, which integrates isothermal amplification with colorimetric detection to yield results within 5 min. However, questions were soon raised about its utility as a singular diagnostic test due to its low positive predictive value (PPV) and high false-negative rates, especially in samples with low viral load [10]. More rapid devices based on isothermal amplification with improved performance were devised. Thus, although the integration of isothermal amplification in POC devices has gained some success, they are not as successful as RT-PCR for COVID-19 detection. In general, the high temperature requirements of RT-PCR prevent non-specific amplification, which is more common in isothermal amplification techniques. Conversely, these temperature requirements somewhat complicate the development of PCR-based rapid devices.

Nonetheless, efforts have been directed toward miniaturizing PCR to make it an automated, high-throughput device that can be applied at point-of-use. In this review, we summarize studies related to the development of miniaturized, high-throughput PCR biosensors for COVID-19 detection. The distinct features, limitations, and advantages of various types of PCR biosensors and chips are discussed. The advantages and limitations of PCR chips over biosensors based on other amplification assays are listed. The potential of biosensing formats to be integrated with RT-PCR is explored, along with the path-breaking integration of CRISPR/Cas technology with amplification assays toward the development of faster, miniaturized devices and chips.

2. RT-PCR: The Gold Standard

RT-PCR is the first molecular diagnostic test to be employed for detecting SARS-CoV-2 RNA in patient samples and is currently considered the gold standard for COVID-19 diagnosis. Different RT-PCR assays have been designed for detecting SARS-CoV-2 virus RNA in different body fluids, such as nasopharyngeal swabs, lower respiratory tract fluid, sputum, saliva, etc. [12–14]. However, RT-PCR is prone to false-negative results that reduce the overall sensitivity of the diagnosis. This may be because of various reasons such as low viral load in the pharyngeal, nasal, and sputum samples; storage and transport of samples; and improper handling [15,16]. Moreover, any mismatches between the primers and probe–target regions compromise the assay performance, leading to false-negative results [15,17]. Another major challenge faced by RT-PCR is that it can yield false-positive results by amplifying RNA from dead, noninfectious viruses as well [18]. Thus, recovered patients that no longer hold the threat of transmitting the disease may be positive per RT-PCR tests.

The current challenges of the qRT-PCR method include the use of fluorescent label binding to the source signal produced by the amplified DNA, which not only increases the cost of the instrument, but also the complexities. This technology is less appealing to developing nations or remote locations with limited resources. Commercial RT-PCR kits have not been subject to rigorous quality control. Personnel skills and good laboratory practice play an important role in Biosafety Level 3. Optimum sample types and timing for peak viral load remain to be fully investigated as sputum or nasal swabs are the most accurate sample for diagnosis of COVID-19, but not throat swabs.

Despite these limitations, RT-PCR remains the gold standard for confirming the diagnosis of COVID-19. There have been multiple attempts to develop portable PCR systems since the inception of the pandemic. Lab-in-tube systems incorporating lysis, reverse transcription, amplification, and detection in a single tube within 36 min were demonstrated in May 2020 [19]. A lab-on-chip device, CovidNudge, can be used to perform sample processing and real-time RT-PCR outside of a laboratory setting [20] (Figure 1). The chip consists of detection arrays for seven SARS-CoV-2 genes and one host gene as a sample adequacy control. This device detects the virus in 90 min and reduces the collection-to-result turnaround time significantly by eliminating the requirement of sample transport from the site of collection to a centralized lab. The sensitivity of this POC test (94%) is comparable to that of lab-based tests in clinical settings. As of September 2020, over 5 M CovidNudge kits had been deployed in the U.K. for COVID-19 testing. Of note is a portable RT-PCR workstation for COVID-19 detection in under-served and remote areas [21]. This workstation is a chip-based, battery-operated qRT-PCR system with the capability of network data transfer and automated reporting. Almost 3.8% (2.7 million) of the total tests conducted in India were performed on these workstations (as of September 2020). The average cost of an RT-PCR varies in different parts of the world. In India, for example, the cost of a conventional RT-PCR test currently varies from INR 400 (~USD 5.30) to INR 950 (~USD 12.6), and POC rapid antigen tests are free. While the CovidNudge test (Table 1) deployed in the U.K. costs around GBP 10 (per test) (equivalent to ~USD 13.80), which is almost 10 times cheaper than the average cost (~GBP 100) of a conventional RT-PCR test in the country.

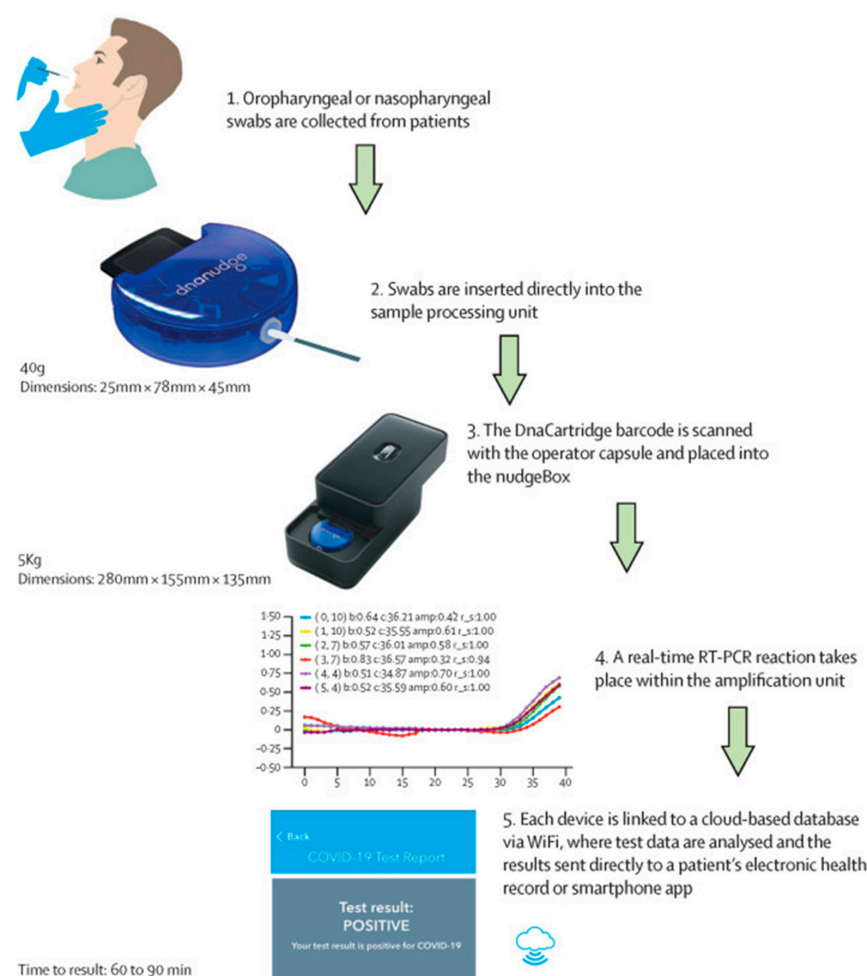


Figure 1. Schematic diagram depicting the various steps performed by the CovidNudge assay for automated detection of SARS-CoV-2 RNA (Reprinted with permission from Ref. [20]).

Table 1. List of commercial, automated RT-PCR systems authorized under emergency use.

Name of the Kit	Target Genes	Type	Sample Preparation	No. of Tests	Time	LOD	Sensitivity	Specificity	Cost (Per Test)	Reference
CovidNudge	rdrp1, rdrp2, E gene, N gene, n1, n2, and n3	RT-PCR	Automated	NA	~90 min	5 copies/ μ L	>94%	100%	GBP 10	[20]
Accula SARS-CoV-2 Test	N gene	RT-PCR	Automated	NA	~30 min	NA	100%	100%	USD 20	[22]
Cepheid Xpert Xpress SARS-CoV-2 assay	N2 and E	RT-PCR (real time)	Automated	10 per kit		0.02 PFU/mL			USD 19.8	[23]
FastPlex Triplex SARS-CoV-2 Detection Kit	ORF1ab, N, RPP30	RT-dPCR	Manual	96 test per kit	90 min	285.7 copies/mL	>95%	95.7%	USD 1152	[23]
Gnomegen COVID-19 RT-Digital PCR Detection Kit	N1, N2	RT-dPCR	Manual	48 samples per day	180 min	2.5 copies per reaction	>95%	99%	NA	[23]
Bio-Rad SARS-CoV-2 ddPCR Test	N1, N2	RT-dPCR	Manual	96 samples	NA	400 copies/mL			NA	[23]
ePlexSARS-CoV-2 Test	N gene	End-point RT-PCR with electrochemical Detection	Automated	12 tests/kit	NA	1×10^3 copies/mL	99.02%	98.41%	NA	[23]

There have been several other innovations related to the fabrication of PCR chips and biosensors for COVID-19 detection. The following sections cover some of these studies and discuss the potential and challenges faced by such devices in emerging as viable commercial products.

3. RT-PCR Biosensors

3.1. Digital RT-PCR

The concept of digital PCR (dPCR) was pioneered by Vogelstein and Kinzler in 1999 [24]. The principle of dPCR is to partition the reaction mixture into many sub-reactions before amplification; the original numbers are determined by counting the partition showing negative and positive reactions [25] (Figure 2). It does not require a standard curve or reference genes and is more resistant to interference factors such as specific template amplification inhibitors [26,27]. The quantification results are analyzed from Poisson's distribution and can achieve an accurate estimation of low concentrations of nucleic acid samples [26]. Therefore, a method like dPCR offers high sensitivity, higher precision, and resistance to inhibitors, which are required for an accurate SARS-CoV-2 diagnosis. The dPCR method can be classified into three types based on liquid separation: droplet-based (ddPCR), chip-based (cdPCR), and microfluidic digital PCR (mdPCR). The primary difference between these three types of digital PCR is the design of the sample partitioning system in the detection platform: ddPCR combines several millions partitioning of the PCR test into individual droplets in a water-in-oil emulsion [26,28], whereas cdPCR uses an active partitioning approach. It has two chip halves with two arrays of microwells. The chambers are aligned so that the opposite halves form continuous channels [28,29]. In mdPCR, microfluidic chambers are used to split the samples. These chambers are fluidically designed such that each sample can be partitioned into tens of thousands of wells [30].

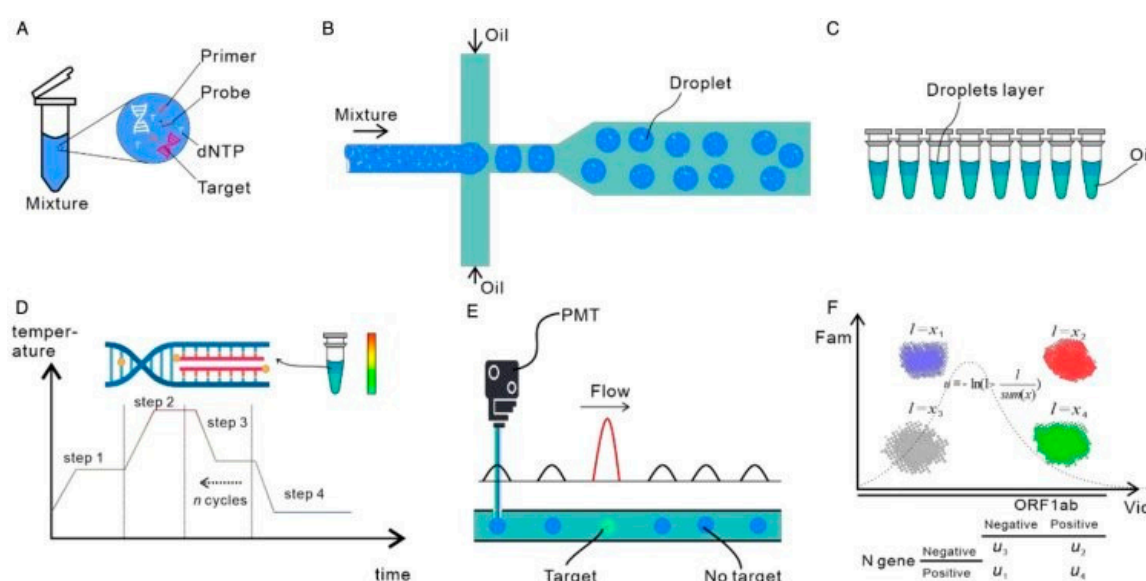


Figure 2. Schematic depicting workflow of a ddPCR system: (A) preparation for amplification, (B) generation of water-in-oil droplets using a microfluidic flow system, (C) collection of the droplets in PCR tubes, (D) PCR amplification, (E) analysis of fluorescence in the droplets after amplification, and (F) fitting to Poisson distribution to determine the absolute copy numbers of the target molecules (Reprinted from Ref. [28]).

ddPCR can be used for the quantification of a low viral load, monitoring of the virus in the environment, evaluation of anti-SARS-CoV-2 drugs [28], and the detection of viral mutations [31]. Many types of clinical samples can be used for COVID-19 testing using dPCR, including blood, urine, sputum, stool, nasal swabs, and throat swabs. Studies have compared RT-PCR with RT-dPCR for the presence of SARS-CoV-2 in pharyngeal swab samples and found RT-dPCR to be more sensitive and accurate than RT-PCR [32,33].

Lu and group showed that RT-dPCR has a detection limit ten-fold lower than that of RT-PCR [34]. They compared the RT-dPCR and RT-PCR of 36 COVID-19 patients with 108 specimens, including blood, pharyngeal swab, and stool, in which four pharyngeal samples yielding negative results in RT-PCR were positive per RT-dPCR. Another study demonstrated that suspected patients who tested negative by RT-PCR were found to be positive by ddPCR [35]. The results of ddPCR were validated by the serological testing of anti-COVID-19 antibodies in the samples. The ddPCR can yield better and more precise quantitation of viral loads of SARS-CoV-2 [36–38]. However, most of the reported ddPCR procedures included an RNA extraction and purification step, which can lead to potential amplification errors [38]. Moreover, direct quantification by ddPCR targeting the envelope (E) gene [39], ORF1ab gene [40], and nucleocapsid (N) [41] region have also been reported. The viral load can be quantified in throat swabs, sputum, nasal swabs, blood, and urine [37]. Droplet-based dPCR was also used to detect SARS-CoV-2 RNA in airborne aerosols [42], in which the viral load in the toilets used by some medical personnel and patients was found to be high. This study indicated the significance of sanitization and room ventilation for limiting COVID-19 spread. The primary advantage of dPCR is its good sensitivity and high-throughput analysis, which has been the key requirement for COVID-19 detection. Currently, there are three commercial dPCR tests authorized for emergency use by the USFDA (Table 1).

However, a few challenges require the utmost attention before dPCR can be used in routine diagnostics. Particularly, much like conventional PCR tests, dPCR also requires expensive instruments, reagents, and professional experts to operate the system. The fabrication of the dPCR chips requires complex steps, making it a costly operation. Moreover, much like other POC tests, strict standards and guidelines need to be followed to assure the quality of results obtained from dPCR systems.

3.2. Electrochemical PCR: Unexplored Potential

The integration of electrochemistry with RT-PCR aims to provide a rapid, miniaturized, hand-held instrument. Electrochemical biosensors work by modification of a working electrode with a biomolecule that interacts with a specific target analyte present in an aqueous electrolyte and generates an electrical signal corresponding to its concentration. In the case of an electrochemical PCR, there is an electroactive species whose oxidation or reduction signal is correlated to the amount of PCR amplified product. A more challenging approach is the use of nanomaterials to tag the DNA primers used in the PCR amplification step, such as gold nanoparticles (AuNPs) or semiconductor quantum dots (QDs). The labeled amplified products are then further quantified via the generation of electrochemical signals.

Electrochemical systems offer the benefits of being seamlessly implemented into compact and intelligent systems, enabling high versatility and real-time detection. Moreover, electrochemically active labels (such as metal-complex, organic molecules, etc.) are more durable than fluorescent dyes (Cy5, FAM, etc.) and are a notable factor toward the commercial applications of electrochemical-RT-PCR (EPCR). The power and sample volume requirements are lower for electrochemical biosensors compared with RT-PCR. Despite the considerable interest, electrochemical biosensors have garnered in the context of COVID-19 detection, the clinical industry appears reluctant to adopt this technology for practical and commercial use.

The pre-COVID era witnessed the emergence of PCR-free electrochemical assays for detecting different nucleic acid targets, including microRNA, viral RNA and DNA, and cancer-related genes [43–45]. Perhaps the research community has been confident that electrochemical assays can compete with the existing PCR technology in terms of sensitivity and turnaround times and eliminate the use of costly reagents and dyes [46]. There have been some studies on PCR-integrated electrochemical biosensors in the last 5 years. Some of the recent studies have demonstrated innovative PCR-free electrochemical sensors for

SARS-CoV-2 RNA detection with remarkable detection limits [47,48]; however, none has yet achieved a commercial or authorized status.

Integrating PCR with electrochemical transducers poses various challenges; the primary challenge includes the capability of the sensing surface to withstand the harsh temperature changes and salt concentrations required during PCR [49]. Isothermal amplification techniques are preferred over PCR for integration with electrochemical sensors. A rapid electrochemical detection system based on rolling circle amplification (RCA) was demonstrated for multiplex detection of the S and N genes of SARS-CoV-2 [50] (Figure 3). Sandwich hybridization was employed in this study, with oligonucleotide probes consisting of redox-active labels (methylene blue-and-acridine orange) for electrochemical detection using differential pulse voltammetry. This assay detects the N or S viral gene at a concentration as low as 1 copy/ μL within 2 h with high selectivity and sensitivity.

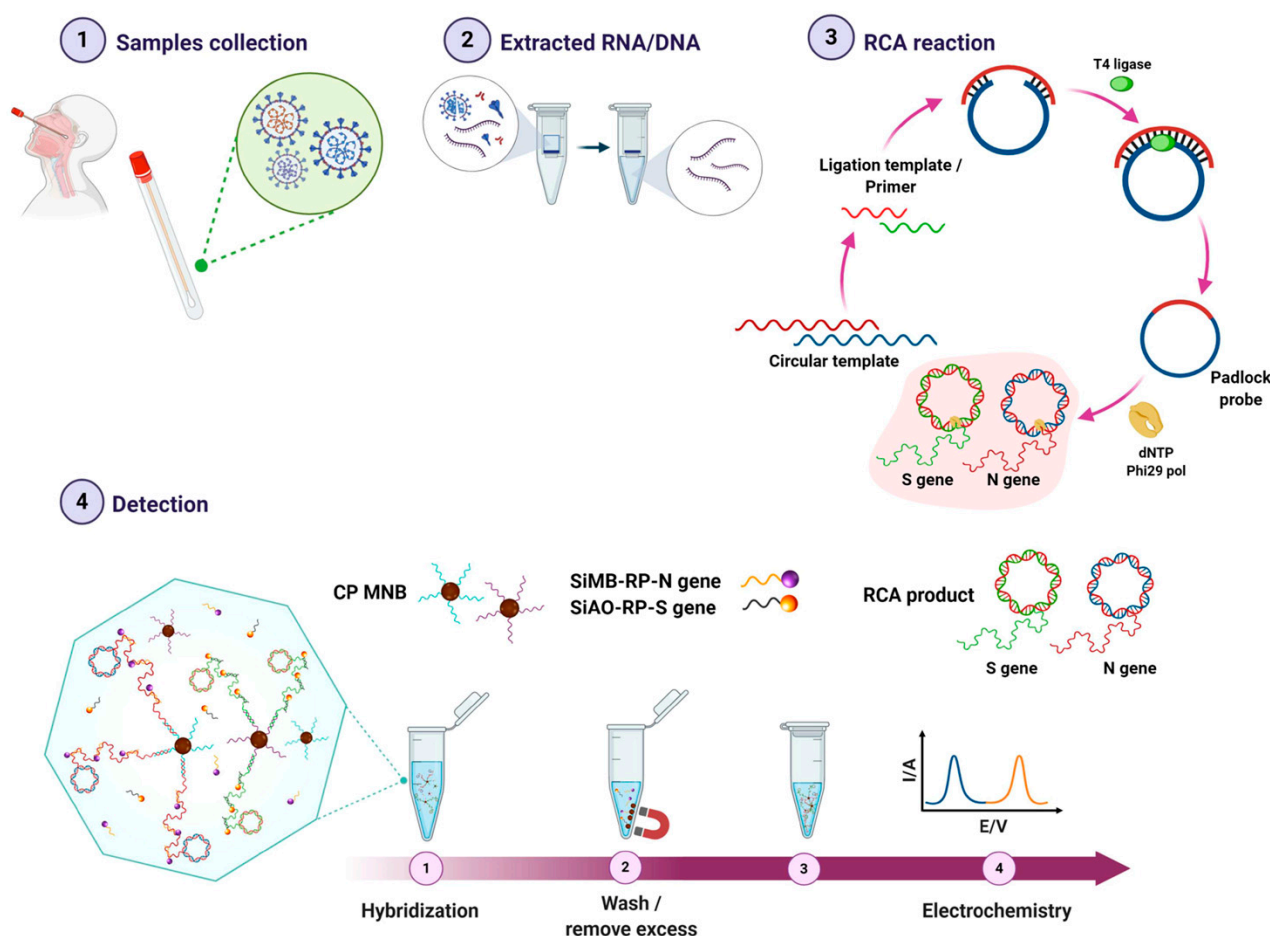


Figure 3. Workflow of the RCA-based electrochemical sensor for SARS-CoV-2 detection (Reprinted from Ref. [50]).

The recent advances in microfluidics technology have enabled the integration of electrochemical electrodes with miniaturized reaction chambers (or chips) designed for PCR. The USFDA recently approved the GenMark ePlex[®] SARS-CoV-2 test, which automates RNA extraction and amplification, and then further integrates it into competitive hybridization-based electrochemical detection [51]. This system uses the principle of electrowetting (digital microfluidics) to manipulate the movement of samples and reagents on a printed circuit board (PCB) (Table 1).

4. CRISPR/Cas-Based Sensors: The New Alternative

CRISPR stands for clustered regularly interspaced short palindromic repeat, which utilizes genetic information of bacterial species as a part of an antiviral process. CRISPR/Cas

is a genetic editing technology whose precise and specific DNA and RNA cleavage ability makes it a useful tool in nucleic acid diagnostics. CRISPR/Cas-based sensors mainly utilize single guide RNA in conjunction with the Cas system to bind to a target sequence or cleave target DNA and RNA, resulting in signal generation. Owing to their high specificity, they are an attractive alternative to POC RT-PCR devices. CRISPR/Cas-based diagnostics circumvents the issue of long turnaround times and enhances the assay specificity [52]. Recently, Hou et al. developed a rapid assay known as CRISPR–COVID for detecting SARS-CoV-2 with less turnaround time (~40 min) compared with RT-PCR and metagenomics sequencing [53]. Another advantage of using CRISPR/Cas systems is the exclusion of RNA isolation and amplification, making it a faster analysis method. An ultrasensitive RT-RPA CRISPR–fluorescence detection system (FDS) assay can eliminate the need for RNA isolation for SARS-CoV-2 detection [54]. It uses a saliva sample that is subject to a mix of chemicals that amplify the viral RNA, which is then subjected to CRISPR/Cas12a-based fluorescence signal amplification. The linear range of this handheld CRISPR-based test was found to be 1 to 10^5 copies/mL with a limit of detection of 0.38 copies/mL, which is consistent with the result obtained using qRT-PCR. In another approach, the need for SARS-CoV-2 RNA pre-amplification was eliminated with the use of CRISPR-Cas13a, which aids the detection of SARS-CoV-2 RNA from nasal swabs [55]. The main highlight of this study was the use of different sets of crRNAs to increase the sensitivity by activation of a greater number of Cas13a per target RNA. Additionally, the study reported the ability to directly translate the fluorescent signal into viral loads, thus resulting in remarkable sensitivity compared with other CRISPR-based assays for COVID detection.

5. Future Outlook

COVID-19 diagnostics has evolved significantly since its first appearance. The range and types of diagnostic devices that have emerged in the past year are immensely diverse. Several earlier diagnostic devices and assays were only the subject of academic interest and research but are now commercially available for use. However, since most of the POC devices for COVID-19 detection have been authorized under emergency use, caution should be taken when extrapolating the use of such devices for the diagnosis of other diseases.

Despite the advances, there are limitations associated with RT-PCR POC devices and biosensors concerning sample preparation in ePCR, false negatives and positives, and reagent evaporation in dPCR. Efforts to identify the limitations in current PCR devices for COVID-19 detection can soon help in the design of improved diagnostic devices. Additionally, different detection strategies and platforms can be integrated to develop new, hybrid devices for improved performance. For example, electrokinetic focusing on microfluidic chips was used to automate the process of nucleic acid purification and amplification with a reduction in non-specific amplification [56]. A recent study used isotachopheresis (ITP), an ionic focusing technique, on a microfluidic chip to automate SARS-CoV-2 RNA purification and subsequent detection by CRISPR-based technique within 35 min [57]. This on-chip device uses a smaller volume of reagents (<100 times lower) and automates sample preparation and subsequent detection. Reduction in bubble generation and reagent evaporation in dPCR systems was also demonstrated by creating a vertical polymeric barrier leading to ultrafast PCR amplification [58].

Centrifugal microfluidic platforms (or lab-on-a-disc) for automated sample preparation and subsequent RT-PCR can also be conceived. These devices use different layers of polymeric substrates to integrate multiple steps involving complex fluid flow. These centrifugal systems were shown to improve reaction rates using efficient mixing, thus enabling high sensitivity and reduced hybridization times [59]. Paperfluidic devices that involve the creation of microfluidic channels on paper can also be realized for SARS-CoV-2 RNA detection. Apart from being inexpensive, paperfluidic devices do not require any additional step to render the channels hydrophilic for fluid flow; the intrinsic hydrophilicity of paper allows fluid flow via capillary action, thus eliminating the need for external pumps. This allows their use in resource-limited, POC settings. These devices, much like LFAs, can

be batch fabricated at minimal cost and can thus be used in mass screening operations in resource-limited settings. A paper-based assay, FnCas9 editor-linked uniform detection assay (FELUDA), was developed in India, which enables detection of single nucleotide variants [60]. This test uses RT-PCR followed by CRISPR-based detection in a lateral flow format. Similarly, paperfluidic devices that can integrate RNA extraction, amplification, and subsequent detection can be realized [61].

COVID-19 diagnostics has provided new opportunities and advances in the clinical diagnostic sector. It will be interesting to see how these developments affect the overall diagnostics landscape over time.

6. Conclusions

Molecular diagnostics has been the cornerstone in controlling the ongoing COVID-19 pandemic. RT-PCR is currently the primary gold standard for COVID-19 diagnosis. Simultaneously, this crisis has brought us to realize the importance of low-cost, sensitive, and high-throughput devices that can be deployed in POC settings. On-site analysis that is fast, reliable, and helps to reduce the economic costs of infection transmission and potential quarantine is required. Different rapid POC tests have been authorized and deployed for mass screening and diagnostic purposes. Yet, RT-PCR has remained the primary and the only method for COVID-19 confirmation. Miniaturized PCR and PCR biosensors, devices that integrate PCR with different detection modalities, have emerged as tools that can address the issue of the low sensitivity of the current rapid POC tests and simultaneous analysis of samples in a high-throughput manner outside of a centralized lab. Digital PCR has emerged as an efficient high-throughput system. However, it does not eliminate the use of expensive reagents and often requires professional involvement in its operation. Electrochemical PCR is also a viable option for faster, cost-effective, and sensitive COVID-19 detection. However, the difficulty of the integration of PCR with electrochemical systems still creates formidable challenges in realizing a commercially adaptable system. CRISPR/Cas-based systems have further created a scope for diagnostic devices that do not require RNA extraction and amplification before detection.

The active transition from routine diagnostic laboratories to the realm of high sensitivity molecular diagnostics can significantly increase the efficiency and responsiveness of POCTs and facilitate the management of outbreaks in difficult settings. Devices such as those mentioned above can readily aid healthcare professionals in making faster medical decisions. However, there are still limitations to be addressed in such systems. Sample preparation errors and false positives and negatives need to be addressed before these assays can eventually be used for other diagnostic applications as well. Although different formats of POC RT-PCR assays have emerged, there is still scope for the development of hybrid, integrated systems that have better performance in terms of specificity and response time. Rigorous validation protocols and a high sampling rate would determine whether these devices are capable of use in the long run.

Author Contributions: Conceptualization, B.D.M., J.H.T.L., and N.G.; writing—original draft writing, N.G., S.A., and T.N.; writing—review and editing, B.D.M., J.H.T.L., A.D., D.K., and A.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: N.G., S.A., A.D., D.K., and B.D.M. acknowledge Yogesh Singh, Vice-Chancellor, Delhi Technological University, Delhi, India, for providing necessary facilities. N.G. and S.A. thank Delhi Technological University, Delhi, India; and the Council of Scientific and Industrial Research (CSIR; 08/133/(0013)/2018-EMR-I), India, respectively, for a fellowship award. T.N. is thankful for funding from the European Union's Horizon 2020 Research & Innovation Programme under the

Marie Skłodowska-Curie grant agreement no.: H2020-MSCA-ITN-813680. B.D.M. thanks the Science & Engineering Research Board (SERB), Govt. of India, for the award of a Distinguished Fellowship (SB/DF/011/2019).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. COVID-19 Weekly Epidemiological Update, 9 March 2021; World Health Organization: Geneva, Switzerland, 2021.
2. Arnaut, R.; Lee, R.A.; Lee, G.R.; Callahan, C.; Yen, C.F.; Smith, K.P.; Arora, R.; Kirby, J.E. SARS-CoV2 testing: The limit of detection matters. *bioRxiv* **2020**. [\[CrossRef\]](#)
3. Grant, B.D.; Anderson, C.E.; Williford, J.R.; Alonzo, L.F.; Glukhova, V.A.; Boyle, D.S.; Weigl, B.H.; Nichols, K.P. SARS-CoV-2 coronavirus nucleocapsid antigen-detecting half-strip lateral flow assay toward the development of point of care tests using commercially available reagents. *Anal. Chem.* **2020**, *92*, 11305–11309. [\[CrossRef\]](#)
4. Ragnosola, B.; Jin, D.; Lamb, C.C.; Shaz, B.H.; Hillyer, C.D.; Luchsinger, L.L. COVID19 antibody detection using lateral flow assay tests in a cohort of convalescent plasma donors. *BMC Res. Notes* **2020**, *13*, 1–7. [\[CrossRef\]](#)
5. Zhu, X.; Wang, X.; Han, L.; Chen, T.; Wang, L.; Li, H.; Li, S.; He, L.; Fu, X.; Chen, S. Multiplex reverse transcription loop-mediated isothermal amplification combined with nanoparticle-based lateral flow biosensor for the diagnosis of COVID-19. *Biosens. Bioelectron.* **2020**, *166*, 112437. [\[CrossRef\]](#)
6. Huang, C.; Wen, T.; Shi, F.-J.; Zeng, X.-Y.; Jiao, Y.-J. Rapid detection of IgM antibodies against the SARS-CoV-2 virus via colloidal gold nanoparticle-based lateral-flow assay. *ACS Omega* **2020**, *5*, 12550–12556. [\[CrossRef\]](#)
7. Cai, X.-f.; Chen, J.; Hu, J.-l.; Long, Q.-x.; Deng, H.-j.; Liu, P.; Fan, K.; Liao, P.; Liu, B.-z.; Wu, G.-c. A peptide-based magnetic chemiluminescence enzyme immunoassay for serological diagnosis of coronavirus disease 2019 (COVID-19). *J. Infect. Dis.* **2020**, *222*, 189–193. [\[CrossRef\]](#)
8. Padoan, A.; Cosma, C.; Sciacovelli, L.; Faggian, D.; Plebani, M. Analytical performances of a chemiluminescence immunoassay for SARS-CoV-2 IgM/IgG and antibody kinetics. *Clin. Chem. Lab. Med.* **2020**, *58*, 1081–1088. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Yu, L.; Wu, S.; Hao, X.; Dong, X.; Mao, L.; Pelechano, V.; Chen, W.-H.; Yin, X. Rapid detection of COVID-19 coronavirus using a reverse transcriptional loop-mediated isothermal amplification (RT-LAMP) diagnostic platform. *Clin. Chem.* **2020**, *66*, 975–977. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Basu, A.; Zinger, T.; Inglima, K.; Woo, K.-m.; Atie, O.; Yurasits, L.; See, B.; Agüero-Rosenfeld, M.E. Performance of Abbott ID Now COVID-19 rapid nucleic acid amplification test using nasopharyngeal swabs transported in viral transport media and dry nasal swabs in a New York City academic institution. *J. Clin. Microbiol.* **2020**, *58*. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Pray, I.W. Performance of an Antigen-Based Test for Asymptomatic and Symptomatic SARS-CoV-2 Testing at Two University Campuses—Wisconsin, September–October 2020; Centers for Disease Control and Prevention: Atlanta, GA, USA, 2021.
12. Wang, X.; Yao, H.; Xu, X.; Zhang, P.; Zhang, M.; Shao, J.; Xiao, Y.; Wang, H. Limits of detection of 6 approved RT-PCR kits for the novel SARS-coronavirus-2 (SARS-CoV-2). *Clin. Chem.* **2020**, *66*, 977–979. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Babady, N.E.; McMillen, T.; Jani, K.; Viale, A.; Robilotti, E.V.; Aslam, A.; Diver, M.; Sokoli, D.; Mason, G.; Shah, M.K. Performance of severe acute respiratory syndrome coronavirus 2 real-time RT-PCR tests on oral rinses and saliva samples. *J. Mol. Diagn.* **2021**, *23*, 3–9. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Patel, M.R.; Carroll, D.; Ussery, E.; Whitham, H.; Elkins, C.A.; Noble-Wang, J.; Rasheed, J.K.; Lu, X.; Lindstrom, S.; Bowen, V. Performance of Oropharyngeal Swab Testing Compared with Nasopharyngeal Swab Testing for Diagnosis of Coronavirus Disease 2019—United States, January 2020–February 2020. *Clin. Infect. Dis.* **2021**, *72*, 482–485. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Zhou, Y.; Pei, F.; Ji, M.; Wang, L.; Zhao, H.; Li, H.; Yang, W.; Wang, Q.; Zhao, Q.; Wang, Y. Sensitivity evaluation of 2019 novel coronavirus (SARS-CoV-2) RT-PCR detection kits and strategy to reduce false negative. *PLoS ONE* **2020**, *15*, e0241469. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Arevalo-Rodriguez, I.; Buitrago-Garcia, D.; Simancas-Racines, D.; Zambrano-Achig, P.; Del Campo, R.; Ciapponi, A.; Sued, O.; Martinez-Garcia, L.; Rutjes, A.W.; Low, N. False-negative results of initial RT-PCR assays for COVID-19: A systematic review. *PLoS ONE* **2020**, *15*, e0242958. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Tahamtan, A.; Ardebili, A. Real-time RT-PCR in COVID-19 detection: Issues affecting the results. *Expert Rev. Mol. Diagn.* **2020**, *20*, 453–454. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Singanayagam, A.; Patel, M.; Charlett, A.; Bernal, J.L.; Saliba, V.; Ellis, J.; Ladhani, S.; Zambon, M.; Gopal, R. Duration of infectiousness and correlation with RT-PCR cycle threshold values in cases of COVID-19, England, January to May 2020. *Euro Surveill.* **2020**, *25*, 2001483. [\[CrossRef\]](#)
19. Wee, S.K.; Sivalingam, S.P.; Yap, E.P.H. Rapid direct nucleic acid amplification test without RNA extraction for SARS-CoV-2 using a portable PCR thermocycler. *Genes* **2020**, *11*, 664. [\[CrossRef\]](#)
20. Gibani, M.M.; Toumazou, C.; Sohbat, M.; Sahoo, R.; Karvela, M.; Hon, T.-K.; De Mateo, S.; Burdett, A.; Leung, K.F.; Barnett, J. Assessing a novel, lab-free, point-of-care test for SARS-CoV-2 (CovidNudge): A diagnostic accuracy study. *Lancet Microbe* **2020**, *1*, e300–e307. [\[CrossRef\]](#)
21. Gupta, N.; Rana, S.; Singh, H. Innovative point-of-care molecular diagnostic test for COVID-19 in India. *Lancet Microbe* **2020**, *1*, e277. [\[CrossRef\]](#)
22. Accula SARS-CoV-2 Test-Letter of Authorization; The U.S. Food and Drug Administration: Silver Spring, MD, USA, 2021.

23. In Vitro Diagnostics EUAs. Available online: <https://www.fda.gov/medical-devices/coronavirus-disease-2019-covid-19-emergency-use-authorizations-medical-devices/vitro-diagnostics-euas> (accessed on 10 February 2021).
24. Vogelstein, B.; Kinzler, K.W. Digital PCR. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 9236–9241. [[CrossRef](#)]
25. Quan, P.-L.; Sauzade, M.; Brouzes, E. dPCR: A technology review. *Sensors* **2018**, *18*, 1271. [[CrossRef](#)] [[PubMed](#)]
26. Hindson, B.J.; Ness, K.D.; Masquelier, D.A.; Belgrader, P.; Heredia, N.J.; Makarewicz, A.J.; Bright, I.J.; Lucero, M.Y.; Hiddessen, A.L.; Legler, T.C. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal. Chem.* **2011**, *83*, 8604–8610. [[CrossRef](#)] [[PubMed](#)]
27. White, R.A.; Blainey, P.C.; Fan, H.C.; Quake, S.R. Digital PCR provides sensitive and absolute calibration for high throughput sequencing. *BMC Genomics* **2009**, *10*, 1–12.
28. Tan, C.; Fan, D.; Wang, N.; Wang, F.; Wang, B.; Zhu, L.; Guo, Y. Applications of digital PCR in COVID-19 pandemic. *View* **2021**, *2*. [[CrossRef](#)]
29. Nykel, A.; Kaszkowiak, M.; Fendler, W.; Gach, A. Chip-based digital PCR approach provides a sensitive and cost-effective single-day screening tool for common fetal aneuploidies—A proof of concept study. *Int. J. Mol. Sci.* **2019**, *20*, 5486. [[CrossRef](#)] [[PubMed](#)]
30. Dueck, M.E.; Lin, R.; Zayac, A.; Gallagher, S.; Chao, A.K.; Jiang, L.; Datwani, S.S.; Hung, P.; Stieglitz, E. Precision cancer monitoring using a novel, fully integrated, microfluidic array partitioning digital PCR platform. *Sci. Rep.* **2019**, *9*, 1–9. [[CrossRef](#)]
31. Wong, Y.C.; Lau, S.Y.; Wang To, K.K.; Mok, B.W.Y.; Li, X.; Wang, P.; Deng, S.; Woo, K.F.; Du, Z.; Li, C. Natural transmission of bat-like SARS-CoV-2_{ΔPRRA} variants in COVID-19 patients. *Clin. Infect. Dis.* **2020**. [[CrossRef](#)]
32. Suo, T.; Liu, X.; Feng, J.; Guo, M.; Hu, W.; Guo, D.; Ullah, H.; Yang, Y.; Zhang, Q.; Wang, X. ddPCR: A more accurate tool for SARS-CoV-2 detection in low viral load specimens. *Emerg. Microbes Infect.* **2020**, *9*, 1259–1268. [[CrossRef](#)] [[PubMed](#)]
33. Dong, L.; Zhou, J.; Niu, C.; Wang, Q.; Pan, Y.; Sheng, S.; Wang, X.; Zhang, Y.; Yang, J.; Liu, M. Highly accurate and sensitive diagnostic detection of SARS-CoV-2 by digital PCR. *Talanta* **2021**, *224*, 121726. [[CrossRef](#)] [[PubMed](#)]
34. Lu, R.; Wang, J.; Li, M.; Wang, Y.; Dong, J.; Cai, W. SARS-CoV-2 detection using digital PCR for COVID-19 diagnosis, treatment monitoring and criteria for discharge. *MedRxiv* **2020**. [[CrossRef](#)]
35. Alteri, C.; Cento, V.; Antonello, M.; Colagrossi, L.; Merli, M.; Ughi, N.; Renica, S.; Matarazzo, E.; Di Ruscio, F.; Tartaglione, L. Detection and quantification of SARS-CoV-2 by droplet digital PCR in real-time PCR negative nasopharyngeal swabs from suspected COVID-19 patients. *PLoS ONE* **2020**, *15*, e0236311. [[CrossRef](#)]
36. Liu, X.; Feng, J.; Zhang, Q.; Guo, D.; Zhang, L.; Suo, T.; Hu, W.; Guo, M.; Wang, X.; Huang, Z. Analytical comparisons of SARS-COV-2 detection by qRT-PCR and ddPCR with multiple primer/probe sets. *Emerg. Microbes Infect.* **2020**, *9*, 1175–1179. [[CrossRef](#)]
37. Yu, F.; Yan, L.; Wang, N.; Yang, S.; Wang, L.; Tang, Y.; Gao, G.; Wang, S.; Ma, C.; Xie, R. Quantitative detection and viral load analysis of SARS-CoV-2 in infected patients. *Clin. Infect. Dis.* **2020**, *71*, 793–798. [[CrossRef](#)]
38. Lv, J.; Yang, J.; Xue, J.; Zhu, P.; Liu, L.; Li, S. Detection of SARS-CoV-2 RNA residue on object surfaces in nucleic acid testing laboratory using droplet digital PCR. *Sci. Total Environ.* **2020**, *742*, 140370. [[CrossRef](#)]
39. Mio, C.; Cifù, A.; Marzinotto, S.; Bergamin, N.; Caldana, C.; Cattarossi, S.; Cmet, S.; Cussigh, A.; Martinella, R.; Zucco, J.; et al. A streamlined approach to rapidly detect SARS-CoV-2 infection avoiding RNA extraction: Workflow validation. *Dis. Markers* **2020**, *2020*. [[CrossRef](#)] [[PubMed](#)]
40. Ternovoi, V.; Lutkovsky, R.Y.; Ponomareva, E.; Gladysheva, A.; Chub, E.; Tupota, N.; Smirnova, A.; Nazarenko, A.; Loktev, V.; Gavrilova, E. Detection of SARS-CoV-2 RNA in nasopharyngeal swabs from COVID-19 patients and asymptomatic cases of infection by real-time and digital PCR. *Klin. Lab. Diagn.* **2020**, *65*, 785–792. [[CrossRef](#)]
41. Deiana, M.; Mori, A.; Piubelli, C.; Scarso, S.; Favara, M.; Pomari, E. Assessment of the direct quantitation of SARS-CoV-2 by droplet digital PCR. *Sci. Rep.* **2020**, *10*, 1–7. [[CrossRef](#)]
42. Liu, Y.; Ning, Z.; Chen, Y.; Guo, M.; Liu, Y.; Gali, N.K.; Sun, L.; Duan, Y.; Cai, J.; Westerdahl, D. Aerodynamic analysis of SARS-CoV-2 in two Wuhan hospitals. *Nature* **2020**, *582*, 557–560. [[CrossRef](#)] [[PubMed](#)]
43. Chen, Y.-X.; Zhang, W.-J.; Huang, K.-J.; Zheng, M.; Mao, Y.-C. An electrochemical microRNA sensing platform based on tungsten diselenide nanosheets and competitive RNA–RNA hybridization. *Analyst* **2017**, *142*, 4843–4851. [[CrossRef](#)]
44. Lynch III, C.A.; Foguel, M.V.; Reed, A.J.; Balcarcel, A.M.; Calvo-Marzal, P.; Gerasimova, Y.V.; Chumbimuni-Torres, K.Y. Selective Determination of Isothermally Amplified Zika Virus RNA Using a Universal DNA-Hairpin Probe in Less than 1 Hour. *Anal. Chem.* **2019**, *91*, 13458–13464. [[CrossRef](#)] [[PubMed](#)]
45. Feng, D.; Su, J.; He, G.; Xu, Y.; Wang, C.; Zheng, M.; Qian, Q.; Mi, X. Electrochemical DNA Sensor for Sensitive BRCA1 Detection Based on DNA Tetrahedral-Structured Probe and Poly-Adenine Mediated Gold Nanoparticles. *Biosensors* **2020**, *10*, 78. [[CrossRef](#)]
46. Santhanam, M.; Algov, I.; Alfonta, L. DNA/RNA electrochemical biosensing devices a future replacement of PCR methods for a fast epidemic containment. *Sensors* **2020**, *20*, 4648. [[CrossRef](#)]
47. Zhao, H.; Liu, F.; Xie, W.; Zhou, T.-C.; OuYang, J.; Jin, L.; Li, H.; Zhao, C.-Y.; Zhang, L.; Wei, J. Ultrasensitive supersandwich-type electrochemical sensor for SARS-CoV-2 from the infected COVID-19 patients using a smartphone. *Sens. Actuators B Chem.* **2021**, *327*, 128899. [[CrossRef](#)] [[PubMed](#)]
48. Alafeef, M.; Dighe, K.; Moitra, P.; Pan, D. Rapid, ultrasensitive, and quantitative detection of SARS-CoV-2 using antisense oligonucleotides directed electrochemical biosensor chip. *ACS Nano* **2020**, *14*, 17028–17045. [[CrossRef](#)]

49. Patterson, A.S.; Hsieh, K.; Soh, H.T.; Plaxco, K.W. Electrochemical real-time nucleic acid amplification: Towards point-of-care quantification of pathogens. *Trends Biotechnol.* **2013**, *31*, 704–712. [\[CrossRef\]](#)
50. Chaibun, T.; Puenpa, J.; Ngamdee, T.; Boonapatcharoen, N.; Athamanolap, P.; O'Mullane, A.P.; Vongpunsawad, S.; Poovorawan, Y.; Lee, S.Y.; Lertanantawong, B. Rapid electrochemical detection of coronavirus SARS-CoV-2. *Nat. Commun.* **2021**, *12*, 1–10. [\[CrossRef\]](#)
51. *ePlex®SARS-CoV-2 Test Assay Manual*; The United States Food and Drug Administration: Silver Spring, MD, USA, 2020.
52. Kumar, P.; Malik, Y.S.; Ganesh, B.; Rahangdale, S.; Saurabh, S.; Natesan, S.; Srivastava, A.; Sharun, K.; Yattoo, M.I.; Tiwari, R. CRISPR-Cas system: An approach with potentials for COVID-19 diagnosis and therapeutics. *Front. Cell Infect. Microbiol.* **2020**, *10*, 576875. [\[CrossRef\]](#) [\[PubMed\]](#)
53. Hou, T.; Zeng, W.; Yang, M.; Chen, W.; Ren, L.; Ai, J.; Wu, J.; Liao, Y.; Gou, X.; Li, Y. Development and evaluation of a rapid CRISPR-based diagnostic for COVID-19. *PLoS Pathog.* **2020**, *16*, e1008705. [\[CrossRef\]](#)
54. Ning, B.; Yu, T.; Zhang, S.; Huang, Z.; Tian, D.; Lin, Z.; Niu, A.; Golden, N.; Hensley, K.; Threeton, B. A smartphone-read ultrasensitive and quantitative saliva test for COVID-19. *Sci. Adv.* **2021**, *7*, eabe3703. [\[CrossRef\]](#)
55. Fozouni, P.; Son, S.; de León Derby, M.D.; Knott, G.J.; Gray, C.N.; D'Ambrosio, M.V.; Zhao, C.; Switz, N.A.; Kumar, G.R.; Stephens, S.I. Amplification-free detection of SARS-CoV-2 with CRISPR-Cas13a and mobile phone microscopy. *Cell* **2021**, *184*, 323–333.e329. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Ouyang, W.; Han, J. One-step nucleic acid purification and noise-resistant polymerase chain reaction by electrokinetic concentration for ultralow-abundance nucleic acid detection. *Ang. Chem.* **2020**, *132*, 11074–11081. [\[CrossRef\]](#)
57. Ramachandran, A.; Huyke, D.A.; Sharma, E.; Sahoo, M.K.; Huang, C.; Banaei, N.; Pinsky, B.A.; Santiago, J.G. Electric field-driven microfluidics for rapid CRISPR-based diagnostics and its application to detection of SARS-CoV-2. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 29518–29525. [\[CrossRef\]](#) [\[PubMed\]](#)
58. Lee, S.H.; Song, J.; Cho, B.; Hong, S.; Hoxha, O.; Kang, T.; Kim, D.; Lee, L.P. Bubble-free rapid microfluidic PCR. *Biosens. Bioelectron.* **2019**, *126*, 725–733. [\[CrossRef\]](#)
59. McArdle, H.; Jimenez-Mateos, E.M.; Raoof, R.; Carthy, E.; Boyle, D.; ElNaggar, H.; Delanty, N.; Hamer, H.; Dogan, M.; Huchtemann, T.; et al. "TORNADO"—Theranostic One-Step RNA Detector; microfluidic disc for the direct detection of microRNA-134 in plasma and cerebrospinal fluid. *Sci. Rep.* **2017**, *7*, 1–11. [\[CrossRef\]](#) [\[PubMed\]](#)
60. Azhar, M.; Phutela, R.; Ansari, A.H.; Sinha, D.; Sharma, N.; Kumar, M.; Aich, M.; Sharma, S.; Singhal, K.; Lad, H.; et al. Rapid, field-deployable nucleobase detection and identification using FnCas9. *bioRxiv* **2020**. [\[CrossRef\]](#)
61. Deng, H.; Zhou, X.; Liu, Q.; Li, B.; Liu, H.; Huang, R.; Xing, D. Paperfluidic chip device for small RNA extraction, amplification, and multiplexed analysis. *ACS Appl. Mater. Interfaces* **2017**, *9*, 41151–41158. [\[CrossRef\]](#)



Prevalence and risk analysis of fluoride in groundwater around sandstone mine in Haryana, India

Saurav Kumar Ambastha¹ · A. K. Haritash¹

Received: 30 December 2020 / Accepted: 4 May 2021
© Accademia Nazionale dei Lincei 2021

Abstract

Groundwater contamination by fluoride is a typical problem associated with most of the regions in India. Mining of minerals can accelerate the dissolution of fluoride resulting in the further contamination of groundwater resources. The present study was undertaken to determine the concentration of fluoride in groundwater around the Bakhrija sandstone mine located in Haryana state, India. It was observed that the groundwater in immediate vicinity of the mine had relatively higher level of dissolved fluoride. The risk associated with consumption of fluoride contaminated groundwater was also observed to be higher in villages adjacent to the mines. The geochemical investigation suggested that dissolution of carbonate minerals may have resulted in solubilisation of fluoride in groundwater through the process of ion-exchange. The study concluded that fluoride level may rise in the other nearby regions if the intensity of mining increases. It may result in further spread of fluoride to other aquifers located around Bakhrija mine, if suitable environmental management plan is not developed.

Keywords Fluoride · Risk analysis · Groundwater · Sandstone · Mining

1 Introduction

Fluoride in groundwater is one among the major pollutants that can affect human health adversely. Since fluorine is an abundantly present element in earth's crust, it is prevalent as dissolved fluoride in groundwater around the globe. Whereas 0.7–1.0 mg/l of fluoride in drinking water is essential to prevent dental cavities and tooth decay, excess of fluoride (≥ 1.5 mg/l) may result in dental and skeletal fluorosis. Although the target organ of fluoride is bones, it is also known to interfere with brain development in children, reduced IQ (Xu et al. 2020), hypothyroidism, hyperglycaemia, infertility (Dey and Giri 2016), and osteosarcoma (Cohn 1992). The accumulation of fluoride in human body may result due to exposure through drinking water, fluoride-rich milk (Ullah et al. 2017), meat, tobacco (Yadav et al. 2007), dentifrice (Kanduti et al. 2016), and other food materials (Fein and Cerklewski 2001). There are a number of reports on fluoride exposure and risk analysis through food or drinking

water, but most of the reports investigate the adverse effects related to fluoride-rich drinking water. Fluorite (CaF_2) and/or fluorapatite ($\text{Ca}_5(\text{PO}_4)_3\text{F}$) present as natural minerals in soil react with ground water, thereby resulting in contamination of drinking water, particularly in rural and remote areas (Haritash et al. 2018).

Most of the sources of groundwater in India have relatively higher concentration of fluoride since geographical distribution of fluoride-rich mineral is higher in Indian soil. The states like Andhra Pradesh (Adimalla et al. 2019), Rajasthan (Arif et al. 2013), Gujarat (Gupta et al. 2005), and Madhya Pradesh (Avtar et al. 2013) represent exceedance of fluoride level (> 1.0 mg/l), but the other states like Jharkhand (Pandey et al. 2012), Bihar (Kumar et al. 2018), Uttar Pradesh (Ali et al. 2017), and Haryana (Haritash et al. 2008) also represent dispersed pockets of fluoride-rich groundwater. Studies have revealed that almost 80 percent of total fluoride accumulated in human body (mg/kg/day) is through drinking water. Therefore, assessment of health risk associated with fluoride-rich ground water is pertinent. Out of the total fluoride ingested, about 60% is absorbed; while the absorption on empty stomach is about 100% (WHO 2004). The rate of dissolution of fluoride from soil increases if the conditions are acidic or time of soil–water interaction is more. Such conditions are found due to the

✉ Saurav Kumar Ambastha
saurav.ambastha@gmail.com

¹ Department of Environmental Engineering, Delhi Technological University, Shahbad Daultapur, Delhi 110042, India

release of acid-mine drainage and are seldom observed in mining areas. Sometimes, the collection of surface runoff in mine pits results in enhanced soil water interaction and higher percolation as well. Different mining operations (drilling and blasting) may result in formation of vertical cracks in subsurface impervious rock stratum resulting in an easy introduction of contaminants into the ground water (Armiento et al. 2016). Sandstone mining involves the use of heavy mechanical drills and explosives to fracture/fragment the rocks. Since fluoride is associated with such geological material, its interaction with outside environment increases upon excavation. It has been reported that the groundwater around sand stone mining remains contaminated by nitrate (NO_3^-), fluoride (F^-), or pathogens (Kumar et al. 2017). Therefore, the present study was undertaken to determine fluoride level in groundwater around sandstone quarries located in Mahendragarh district of Haryana state, India. Further, the exposure assessment and risk were calculated as per the methodology suggested by USEPA (1993).

2 Materials and methods

The present study was undertaken in Mahendragarh district, Haryana, India. The district has area of 1939 Km^2 with total population of 922,088 (Census of India 2011). Climate of Mahendragarh district is hot in summers and cold in winters with unevenly distributed rainfall of 500 mm during monsoon. Mahendragarh district has nine major mining sites and Bakhrija stone mine is the largest with an area of about 66 km^2 . The location of Bakhrija mines is between $27^\circ 55' 1''$ and $27^\circ 54' 6''$ North latitude and $76^\circ 03' 28.34''$ to $76^\circ 03' 27.56''$ East longitude and it is famous for quarrying of calcite, limestone, and mica as chief minerals. Bakhrija stone mines are surrounded by Bakhrija, Dholera, Meghot Binja, Nujota, Meghot Halla and Khojpur Naglia villages (Fig. 1). In the present study, a total number of fifteen (15) groundwater samples were collected from bore wells located around the mining area. Samples were collected in pre-rinsed fresh Polypropylene bottles of 1.0 L capacity each. The samples were stored at low temperature in an ice box, and transported to the laboratory within 6 h for their chemical analysis. The samples were characterized for different parameters, in triplicates, following the standard methods as prescribed by APHA (2012). The fluoride concentration

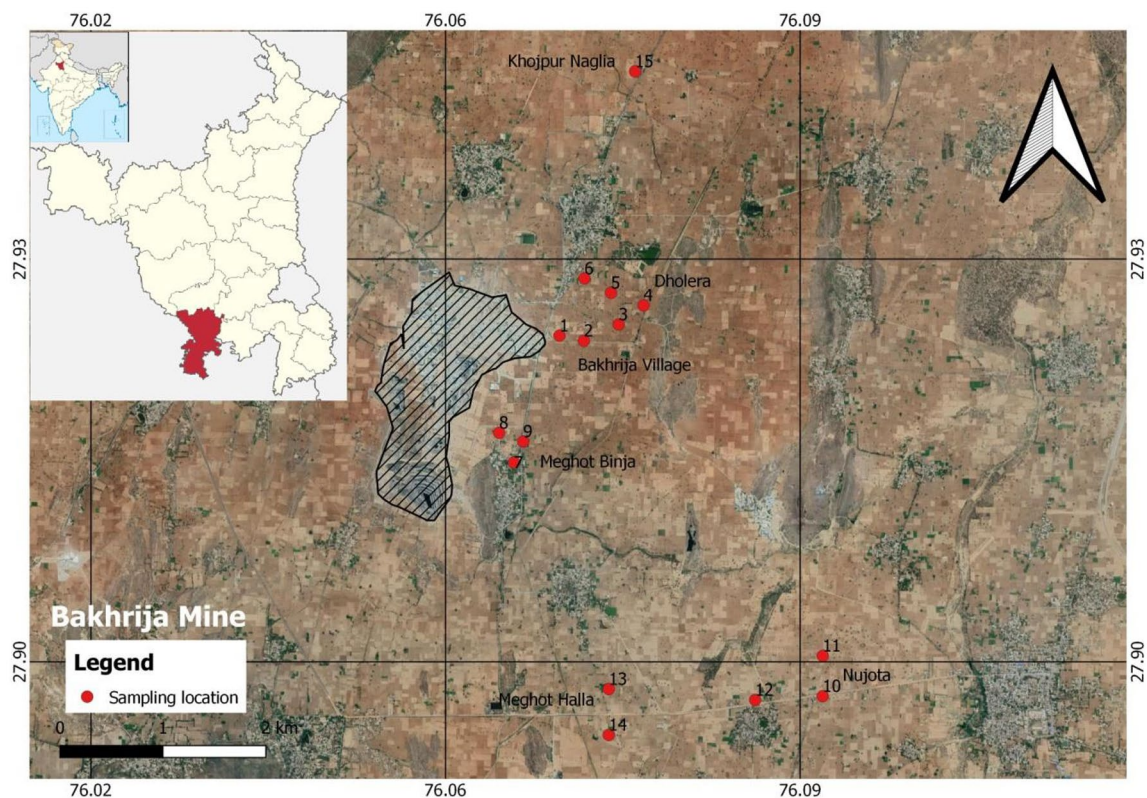


Fig. 1 Location of Bakhrija Sandstone mine and borewells for collection of groundwater samples

was determined using an ion-specific electrode (ISE—Orion Scientific, USA). Later, the exposure assessment and risk analysis were performed using the methodology as given by USEPA (Eq. 1). The suitability of sampled ground water sources was also evaluated comparing the observed values against standard prescribed values of Bureau of Indian Standards (BIS 2012) and WHO (2004).

2.1 Fluoride and human health

To estimate the risk and content of adverse effects of fluoride on human health, the procedure is divided into four phases (Selinus et al. 2016). These four phases are (i) Identification of hazard, (ii) Value selection on toxicity reference, (iii) Assessment of exposure and (iv) Characterisation of risk. Although there are alternate ways of exposure to fluoride i.e. drinking water, intake of other beverages and food, toothpaste, tea, pan masala and tobacco, etc. (Yadav et al. 2007), but drinking water is the prominent source of fluoride. For the study area, the health risk associated with the daily intake of fluoride-rich water was estimated using the following formula

$$\text{Chronic daily intake (CDI)} = [\text{C}_w * \text{IR} * \text{EF} * \text{ED}] / \text{BW} * \text{AT} \quad (1)$$

where CDI is chronic daily intake of fluoride through drinking water (mg/Kg/day), C_w is the fluoride concentration in drinking water (mg/l), IR is ingestion rate of drinking water (3 L/day), EF is frequency exposure (365 day/year) and ED is exposure period (70 years), BW is average weight of the body (60 kg) and AT is average time of exposure (365×70 Days).

The non-carcinogenic risk to human health from fluoride toxicity is determined with the use of formula for hazard quotient (HQ).

$$\text{HQ Fluoride} = \text{CDI} / \text{RfD} \quad (2)$$

RfD indicates the reference fluoride dosage in the formula above in mg/kg/day. RfD is used to determine fluoride's risk to health during a defined pathway of exposure. According to USEPA's Integrated Risk Information System (IRIS), RfD for drinking water is 0.05 mg/kg/day. The HQ value less than one is considered safe, while the HQ more than unitary value has potential possibility of non-carcinogenic health effects that can arise due to the consumption of water contaminated with fluoride.

3 Results and discussion

Based on the physico-chemical characterisation, it is noticed that the groundwater contains most of the cations and anions at or above measurable concentration (Table 1). All the samples of groundwater were observed to be slightly alkaline with respect to pH. Since natural minerals with basic nature (calcium, magnesium, carbonate and bicarbonate) dominantly get dissolved, the pH of ground water is generally alkaline. Based on TDS, most of the samples (93%) were found to be exceeding the prescribed limit of 500 mg/l; while based on total hardness, all the collected samples were classified as hard. Similar to this, all the collected samples exceeded the prescribed limit for calcium and magnesium, thus, confirming the hard nature of the groundwater. Unlike calcium and magnesium, chloride and sulphate exceeded the specified limit in 33% of samples collected, indicating that the hardness was dominantly contributed by carbonate and bicarbonate salts of the cations. Nitrate was within the permissible limit (< 45 mg/l) in all the ground water samples indicating that anthropogenic addition through fertilizer or waste water disposal is not resulting in ground water contamination in the study area. Fluoride is another important

Table 1 Physico-Chemical characteristics, suitability for drinking, and potential effects of fluoride in groundwater in the study area

Parameters	Minimum	Maximum	Mean \pm SD	Desirable limit	Exceedance (%)	Potential effect
pH	7.1	8.6	7.6 ± 0.36	6.5–8.5	–	Taste, corrosion
TDS (mg/l)*	366	2610	1083 ± 541	500	93*	Gastrointestinal irritation
TH (mg/l)	300	3020	886 ± 688	300	100	Kidney stones
NO_3^- (mg/l)	6	18	11 ± 5	45	0	Methaemoglobin-aemia
SO_4^{2-} (mg/l)	33	597	129 ± 140	150	33	Laxative effect
Cl^- (mg/l)	39	900	231 ± 229	250	33	Anaesthetic effect, Salty taste
K^+ (mg/l)*	3	22	8 ± 5	–	–	Bitter taste
F^- (mg/l)	0.5	11	2.5 ± 2.7	1.0	66	Dental and Skeletal fluorosis
Ca^{2+} (mg/l)	80	344	117 ± 55	75	100	Scale formation
Mg^{2+} (mg/l)	196	2676	769 ± 648	30	100	Nausea and vomiting
Na^+ (mg/l)*	159	681	326 ± 150	30–60	100*	Hypertensive effects

*As per WHO (2004); TH Total hardness as CaCO_3 ; TDS Total dissolved solids

anion which may induce potential health effects over the exposed population. It was observed to be exceeding the level of 1.0 mg/l in 66% of the collected samples of ground water. Relatively higher values of fluoride at some locations are a cause of concern considering its toxicity and its health effect.

3.1 Fluoride exposure and health implication

Health risk assessment is important for determining the extent and possibility of health consequences over the people living in the region since they are vulnerable to life-threatening contaminants in drinking water. The oral ingestion of excess fluoride through drinking water plays a crucial role and can pose a non-carcinogenic health risk to

the population. Accordingly, the non-carcinogenic risk of fluoride to human health is estimated in terms of daily intake or hazard quotient (Table 2). Hazard quotient of the fluoride for the collected samples lies between 0.4 and 8.8 and the mean F^- concentration of the all collected samples is 2 mg/l which is higher than the permissible limit of BIS (2012).

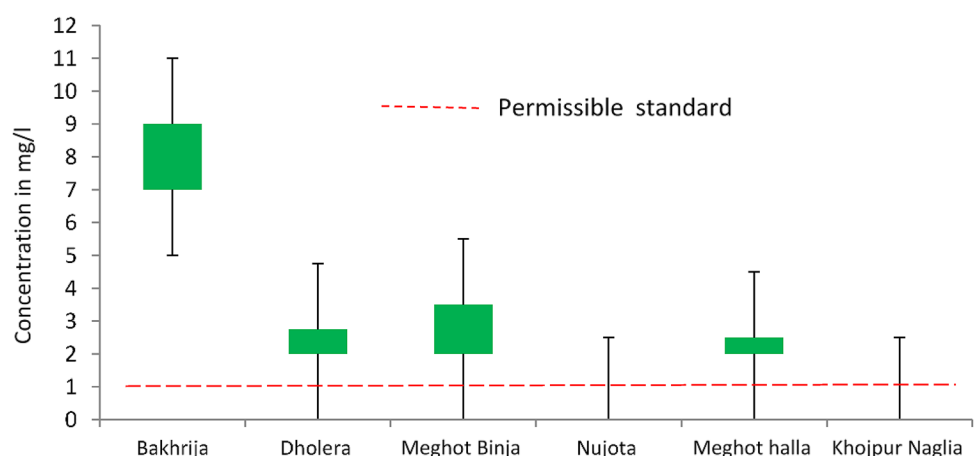
Fluoride concentration more than 1.5 mg/l has been found in groundwater of Bakhrija village (Fig. 2) with maximum concentration of 11 mg/l and significantly high value of HQ. Bakhrija is the nearest village from the mining area. Other villages viz. Dholera, Meghot Binja, and Meghot Halla represented groundwater fluoride concentration between 2.5 and 4.0 mg/l; and maximum concentration of 5.0 mg/l (Meghot Binja). Hazard quotient was in between 0.8 and 4.0 for these villages. Nujota and Khojpur Nagalia represented

Table 2 Prevalence of fluoride and associated health risk in groundwater of different villages around Bhakrija mine, Mahendragarh

S. No	Village	pH	TDS	Risk assessment			
				F^- (mg/L)	CDI (Water)	HQ	Inference
1	Bakhrija	8.6	340	11	0.44	8.8	Severe dental and skeletal fluorosis
2		7.4	440	3	0.12	2.4	
3	Dholera	7.7	720	2	0.08	1.6	Dental and skeletal Fluorosis on prolonged exposure
4		7.9	300	5	0.2	4	
5		7.8	760	2	0.08	1.6	
6		7.6	840	2	0.08	1.6	Dental and skeletal fluorosis on prolonged exposure
7	Meghot Binja	7.6	920	2	0.08	1.6	
8		7.3	1020	2	0.08	1.6	
9		8.0	320	5	0.2	4	No Effect
10	Nujota	7.4	660	0.5	0.02	0.4	
11		7.5	1280	0.5	0.02	0.4	
12		7.0	1740	0.5	0.02	0.4	No Effect
13	Meghot Halla	7.1	640	1	0.04	0.8	
14		7.6	880	3	0.12	2.4	Dental and skeletal fluorosis
15	Khojpur Naglia	7.6	3020	0.5	0.02	0.4	No Effect

CDI Chronic daily intake; HQ Hazard quotient

Fig. 2 Spatial variation of fluoride in groundwater around Bakhrija mine, Mahendragarh



groundwater fluoride concentration within the permissible range. Human population living in the villages where the concentration of fluoride is recorded above the permissible limit is more likely to be exposed to potential health risks.

Fluorosis is a chronic disease involving the human population and is exacerbated through the intake of a higher concentration of fluoride through food and drink (Nuccio 2016). Children are more susceptible to higher fluoride as compared to adults. Lower body weight than adults could be the cause of higher risk from the exposure of fluoride (Kumar et al. 2016). Intake of fluoride (0.5–1.0 mg/l) is very important in early phase of life as it can stop dental caries, but higher level (> 1.5 mg/l) can lead to dental and skeletal fluorosis. Fluoride consumption for the first three years of life is the most important in fluorosis aetiology (Levy et al. 2002).

The groundwater of the study area was classified into three classes viz. 0–1 mg/l as safe; 1–4 mg/l causing dental fluorosis; and more than 5 mg/l causing skeletal fluorosis. About 33% samples fall in Class-I and can be considered safe for drinking; while 46% and 20% of groundwater samples come under Class II and Class III, respectively, and can cause dental and skeletal fluorosis (Fig. 3).

Chronic intake of excessive F^- (i.e. 1.5–4.0 mg/l) can lead to fluorosis of the enamel and bone, and in severe cases (i.e. 4–10 mg/l), skeletal fluorosis associated with joint weakness, ligament calcification, and some osteosclerosis of the pelvis and vertebrae may be observed (Liang et al. 2017; Narsimha and Sudarshan 2016; Podgamy and McLaren 2015). This occurs mostly because F^- is highly electronegative and has a comparable ionic radius (133 pm) to that of hydroxyl ion (140 pm), which contributes to hydrogen fluoride formation (Kumar et al. 2016). Fluoride in the human body, in fact, easily diffuses through the intestines, dissolves in the blood and accumulates in calcified tissues (Dey and Giri 2016). Health-related problems in and around the areas are primarily non-carcinogenic in nature, especially in the areas examined, where fluoride in drinking water does not reach 10 mg/l. However, higher doses (> 10 mg/l) can be correlated

with debilitating fluorosis and carcinogenic risk (Ali et al. 2019). Confined studies on this aspect indicate that fluoride may allow cells to develop faster enough that will become cancerous over time, but it is controversial since there is no reliable correlation between fluoride and the influence of carcinogenicity (Bajpai 2013).

3.2 Genesis of groundwater and contamination

Based on the concentration (in meq/l) of selective dominant anions (Cl^- and SO_4^{2-}) and a cation (Na^+), the prevailing dominant soil water interactions in the study area were identified based on the base-exchange indices. The classification of groundwater was done using the following equation (all units are in meq/l).

Base Exchange (base exch)

$$= Na^+ - Cl^- / SO_4^{2-} \text{ meq/l (Matthess 1982)} \quad (3)$$

As stated above (Eq. 3), the base-exchange values more than unitary (> 1) suggest that groundwater is $NaHCO_3^-$ type; on the opposite, base exchange less than unitary (< 1), the groundwater represents $Na^+-SO_4^{2-}$ type. The base exchange index plot shows that most of the samples (87%) belong to the $Na^+-SO_4^{2-}$ type, while only 13% belongs to the $Na^+-HCO_3^-$ type in the study area. It is well known that the $Na^+-SO_4^{2-}$ form of water accelerates the carbonate mineral deposition of calcite, and is linked with the release of fluoride from minerals in gneissic basement rocks and granite accumulation due to the dissolution of silicates and, therefore, a rise in the concentration of fluoride in groundwater. $NaF(s)$ dissolves to release $Na^+(aq)$, a conjugate base of strong acid fluoride that does not react with water. When the salt, NaF , is dissolved in water, the F^- ion is formed (Chitrakshi and Haritash 2018; Mamatha and Rao 2010). The ion-exchange reactions followed by the mineral

Fig. 3 Classification of groundwater sample with respect to risk of fluorosis in village around Bakhrija mine, Mahendragarh

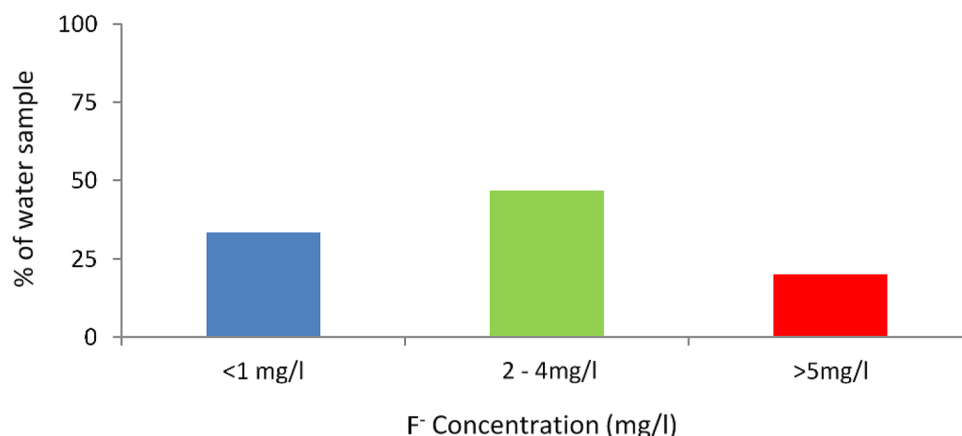
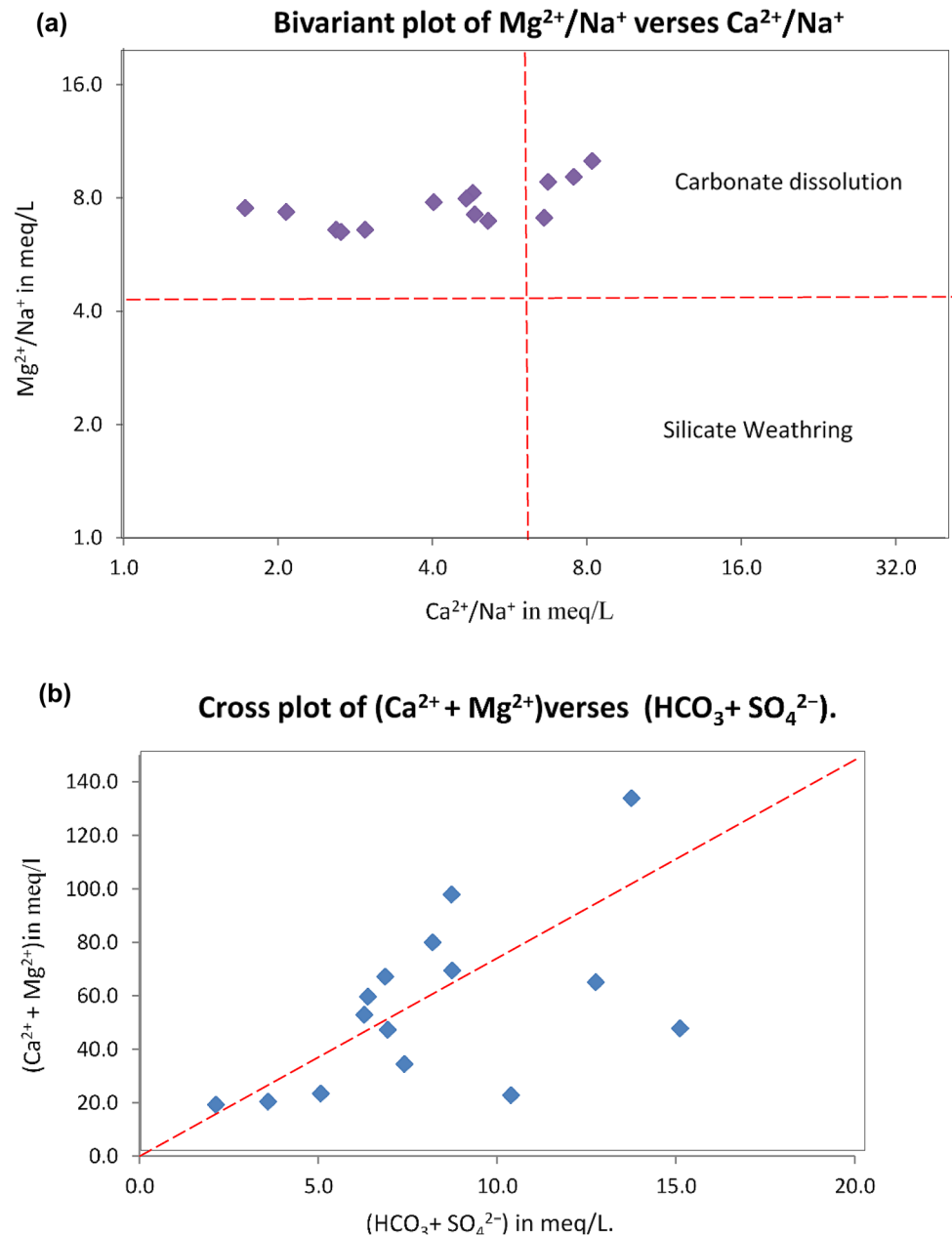
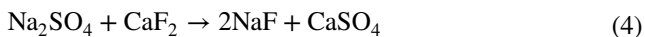


Fig. 4 **a** Bivariant plot of $\text{Mg}^{2+}/\text{Na}^+$ versus $\text{Ca}^{2+}/\text{Na}^+$. **b** Cross plot of $(\text{Ca}^{2+} + \text{Mg}^{2+})$ versus $(\text{HCO}_3^- + \text{SO}_4^{2-})$



reaction are key factors in the study area that are responsible for a high level of fluoride.



Carbonate-rich rocks, such as limestone and dolomite, are major starting materials for carbonate weathering. The carbonates present in these rocks are dissolved in groundwater during water infiltration. Calcium (Ca^{2+}), magnesium (Mg^{2+}), their molar ratio (Ca/Mg), and hydrogen carbonate (HCO_3^-) are the major chemical parameters describing the groundwater carbonate equilibrium. Generally, the molar ratio in groundwater between calcium and magnesium depends on the lithological composition of groundwater

recharge areas, i.e., if Ca/Mg molar ratio is equal to 1, it poses presence of dolomite while the higher molar ratio suggests dissolution of calcite minerals. The bivariant plot (Fig. 4) of $\text{Mg}^{2+}/\text{Na}^+$ versus $\text{Ca}^{2+}/\text{Na}^+$ clearly shows that carbonate dissolution is the dominant process which contributes to the chemical quality of the groundwater in the study area. Once again, the cross plot of versus $(\text{Ca}^{2+} + \text{Mg}^{2+})$ ($\text{HCO}_3^- + \text{SO}_4^{2-}$) (Fig. 4b) is observed and implies that most of the samples with more than one fluoride concentration are found above equiline, which further suggests the carbonate dissolution, weathering and the movement of ions in the region are responsible for higher F^- and HCO_3^- concentrations in the groundwater. It is well known that an

increase in pH (*i.e.*, alkaline condition), sodium, and bicarbonate ion concentrations eventually raises the concentration of fluoride in groundwater as a result of the above reactions and mechanisms. In general, the longer interaction of rock water implies the weathering of fluoride-bearing minerals under alkaline conditions resulting in higher concentrations of fluoride in groundwater (Raj and Shaji 2017; Adimalla and Venkatayogi 2017; Cremisini and Armiento 2016).

4 Conclusion

The study indicates that groundwater adjacent to the mining area in Mahendragarh is rich in fluoride especially in the villages located in southeast direction. The exposure assessment study reveals high exposure in these regions resulting in the high risk of non-carcinogenic effects due to fluoride. The geochemical analysis reveals presence of silicate weathering in areas with higher fluoride level indicating that natural processes are resulting in dissolution of fluoride in groundwater and ion exchange is dominantly responsible. Further, it is important to mention that contamination of groundwater by fluoride may increase with time and with more intense rate of quarrying. It is recommended for regular monitoring of groundwater in the villages around mining area so that any possibility of fluoride contamination is timely noticed and checked.

Acknowledgements The authors acknowledge the help of Mr. Rajesh Sehrawat, Mine inspector, Govt. of Haryana in several ways during this study.

Authors' contributions SKA: Conceptualization, Methodology, Software, SKA: Data curation, Writing- Original draft preparation. AKH: Visualization, Investigation. AKH: Supervision: SKA: Software, Validation: SKA/AKH: Writing-Reviewing and Editing.

Funding The authors have no relevant financial or non-financial interests to disclose. The authors have no financial or proprietary interests in any material discussed in this article.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Human and animal rights Authors declare that no involvement of Human and Animals in the research work.

References

Adimalla N, Venkatayogi S (2017) Mechanism of fluoride enrichment in groundwater of hard rock aquifers in Medak, Telangana State South India. *Environ Earth Sci* 76:45. <https://doi.org/10.1007/s12665-016-6362-2>

- Adimalla N, Venkatayogi S, Das S (2019) Assessment of fluoride contamination and distribution: a case study from the rural part of Andhra Pradesh India. *Appl Water Sci* 9:94. <https://doi.org/10.1007/s13201-019-0968-y-y>
- Ali S, Kumari M, Gupta S, Sinha A, Mishra B (2017) Investigation and mapping of fluoride-endemic areas and associated health risk—A case study of Agra Uttar Pradesh, India. *Hum Ecol Risk Assess Internat J* 23(3):590–604. <https://doi.org/10.1080/10807039.2016.1255139>
- Ali S, Fakhri Y, Golbini M, Thakur S, Alinejad A, Parseh I, Shekhar S, Bhattacharya P (2019) Concentration of fluoride in groundwater of India: a systematic review, meta-analysis and risk assessment, *Groundwater for Sustainable Development*, 9. ISSN 100224:2352–2801. <https://doi.org/10.1016/j.gsd.2019.100224>
- APHA (2012) Standard methods for the examination of water and waste water, 22nd edn. American Public Health Association American Water Works Association, Water Environment Federation
- Arif M, Husain I, Hussain J, Kumar S (2013) Assessment of fluoride level in groundwater and prevalence of dental fluorosis in Didwana block of Nagaur district, Central Rajasthan India. *Int J Occup Environ Med* 4(4):178–184 (PMID: 24141866)
- Armiento G, Angelone M, De Cassan M, Nardi E, Proposito M, Cremisini C (2016) Uranium natural levels in water and soils: assessment of the Italian situation in relation to quality standards for drinking water. *Rend Fis Acc Lincei* 27:39–50. <https://doi.org/10.1007/s12210-015-0462-x>
- Avtar R, Kumar P, Surjan A (2013) Geochemical processes regulating groundwater chemistry with special reference to nitrate and fluoride enrichment in Chhatarpur area, Madhya Pradesh, India. *Environ Earth Sci* 70:1699–1708. <https://doi.org/10.1007/s12665-013-2257-7>
- Bajpai J (2013) Fluoride carcinogenesis: the jury is still out! *South Asian. J Cancer* 2(4):192. <https://doi.org/10.4103/2278-330X.119881>
- BIS (2012) Bureau of Indian standards, IS: 10500. Indian standard for drinking water specification
- Census of India (2011) Haryana (Series 07), Part XII-B, District Census Handbook-Mahendragarh. Directorate of Census operations, Haryana. https://censusindia.gov.in/2011census/dchb/0616_PART_B_DCHB_MAHENDRAGARH.pdf
- Chitrakshi HA (2018) Hydrogeochemical characterization and suitability appraisal of groundwater around stone quarries in Mahendragarh India. *Environ Earth Sci* 77:252. <https://doi.org/10.1007/s12665-018-7431-5>
- Cohn P (1992) A brief report on the association of drinking water fluoridation and the incidence of osteosarcoma among young males. Department of Health Environment
- Cremisini C, Armiento G (2016) High geochemical background of potentially harmful elements. The “geochemical risk” and “natural contamination” of soils and water: awareness and policy approach in Europe with a focus on Italy. *Rend Fis Acc Lincei* 27:7–20. <https://doi.org/10.1007/s12210-015-0457-7>
- Dey S, Giri B (2015) fluoride fact on human health and health problems: a review. *Med Clin Rev* 2:2. <https://doi.org/10.21767/2471-299X.100011>
- Fein N (2001) Fluoride content of foods made with mechanically separated chicken. *J Agric Food Chem* 49(9):4284–4286. <https://doi.org/10.1021/jf0106300>
- Gupta S, Deshpande R, Agarwal M, Raval B (2005) Origin of high fluoride in groundwater in the North Gujarat-Cambay region, India. *Hydrogeol J* 13:596–605. <https://doi.org/10.1007/s10040-004-0389-2>
- Haritash A, Kaushik C, Kaushik A, Kansal A, Yadav A (2008) Suitability assessment of groundwater in some villages of Rewari district in Haryana. *Environ Monit Assess* 145(1–3):397–406. <https://doi.org/10.1007/s10661-007-0048-x>

- Haritash A, Aggarwal A, Soni J, Sharma K, Sapra M, Singh B (2018) Assessment of fluoride in groundwater and urine, and prevalence of fluorosis among school children in Haryana India. *Appl Water Sci* 8:52. <https://doi.org/10.1007/s13201-018-0691-0>
- Kanduti D, Sterbenk P, Artnik B (2016) Fluoride: a review of use and effects on health. *Materia Socio-Med* 28(2):133–137. <https://doi.org/10.5455/msm.28.133-137>
- Kumar S, Lata S, Yadav J, Yadav JP (2016) Relationship between water, urine and serum fluoride and fluorosis in school children of Jhajjar District, Haryana, India. *Appl Water Sci* 7:3377–3384. <https://doi.org/10.1007/s13201-016-0492-2>
- Kumar B, Singh U, Mukherjee I (2017) Hydrogeological influence on the transport and fate of contaminants in the groundwater India. *JSM Biol* 2(1–11):1009
- Kumar S, Venkatesh A, Singh R, Udayabhanu G, Saha D (2018) Geochemical Signatures and isotopic systematics constraining dynamics of fluoride in groundwater across Jamui district, Indo-Gangetic alluvial plains, India. *Chemosphere* 205:493–505
- Levy S, Hillis S, Warren J, Broffitt B, Mahbubul Islam A, Wefel J, Kanellis M (2002) Primary tooth fluorosis and fluoride intake during the first year of life. *Community Dent Oral Epidemiol* 30(4):286–295. <https://doi.org/10.1034/j.1600-0528.2002.00053.x>
- Liang S, Nie ZW, Zhao M, Niu YJ, Xin KT, Cui XS (2017) Sodium fluoride exposure exerts toxic effects on porcine oocyte maturation. *Sci Rep* 7:17082. <https://doi.org/10.1038/s41598-017-17357-3>
- Mamatha P, Rao S (2010) Geochemistry of fluoride rich groundwater in Kolar and Tumkur districts of Karnataka. *Environ Earth Sci* 61:131–142. <https://doi.org/10.1007/s12665-009-0331-y>
- Matthess G (1982) The properties of ground water. Wiley
- Narsimha A, Sudarshan V (2016) Contamination of fluoride in groundwater and its effect on human health: a case study in hard rock aquifers of siddipet, Telangana state, India. *Water Sci* 7:2501–2512. <https://doi.org/10.1007/s13201-016-0441-0>
- Nuccio M (2016) Pollution of waters and soils by contaminants of magmatic origin. *Rend Fis Acc Lincei* 27:21–28. <https://doi.org/10.1007/s12210-015-0474-6>
- Pandey A, Shekhar S, Nathawat M (2012) Evaluation of fluoride contamination in groundwater sources in Palamu District, Jharkhand India. *J Appl Sci* 12:882–887. <https://doi.org/10.3923/jas.2012.882.887>
- Podgomy P, McLaren L (2015) Public perception and scientific evidence for perceived harms/risks of community water fluoridation: an examination of online comments pertaining to fluoridation cessation in Calgary. *Can J Pub Health* 106(6):413–425. <https://doi.org/10.17269/cjph.106.5031>
- Raj D, Shaji E (2017) Fluoride contamination in groundwater resources of Alleppey Southern India. *Geosci Front* 8(1):117–124. <https://doi.org/10.1016/j.gsf.2016.01.002>
- Selinus O, Centeno J, Finkelman R (2016) Medical geology: impacts of the natural environment on public health. *Geosciences* 6(1):8. <https://doi.org/10.3390/books978-3-03842-198-6>
- Ullah R, Zafar S, Shahani N (2017) Potential fluoride toxicity from oral medicaments: a review. *Iran J Basic Med Sci* 20(8):841–848
- USEPA (1993) US environmental protection agency. Reference dose (RfD): description and use in health risk assessments, Background Document 1A. <https://www.epa.gov/iris/reference-dose-rfd-description-and-use-health-risk-assessments>
- WHO (2004) World health organization guidelines for drinking water quality. World Health Organization
- Xu K, An N, Huang H (2020) Fluoride exposure and intelligence in school-age children: evidence from different windows of exposure susceptibility. *BMC Public Health* 20:1657. <https://doi.org/10.1186/s12889-020-09765-4>
- Yadav A, Kaushik C, Haritash A, Singh B, Raghuvanshi S, Kansal A (2007) Determination of exposure and probable ingestion of fluoride through Tea, Toothpaste, Tobacco and Pan Masala. *J Hazard Mater* 142:77–80. <https://doi.org/10.1016/j.jhazmat.2006.07.051>

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

Realistic face generation using a textual description

Anukriti Kumar

Department of Information
Technology
Delhi Technological University
New Delhi, India
anu1999kriti@gmail.com

Anurag Mudgil

Department of Information
Technology
Delhi Technological University
New Delhi, India
anuragmudgil00@gmail.com

Nakul Dodeja

Department of Information
Technology
Delhi Technological University
New Delhi, India
nkl.dodeja5@gmail.com

Dinesh Kumar Vishwakarma

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

Abstract— A lot of research is going on in the field of GANs and their ability to generate photorealistic images from their textual descriptions. Still the images generated by existing tasks focus primarily on either generating birds or flowers from their descriptions. This paper has successfully implemented a sketch-refinement technique for human face creation from their text descriptions. This problem has a wide variety of industrial applications as well such as creating comics automatically, assisting a movie creator to generate frames, art creation assisting an artist to generate sketches and even for educational purposes. Thus, keeping the need for such a system in mind, a two-staged StackGAN architecture obtained from deep convolutional neural networks is proposed in this paper. The features of faces like blonde hair, arched eyebrows are converted into an image of an actual person with these features. Not limiting to this, facial expressions like a wide smile, happy face are converted from its textual form to the corresponding image of the person with the same expressions. The generation of a high resolution 256×256 image using captions provided in CelebA dataset makes it a valuable contribution in the field of research. Further, the proposed research work has successfully obtained an inception score of 4.04 ± 0.05 over 10 iterations of evaluation and have shown promising results.

Keywords— *StackGAN, Generator, Discriminator, Sentence-BERT, Conditioning Augmentation Network, Inception Score*

I. INTRODUCTION

The ability to imagine things based on their physical description reflects the intellectual capabilities of human beings. Whenever we read a novel we are always curious about how the characters look in reality and we often imagine the appearances of the characters; imagining the complete person is still viable but getting description of some important details is still a daunting task. When a book is converted into a movie it is very important that the cast matches the actual description of characters making it more realistic for the audience.

Apart from it sketch artists often have to draw sketches of people that they have not seen in their life just based on the textual description provided to them. This finds immense use in tracking criminals and finding lost people. Thus generating photo realistic images of people based on their textual description is a problem of critical significance. It has a variety of applications from casting in media to designing faces of actual people to find them. There is also immense potential for usage in the fields of computer aided design and editing of photographs. Vast amount of research has been done in the field of image captioning and it is finding its implementation in commercial products like Google photos. Text to Image generation process is the inverse of the

process of captioning of images that possesses its own variety of important uses like Criminal Face Reconstruction, Face Generation of characters in a story etc. and this field is still not much explored by the researchers.

Prevailing ways for conversion of textual data to images mainly focus on generating images of objects, flowers or birds and only a few have addressed the issue of generating images of actual faces from the description. The ones which are based on generation of faces just reflect the basic meaning of the description and fail to show the details and vivid object parts. While image synthesis is in itself an arduous task in Computer Vision, generating images of high resolution is an even more difficult task because of possible lack of overlay between the substructures of distributions of an implied model and that of natural images in pixel space containing a large number of dimensions. With an increase in resolution further, the problem becomes more severe and starts giving nonsensical outputs.

The primary intention of this research is Text to Face Generation - generating 256×256 resolution images of faces of people on the basis of the textual description provided. We convert features of faces like blonde hair, arched eyebrows into an image of an actual person with these features. Not limiting to this we will also convert facial expressions like a wide smile, happy face from its textual form to its corresponding image of the person with the same expressions.

II. LITERATURE REVIEW

Multimodal deep learning [1] involves data estimation in a mode by means of provision of data in a separate mode. Text to Image generation is an illustration of the same.

DC-GAN [2] employed an end-to-end conditional GAN architecture to produce 64×64 images using features derived from an RNN. 256×256 images synthesized by a dual-stage GAN, known as StackGAN [3], 512×512 images synthesized by an architecture consisting of nested discriminators in hierarchy, known as HDGAN [4] and attention-harnessing AttnGAN [5] are some of the more recent developments in the field of image synthesis by means of descriptive text.

Text to Face generation is a related, under-researched problem that involves synthesis of the human face from the provided textual description of facial features of an individual. Face2Text [6] provides a comprehensive dataset that amalgamates the nature of labels (facial, inferred personal and emotional) present in old datasets and renders rich, well-defined labels of differing complexity in terms of syntax and semantics. However, the small size of the dataset implies that the model may remember the entire dataset and produce pseudo-good results. This is a contributing factor towards the unaddressed nature of the problem. LFW [7] and

CelebA [8] are large-scale datasets possessing images and their corresponding labels. The labels in these datasets contain information pertaining to physical attributes like, face shape and skin tone and personal attributes like, sex and age.

Prior to the advent and subsequent surge in popularity of GAN [9], [10], deep convolutional neural networks [11] and variational auto encoder [12] were employed in the field of face synthesis. An attribute-oriented solution proposed by Li et al., [13] aimed to preserve the identity of facial image by employing a combination of VCG-network and Gradient Descent algorithm. It is constrained by the possibility of facial synthesis via simple features solely. Disentangled representation is used in DC-IGN [14] alongside SGVB algorithm (VAE) for the task of face synthesis. But computational weakness stemming from dealing of uni-attribute per batch and requirement of labeled dataset renders it infeasible.

Rapid development in GANs and their conditional variants are responsible for much of the gradual growth in the quality of generated images. TP-GAN [15] proposes a GAN architecture based on dual pathway. It provides a realistic generation of the frontal view via local and global feature details. However, there is a need for a large, labeled front face dataset for the same. By employing pose code, instead of actual physical features, adjustment of the head was carried out in DR-GAN [16]. PIM [17] improves upon TP-GAN and DR-GAN by carrying out unsupervised training in the presence of a deep architecture.

A completely trainable GAN proposed by Khan et al., [18] is capable of harnessing descriptions for image generation but its performance in terms of actual synthesis and not mere matching from given dataset remains untested. FaceID-GAN [19] involved a competition among the generator, discriminator and identity classifier in terms of preservation and quality of image. It was able to achieve facial modification via expression attributes. However, self-specification of features is restrictive. To overcome this, FaceFeat-GAN [20] involved a dual stage process that produces diverse, synthesized features and identity preserving, high quality images from these features in the two respective stages. However, a broader-scoped, identity-preserving, accurate, synthesis of face from input textual description remains largely unaddressed.

III. METHODOLOGY

The following section aims to explain the work done in detail. It explains how a two-staged StackGAN architecture has been used to generate highly realistic images from textual description. This section explains how relevant data for our problem domain has been created. It also represents our model architecture and its major components. This also contains implementation details of Stage-I and Stage-II GANs, discussing the loss functions associated as well as their corresponding optimizers.

A. Dataset

We used CelebFaces Attributes Dataset (CelebA) for generating textual descriptions required for training our

model. It is a large-scale dataset with facial attributes of around 2,02,599 celebrity images each covering large pose variations and diversities. These cover around 10,177 identities alongside five landmark places and labels comprising of 40 features per image. This aforementioned dataset commonly acts as training & testing subsets, to be employed in a plethora of tasks in the field of Computer Vision.

In our work, we categorized the 40 attributes list in the CelebA dataset corresponding to each image into six different groups, each such group describing some peculiar facial characteristic. These groups were created based on the facial shape, hairstyle, appearance-describing attributes (such as young, pale etc.) accessories worn, hairstyle of the person and other facial features such as eyes, nose, lips etc. The attribute representing gender is used to identify a person as male/female, indicating whether to use 'he' or 'she' for the person in order to generate textual descriptions. Further, a dictionary is created with keys representing attributes from CelebA dataset and values from the text with which we want to replace them in the descriptions. While creating a sentence, we first appended a prefix corresponding to the unique group (Eg: "He has an" for describing oval face attribute) and then we added the corresponding value from the dictionary created. After summing up all these individual sentences about an individual, we obtained the textual description highlighting the person's facial features, outline and relevant accessories based on which we generated images. Table I presents a description of the dataset used.

TABLE I. DESCRIPTION OF DATASET

Face images of various celebrities	202599
Unique identities	10177
Binary attribute annotations per image	40
Landmark locations	5

B. Network Architecture

Our model is based on sketch-refinement methodology for which we implemented a two staged GAN architecture with StackGAN, capable of converting the arduous task into relatively more feasible sub-tasks. Here, the Stage-I GAN makes use of the annotations to sketch the basic object colors and shapes. Consequently, the images generated are of poor resolution. Then, by means of the outputs obtained from Stage-I serving as inputs alongside the annotations used, Stage-II GAN produces detailed, high quality, realistic images of significant resolution. This is possible courtesy of the ability of the Stage-II GAN to improve upon the issues in images of Stage-I GAN, whilst incorporating finer details during the process of refinement.

The subsequent subsections explain each of these in detail. In every stage, C-GANS are responsible for construction of the network of generators. Using them, it is possible to create fake, photo-realistic images after training on image data as they help learn a conditional density model.

Fig. 1 represents the architecture of the proposed model.

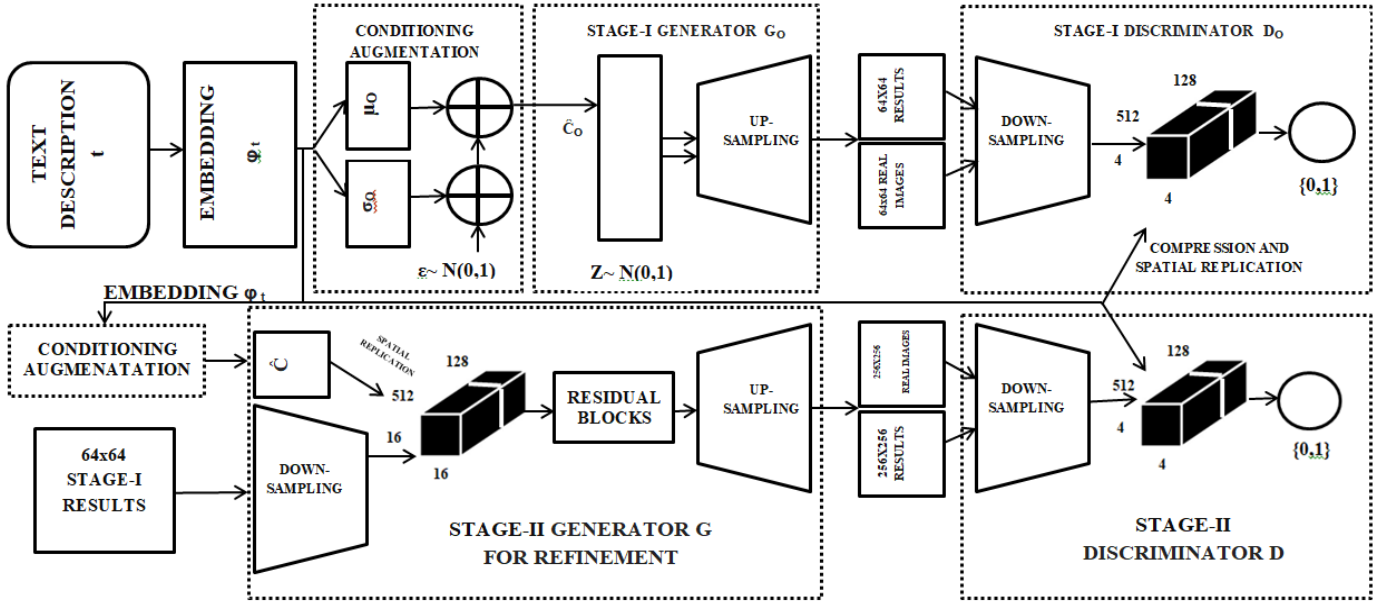


Fig.1. Architecture of proposed model

a) *Textual embedding encoder network*: To encode the text description into an embedding for which we have used a pre-trained Sentence-BERT network. It introduces pooling to the token embeddings generated by BERT in order for creating a fixed size sentence embedding. These can be used for unsupervised tasks to encode the semantics of sentence data. The major purpose of creating embeddings is to create vectors of constant size from input text of random size. The text encoder network encodes a sentence to a 1,024 dimensional text embedding. This network is common to both of the stages. Table II describes the shape of the training and test set embeddings.

TABLE II. DESCRIPTION OF THE SHAPE OF EMBEDDINGS

Train data embeddings shape	(151949, 10, 1024)
Test data embeddings shape	(50650, 10, 1024)

b) *Conditioning augmentation network*: Large number of dimensions is a common feature of text-conditioned latent space that can lead to discontinuities. Hence, we have utilized Conditioning Augmentation technique for extraction of latent variables from text embeddings in order to facilitate creation of supplementary variables for conditioning. In this technique, the conditioning augmentation block represents a single linear layer in which the embeddings of shape (1024,) are fed and in return, it provides a tensor of shape (256,) which is considered as the relevant input to be understood by GAN. This network is used to sample random latent variables from a distribution given by $N(\mu(\varphi_t), \Sigma_o(\varphi_t))$.

This technique adds more randomness in the network. Also, helps in making the generator more robust by capturing objects with much more diversity in terms of poses and appearances. With a higher number of image-text pairs, we can train a robust network that can handle perturbations. Through this, we not only just considered

the training images and the noise inspiration but we also considered some additional information represented as c which in our case was the phrase used to describe an image. This information is used by discriminator to check if the phrase matches with the image and for the generator, it is used as a part of inspiration for creating the image.

After obtaining textual embeddings from Sentence-BERT, a fully-connected layer is provided with them so as to generate values such as the mean as well as standard deviation. These are then used to create a diagonal covariance matrix by placing it in a diagonal of the matrix. Finally, we create a Gaussian distribution using these which can be represented as follows $N(\mu_o(\varphi_t), \Sigma_o(\varphi_t))$.

Then, we sampled from the Gaussian distribution that we just created. To sample \hat{C}_0 , we first take the element-wise multiplication of standard deviation and then add the output to mean. The formula to compute is as follows in Equation (1).

$$\hat{C} \sim N(0, I) \quad (1)$$

c) *Stage-I GAN generator network*: An architecture of Deep Convolutional Neural Network with several 2D up-sampling blocks, each containing an up-sampling and convolutional layer along with a batch normalization layer, has been created by us as is shown in the diagram Fig. 1. The generator network is a Conditional GAN, which is conditioned on the conditioning variable c and the random variable z sampled from a Gaussian distribution with dimension N . After concatenating the text-conditioning variable with the noise variable, we created a dense layer containing (16,384) nodes - two-dimensional tensor into four-dimensional tensor image.

It further generates a low-resolution image of dimensions which might represent just facial outline or

primitive facial features with a lot of defects. The architecture implemented in our work contains several up-sampling blocks made from several convolutional layers followed by batch normalization or activation layers and their parameters are given in Table III.

TABLE III. DESCRIPTION OF STAGE-I GENERATOR PARAMETERS

Total parameters	10,270,400
Trainable parameters	10,268,480
Non-trainable parameters	1920

d) *Stage-I GAN Discriminator Network*: Another Deep Convolutional Neural Network, consisting of various down sampling blocks as convolutional layers, that has been used by us is the Discriminator. These layers create feature maps which are further concatenated to textual embeddings. The text embeddings are compressed to less dimensions (three-dimensional tensor) and the image is generated via a number of down-sampling blocks, till the time it becomes a two-dimensional tensor. It is then concatenated and fed to a convolutional layer to facilitate joint learning of features vis-a-vis the textual description and image. Ultimately, we ascertained the probability of discerning actual data distribution images from those obtained via generators with the help of a fully connected layer containing a single node. Hence, after stage-I, we obtained 64×64 low resolution images with various distortions and vagueness. The training parameters are as given in Table IV.

TABLE IV. DESCRIPTION OF STAGE-I DISCRIMINATOR PARAMETERS

Total parameters	3,097,601
Trainable parameters	3,094,785
Non-trainable parameters	2,816

e) *Stage-II GAN Generator Network*: In this deep convolutional neural network, we have used lower dimensional Gaussian Conditioning Vector which is later transformed into a 3D tensor and the sample generated from the last stage is further compressed to 2D tensor by various down-sampling blocks which generated tensors with shape $16 \times 16 \times 512$ as can be seen from the given architecture to generate image features. Both of these are further concatenated into a single tensor with shape (batch_size,640) and finally, provided to the generator to encode the textual features alongside the image. The output of this is further fed through a series of up-sampling blocks that ultimately generated a high-resolution, more photorealistic image having dimensions 3.

f) *Stage-II GAN Discriminator Network*: Structure of the discriminator is similar to the one used in Stage-I GAN in the sense that it still employs a deep convolutional neural network, albeit with an increased number of down-sampling blocks as the input size in this stage is larger. These blocks are further followed by a concatenation block and then a classifier. This architecture helps obtain

arrangement of better nature between the conditioning text and the image that has been input. By considering real images alongside their corresponding textual descriptions as positive sample pairs and real images with text embeddings that have been mismatched, fake with embeddings of text as negative sample pairs, the Discriminator works.

C. KL Loss Function

Kullback-Leibler divergence (KL divergence) loss function also known as relative entropy function is calculated as the negative sum of probability of each event in A multiplied by the log of the probability of event B over the probability of event A. The KL Divergence loss function formula is given below as in Equation (2).

$$KL(M \parallel N) = -\sum_i M(x_i) \log (M(x_i)/N(x_i)) \quad (2)$$

Similar to any other GAN, training of the generator and discriminator networks in Stack-I and Stage-II GAN is done by minimizing generator network loss and maximizing discriminator network loss. KL_loss is a custom loss function, specified in the Training of generators of both the GANs.

IV. RESULTS AND EXPERIMENTATION

The following section presents the results obtained from the conducted experiments. It discusses Inception Score metric and depicts the variation of discriminator as well as generator loss with the number of training epochs. This section also provides a visual representation of the results obtained.

A. Inception Score

The inception score (in short IS) with its origin from Inception Classifier is a very widely used metric for judging the performance of images generated by General Adversarial Network. Inception score outputs a floating point number which is directly proportional to how realistic are the images generated by the GANs. The upper bound of score is infinity but in practicality there usually emerges a non-infinite ceiling. It gets well correlated with the human evaluation and can act as an alternative to a human actually judging the images for the quality. It takes into account both factors that images have variety and each image actually looks like a data instance.

To calculate both the conditions, the Inception score makes use of the Inception Classifier which classifies the image to a particular label by providing the probability values for each label. Now as similar labels sum to give focused distribution while different labels sum to a uniform distribution so we can use this classifier on generated images from GAN by passing the fake images through the classifier which can predict the label of images in the output.

So the score combines two factors that each image must be distinct (so probability distribution for a single picture must be narrow i.e. focused on one peak) and collectively it must be uniform (for the sum the distribution must be uniform) telling that the collection has variety.

We also used this score for the purpose of checking how our model works and we obtained IS of 4.04 ± 0.05 over 10 iterations.

Our inception score indicates that for every image x , textual encoding z and class label y , the value of marginal distribution $p(y)$ is as given by Equation (3).

$$p(y) = \int p(y|G(z))dz \quad (3)$$

It possesses entropy value that is large in magnitude and of a nature comparable to $p(y|x)$.

B. Loss Variation

The plots obtained from Stage-II can be shown in Fig. 2 and Fig. 3. These help in determining when to stop the training of GANs. If losses are not decreasing further, it indicates that the training can be stopped as there is no chance of improvement or if losses keep on increasing, then it indicates that the training must be stopped. But, if the losses are gradually decreasing, then we must continue training.

C. Visualization of Results

This section represents the results obtained with the help of our architecture. Fig. 4 depicts the conversion of textual description to facial image achieved by the model.

The man has an oval face with high cheekbones. He has wavy hair which is brown colored. He has a marginally open mouth. The youthful appealing man is smiling.



The lady has high cheekbones. She has straight hair which is brown colored. She has arched eyebrows and somewhat open mouth. The youthful appealing lady has makeup. She is wearing lipstick.



The man looks youthful and his hair is brown colored.



The lady has an oval face. She has straight hair which is brown colored. The youthful appealing lady has makeup. She is wearing lipstick.



The lady has an oval face with high cheekbones. She has an enormous lips and small eyes with curved eyebrows and somewhat open mouth. She has straight hair which is brown colored. She is wearing earrings as well as lipstick. The youthful appealing lady has makeup.



Fig. 4. Conversion of textual description to facial image



Fig. 2. Variation of Generator Loss with the Number of Training Epochs

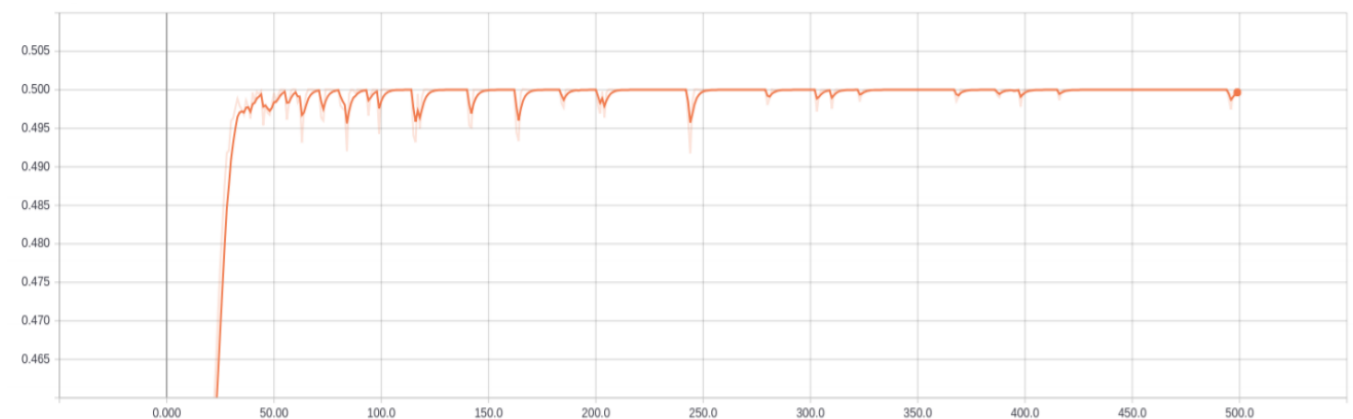


Fig. 3. Variation of Discriminator Loss with the Number of Training Epochs

V. CONCLUSION AND FUTURE WORK

The following section presents the conclusion arrived at from the conducted research along with prospective directions of future work.

The Stack GAN architecture implemented in this paper has shown promising results on our dataset. It is based on the sketch-refinement technique. In short, we obtained IS of 4.04 ± 0.05 over 10 iterations and we successfully synthesized 256×256 photorealistic images. Stage-I GAN takes a sentence as input and outputs a resultant picture containing colors and shapes of rudimentary nature whereas the second one (Stage-II) takes as input an image of lower resolution from Stage-I along with the initial sentence. This way it is able to synthesize an image possessing a resolution larger in magnitude having fine-tuned the intricacies (256×256). We have successfully synthesized people's pictures based on their description.

A lot of research is still going on in the field of GANs to handle the existing challenges such as Mode Collapse and Failure to converge. In future, an appropriate network can be proposed in order to solve these. Apart from this, increasing the size of our dataset further with a greater variety of textual descriptions can be a consideration to enhance the chances of better training. Further, capsule networks over CNN architectures can be considered in future due to their obvious advantages. A better evaluation metric other than inception score should be proposed later so that the similarity in images can be captured without using classes.

VI. ACKNOWLEDGEMENT

I would like to extend my gratitude towards Dr. Dinesh Kumar Vishwakarma (Associate Professor, IT) and all the faculty members of the Department of Information Technology, DTU. They all provided us with immense support and guidance for the project.

I would also like to express my gratitude to the University for providing us with the laboratories, infrastructure, testing facilities and environment which allowed us to work without any obstructions.

I would also like to appreciate the support provided to us by our lab assistants, seniors and our peer group who aided us with all the knowledge they had regarding various topics.

VII. REFERENCES

- [1] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," 2011.
- [2] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," 2016.
- [3] H. Zhang *et al.*, "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks," 2017, doi: 10.1109/ICCV.2017.629.
- [4] Z. Zhang, Y. Xie, and L. Yang, "Photographic Text-to-Image Synthesis with a Hierarchically-Nested Adversarial Network," 2018, doi: 10.1109/CVPR.2018.00649.
- [5] T. Xu *et al.*, "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks," 2018, doi: 10.1109/CVPR.2018.00143.
- [6] A. Gatt *et al.*, "Face2Text: Collecting an annotated image description corpus for the generation of rich face descriptions," 2019.
- [7] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," 2008, Accessed: Dec. 06, 2020. [Online]. Available: <http://vis-www.cs.umass.edu/lfw/>.
- [8] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-celeb-1M: A dataset and benchmark for large-scale face recognition," 2016, doi: 10.1007/978-3-319-46487-9_6.
- [9] I. J. Goodfellow *et al.*, "Generative adversarial nets," 2014, doi: 10.3156/jsoft.29.5_177_2.
- [10] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016.
- [11] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans. Neural Networks*, 1997, doi: 10.1109/72.554195.
- [12] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2014.
- [13] M. Li, W. Zuo, and D. Zhang, "Convolutional Network for Attribute-driven and Identity-preserving Human Face Generation," 2016.
- [14] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. B. Tenenbaum, "Deep convolutional inverse graphics network," 2015.
- [15] R. Huang, S. Zhang, T. Li, and R. He, "Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis," 2017, doi: 10.1109/ICCV.2017.267.
- [16] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," 2017, doi: 10.1109/CVPR.2017.141.
- [17] J. Zhao *et al.*, "Towards Pose Invariant Face Recognition in the Wild," 2018, doi: 10.1109/CVPR.2018.00235.
- [18] M. Z. Khan *et al.*, "A Realistic Image Generation of Face from Text Description using the Fully Trained Generative Adversarial Networks," *IEEE Access*, pp. 1–1, Aug. 2020, doi: 10.1109/access.2020.3015656.
- [19] Y. Shen, P. Luo, J. Yan, X. Wang, and X. Tang, "FaceID-GAN: Learning a Symmetry Three-Player GAN for Identity-Preserving Face Synthesis," 2018, doi: 10.1109/CVPR.2018.00092.
- [20] Y. Shen, B. Zhou, P. Luo, and X. Tang, "FaceFeat-GAN: a Two-stage approach for identity-preserving face synthesis," *arXiv*, 2018.

Recall-based Machine Learning approach for early detection of Cervical Cancer

Apoorva Gupta
Department of Biotechnology
Delhi Technological University
Delhi, India
gupta.apoorva1399@gmail.com

Ashutosh Anand
Department of Biotechnology
Delhi Technological University
Delhi, India
ashutoshanand586@gmail.com

Yasha Hasija
Department of Biotechnology
Delhi Technological University
Delhi, India
yashahasija06@gmail.com

Abstract— With frequent advancements in development of algorithms and need to incorporate them with clinically synthesized medical information is the paramount of modern-day bioinformatics. This aspect of computational study is of great healthcare significance as deduced results could be far fetched to generalize conclusions in regular medicine practice and diagnosis thus fastening up the process of detection. This paper tries to generalize cervical cancer detection approach with random forest regression technique. Unlike other papers which focus on accuracy and precision, this paper emphasizes on recall-based approach and beneficial tenets this approach over former ones. Four diagnostic tests used for early stage detection of Cervical cancer are Hinselmann's test, Schiller's test, Biopsy and Cytology. Each test is studied individually and analysis was made on the basis of confusion matrix, recall score and receiver operator curve (ROC). The basic aim during the entire development was to achieve higher recall scores with reduced false positive values.

Keywords— SMOTE, Expectation Maximization, Recall, ROC curve, Oversampling, Schiller, Cytology, Biopsy, Hinselmann, Cervical Cancer, Random Forest, Classification, SHAP

I. INTRODUCTION

As per the Indian Council of Medical Research (ICMR) estimates, there would be 1.04 lakh cases of cervical cancer in India in 2020. With a woman dying every 8 minutes due to this invasive disease the burden on existing healthcare regimes is immense to prevent its widespread. The disparity arising due to stark variation in socio-economic indicators of nations like every other community ailment has a significant effect on the impact assessment and policy frameworks undertaken to deal with it. Trends revealed from a survey done by the world cancer research fund further consolidate this notion with sub-Saharan African nations of Malawi, Swaziland, and Zambia being the worst-hit nations pan globe [1]. Although the results deduced from more developed regions show less prominence but a global outlook still indicates that the road for its complete eradication is still not half covered.

Oncologically cervical cancer involves the development of invasive malignant tumors in lower squamous regions of female genitalia. The prime reason in the majority of such instances is Human Papillomavirus Infection (HPI) that could have been acquired during sexual activity. Other factors may include environmental triggers or lifestyle complications like smoking or early age sexual activity. Considering the mortality rate especially in developing and underdeveloped countries the early-stage diagnosis of cervical cancer primarily becomes important because unregulated spread in high prominence zones could have

irreversible demographic implications. Some diagnosis methods followed by present-day oncologists include:-

- Biopsy - One of the techniques applied for the diagnosis of cancers in general. In it, a tissue sample from the affected region (here lower narrow portion of the uterus) with a similar kind of sample from the unaffected region is taken. The contrast in results obtained from two different tissue analyses is used to predict cancer prevalence.
- Cytology - It is a preliminary test basically employed to detect pre-invasive cancerous lesions.
- Hinselmann Test - It is a visual diagnostic technique in which a colposcope is used for a magnified examination of the female genitalia. In the case of cervical cancer, premalignant and malignant lesions are largely observed
- Schiller Test - It is a biochemical diagnostic method. Iodine solution is applied to the cervical region. Cells in normal cervical mucosa contain glycogen hence give brown stain which hitherto in the case of cancerous cells is not observed. Positive Schiller test results are generally followed by histological analysis or biopsy.

All the above tests are either done alone or in combination for the accurate detection of cervical cancer. Generally, these are performed as confirmatory tests especially Schiller and Biopsy but after abnormal pap smear test results.

Early-stage prediction techniques with a high specificity could prevent the further widespread of carcinogenic cells due to effective treatment availability at an early stage and lesser malignancy of tumor leading to reduced complications that otherwise increase drastically. At the same time will supplement the present physiological and biochemical diagnostics in yielding faster and effective throughputs. With the healthcare sector becoming more data-driven concordant with recent advancements in Machine Learning (ML) algorithms the scope of development of proposed alternative techniques is very wide. They could act as a possible catalyst to accelerate the process. In oncology, ML-based approaches could enable us to find a needle in a haystack as they rule out possibilities of human-based errors and biases. Classification is the most important ML method that can be employed in the diagnosis, detection and management of cancer. In this paper random forest classifier algorithm is used on the acquired dataset with statistical implications and related visualizations.

II. BACKGROUND

A. Data Imbalance

The class imbalance is one of the major challenges faced in classification problems and has received a lot of attention in ML. A dataset having binary classes is bound to be imbalanced when one of the classes has more number of instances than the other one. The class with fewer instances is under-represented in the dataset. This class imbalance can have a drastic effect on the classification algorithm when applied to real-world problems where it becomes disastrous to wrongly classify the instances from the under-represented class. One such example is the early diagnosis of a disease as such datasets usually have few instances of patients having the disease. This under-representation of patients having the disease translates to the Machine Learning algorithm misclassifying the patients [2] [3].

Numerous solutions, both at data and algorithmic level, have been recommended to solve data imbalance problem, which is encountered quite often. At the data level, several data re-sampling techniques such as random over- and under-sampling, and directed over- and under-sampling are employed [4]. Whereas at the algorithmic level, assigning weights to classes, probability estimate adjustment is some of the proposed solutions [5].

B. Random Oversampling

Random oversampling is a simple, non-heuristic technique to increase the instances of the minority class in order to balance the data by replicating the positive examples [6]. However, this might result in overfitting of the ML model as it just makes identical copies of the instances of minority class. Synthetic Minority Over-Sampling Technique (SMOTE), introduced by Chawla et al. [4], creates new instances of the minority class through interpolation between the already positive instances that are close together. These new instances enable the classifier to build larger decision regions near the minority class.

Oversampling the values in a dataset enhance the computational cost of the algorithm used for training and learning purposes. However, as shown by Batista et al. [6] and Barandela et al. [7], applying oversampling is recommended when minority class has fewer number of samples in the dataset as compared to the majority class, i.e., majority class/minority class ratio is high.

C. Random Forest Classifier

Random forest classifier is a powerful ensemble ML algorithm that has been proven to be very effective in pattern recognition and high-dimensional classification problems. Random forests were first introduced by Ho [8], Amit and Geman [9], and Breiman [10]. Random forest is a collection of decision trees and can be viewed as a classifier constituting various methods. The basic principle of random forests is to construct multiple binary trees using random bootstrap samples coming from a training set. Each tree in the forest makes classification on randomly selected sub-sample of the training data to yield a classification result. Then the forest selects the classification with maximum votes as the final result.

Random forests are based on the bootstrap aggregation concept of Breiman and the random feature selection concept of Ho. Therefore, individual trees instead of being trained on the whole dataset are trained on a subset of the dataset. For

example, let the dataset has M instances then, by bootstrap manner, $\frac{1}{3}$ of the instances are selected at random for individual tree and the rest are considered out-of-the bag observations to evaluate the error. Moreover, a random feature is chosen to be the decision node at each node. Therefore, for n number of features, the size of the feature selected at each split is either \sqrt{n} or $\sqrt{n/2}$ [11].

D. Recall value in early detection

The rationale behind every classification-based machine learning project is to deduce higher accuracy results. For this to bring at evaluation different techniques are deployed from simpler ones like accuracy and precision values to complex ones like F1 score and recall values. The ambiguity arises in the selection of the technique depending and factors affecting the final results whether it be the nature of the dataset, its premise/theme, or every ML algorithm is hardlined with one of these techniques. Considering the data dynamics and veracity the later proposition could be sidelined easily however to deconstruct the former aspect we need to understand these techniques briefly in relation to our dataset:-

- Accuracy - It is the fraction of total right predictions positive and negative from the total aggregated sample space.
- Precision - It is the fraction of right positive cancer predictions from total aggregated sample space whether the patient has cervical cancer or not.
- Recall - It is the fraction of right positive predictions from the total aggregated sample space of patients having cervical cancer.

Evidently recall value thus obtained is a key to higher specificity and should be leveraged over accuracy because it involves wrong predictions as well which are of no prediction significance in this case. So, this leaves us with precision and it should also be outweighed because the former itself involves actual cervical cancer patients clearing the idea that we should not miss any actual positive case with negative prediction over an actual negative case with positive prediction. Thus, the higher the recall value higher the model specificity. Fringing benefits from it is the reduction in mental stress that one undergoes after listening about cancer, costs in medicare, and undergoing cumbersome diagnostic tests [12].

E. ROC Curve

The receiver operator curve abbreviated as ROC is a machine learning-based graphical visualization method area under which (AUC) is used for the determination of binary classifier's tendency to categorize between required classifiers correctly like a correct positive cervical prediction on one side and correct negative on the other [13]. Higher AUC value accounts for more accurate binary differentiation with lesser overlapping. For projects involving multi-algorithm analysis, they also help in the identification of the best algorithm for the used dataset. It can be done by plotting algorithm wise ROCs and selecting the one with maximum AUC. The plot constitutes true positive rates (sensitivity) and false-positive rates (1-specificity) to summarize confusion matrices. Visually ROC curves more peaked towards the upper left-hand corner tend to have a higher discriminant tendency [14].

III. MATERIAL AND METHODS

A. Data Source and Features

Data is obtained from cervical cancer (risk factors) dataset available at the UCI Centre for Machine Learning and Intelligent Systems machine learning repository [15]. It is a multivariate dataset with 858 instances of cervical cancer patients reported at 'Hospital Universitario de Caracas' in Caracas, Venezuela. It acquired information on 36 different attributes ranging from demographic, lifestyle, and physiological to diagnostic features. Bottleneck arose due to missing data as several patients preferred not to share personal information due to privacy concerns. Thus, to deal with missing data a new imputation method based on the concept of expectation maximization was used.

Expectation maximization imputations are better than mean as it preserves the characteristic relationship of missing variables with associated features unlike mean which just makes simple replacement based on average from a single column. Statistically, it finds the maximum likelihood for a variable/missing value in a dataset where the value of related other variables is latent. Since data has missing values associated with variables having a distinct origin and features this algorithm is used.

Further, data was highly imbalanced with only 18 1's in Dx: Cancer (bool value for YES to cancer) column accounting for only 2.09% of true positive value. This imbalance could have led to erroneous values of specificity predictors. SMOTE was incorporated as a random oversampling technique to reduce the imbalance bias in the results. It further complimented the specificity of our alternative approach as well thus, serving the twin benefits. Dx: Cancer column acts as the label but, the final classification of whether the patient has cancer or not is made on the basis of four tests: - Hinselmann, Schiller, Cytology, and Biopsy. Therefore, the dataset was split into four parts, one for each diagnostic test.

B. Machine Learning Workflow

The execution starts with importing the dataset, since data has several missing values and data imbalance, expectation maximization and SMOTE is used to go away with it respectively. After preprocessing the training data is fed to a random forest regression algorithm. Predicted values are evaluated with test values with confusion matrix. Recall scores are obtained and further evaluated. For data pre-processing, in accordance with the general practice of deleting those columns which have less than 5% of the values filled is used to remove two columns, namely - 'STDs: Time since first diagnosis', 'STDs: Time since last diagnosis' - as they had maximum data cells missing. Following this, the data was split into training and testing dataset (70:30 ratio) before imputing the missing values. This is because the test data is supposed to be the data never encountered by the algorithm and hence, was kept aside for testing purposes only. After train-test-split, the missing values were imputed based on expectation maximization. The missing values in the test set were imputed based on the parameters of the training set. Data imputation was followed by data oversampling using SMOTE function of imblearn library. The training set was resampled so that the model could have enough positive instances to learn from the data. And the testing dataset was resampled so that the model could be tested for both the classes unbiased.

A unique method using Shapley Additive Explanations (SHAP) library is used. This recently developed model helps in comprehensive understanding of used machine learning models and output generated from it. The integration of this aspect enabled us to figure out the column attributes that have major involvement in forming the prediction thus affecting the values of ultimate indicators. Thus, the model for SHAP is developed separately for identification of most significant attributes and the results associated with each diagnostic test are evaluated.

IV. RESULTS

To obtain the results, we first cleaned the dataset as explained in the earlier sections. Then the dataset thus obtained was fed to a Random Forest Classifier, whose attributes were set to their default values, in order to see how each diagnostic test performed in combination with a machine learning algorithm. We chose recall as the measure to gauge the success of the model. In addition to accurately detecting the patients having cervical cancer, the detector should also report the least number of false-negative cases because if an early disease detection algorithm is reporting a high number of false negatives then it defeats the core concept of early detection of the disease. Table 1. shows the confusion matrix and the recall score obtained on the test dataset. Biopsy has the highest recall score (0.996) and reports a minimum number of false-negative cases (1) whereas Cytology has the least recall score (0.912) and reports a maximum number of false-negative cases (22). Schiller test closely followed biopsy with 7 false negatives observations and a recall score of 0.972. Further, Hinselmann test gave 20 false-negatives and recall value of 0.920. However, all of the diagnostic tests have a recall score of greater than 0.9.

Further SHAP library is deployed to identify factors affecting the diagnosis of cancer most and the conclusions drawn from evaluation are very interesting. The top 3 factors responsible for development or diagnosis of cervical cancer are 'Dx:CIN'(Abnormal cells that grow on cervix epithelium. These cells are not carcinogenic but may turn invasive leading to cancer), 'Dx:HPV(Human Papillomavirus infection)' and 'First sexual intercourse' in case of Hinselmann's test, biopsy and cytology. The first two factors remained the same for Schiller's test as well but third factors came out to be 'Dx'(Medical Diagnosis). For more information, the Github repository of the developed code could be accessed [16].

TABLE I. CONFUSION MATRIX AND RECALL SCORE FOR VARIOUS CERVICAL CANCER DIAGNOSTIC TESTS

S.no	Diagnostic Test	Confusion Matrix	Recall Score
1.	Hinselmann Test	$\begin{pmatrix} 251 & 0 \\ 20 & 231 \end{pmatrix}$	0.920
2.	Schiller Test	$\begin{pmatrix} 251 & 0 \\ 7 & 244 \end{pmatrix}$	0.972
3.	Cytology	$\begin{pmatrix} 251 & 0 \\ 22 & 229 \end{pmatrix}$	0.912
4.	Biopsy	$\begin{pmatrix} 251 & 0 \\ 1 & 250 \end{pmatrix}$	0.996

Figure 1. shows the ROC curve as well as the area under the ROC curve for both predictions using a random model and predictions using Random Forest Classifier. From Fig. 1.

it is observed that Biopsy has the maximum AUROC value (0.998) whereas Cytology has the minimum AUROC value (0.956).

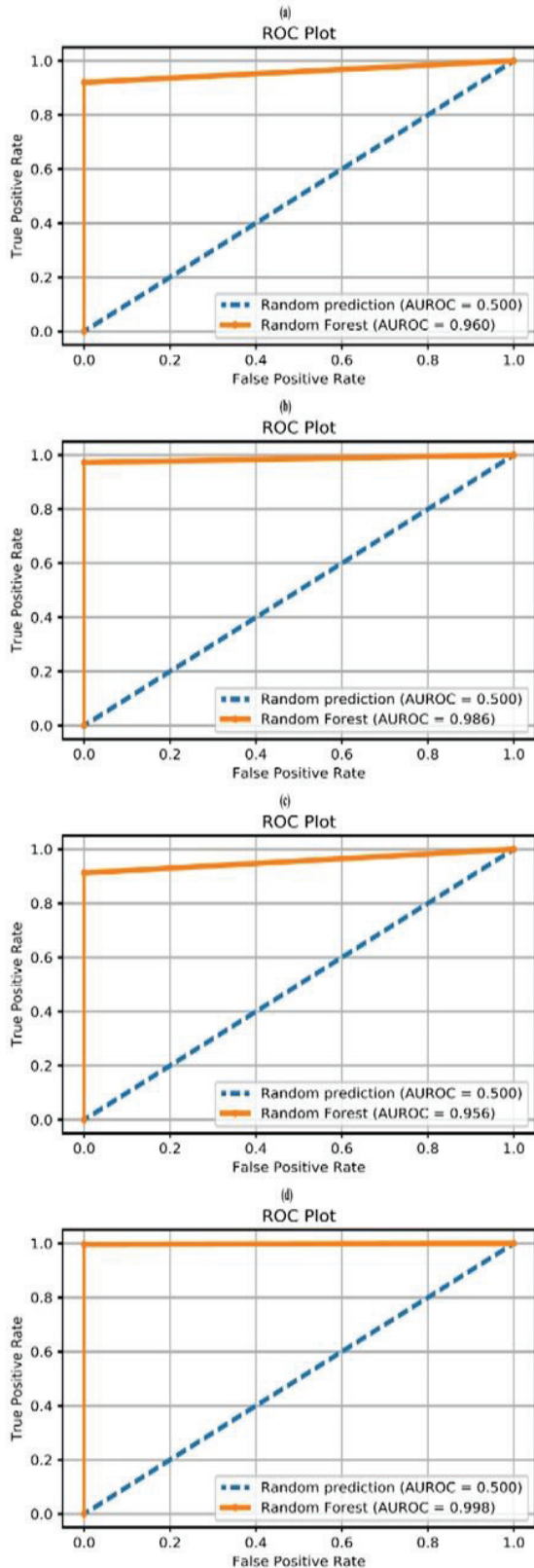


Fig. 1. ROC Curves along with AUROC values obtained from classifying the patients using both random model and Random Forest Classifier (a)Hinselmann (b)Schiller (c)Cytology (d)Biopsy

V. CONCLUSIONS

Our entire study comprehensively covered a major aspect in development of bioinformatics that is early detection of a disease in tested patients (cervical cancer here). Bioinformatic throughputs like these should be taken up at large scale. Since availability of data is the fuel for such studies, policy level interventions for major data storage in healthcare regimes are required. Bottlenecking arises due to lack of efficient data sources both in terms of volumes and versatility. Another main deduction from this study is further promotion of the idea of more applicability of studies focused on recall scores thus reducing false positive cases. The algorithm developed with assistance from SHAP library consolidated the fact that presence of CIN in cervix epidermis and HPV infection are the prime reasons that might lead to the development of cervical cancer and associated complications. Interestingly the age of first sexual intercourse came out to be the another most important factor for such tumour developments. Thus opening uncharted avenues in machine learning for both re-discovering and discovering new principles. However, for this, veracity and variety of the dataset is of the utmost importance. From the experimental process to collect the data to preparation of the final dataset, each and every step in-between plays a crucial role in establishing (or re-establishing) new factors (or already established factors).

Chronic diseases like cervical cancer have serious ramifications not only on patients but also on society and rapid surge in positive instances aggravate the already stressed healthcare and diagnostic infrastructure. Adoption of such studies at mass level with better quality of data inputs could provide eminent relief to these shortcomings. Collaborative inter-governmental initiatives in such directions could also be a good step in waging the gap arising due to financial and infrastructural constraints in developing and underdeveloped countries. The persistence of early sexual intercourse as one of the prime factors further directs as to undergo deliberations to make sex education mandatory with impetus to personal and reproductive health. Thus, making such primary studies a foundation for transforming modern estates of healthcare will not only alleviate untimely loss of life but also enhance the socio-economic output of the nations.

REFERENCES

- [1] "Cervical cancer statistics," 2018. <https://www.wcrf.org/dietandcancer/cancer-trends/cervical-cancer-statistics>.
- [2] V. García, J. S. Sánchez, R. A. Mollineda, R. Alejo, and J. M. Sotoca, "The class imbalance problem in pattern classification and learning."
- [3] D. Ramyachitra and P. Manikandan, "IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS: A REVIEW," International Journal of Computing and Business Research.
- [4] N. v. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
- [5] F. Provost and T. Fawcett, "Robust Classification for Imprecise Environments," Machine Learning, vol. 42, no. 3, pp. 203–231, 2001, doi: 10.1023/A:1007601015854.
- [6] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.

- [7] R. Barandela, R. M. Valdovinos, J. S. Sánchez, and F. J. Ferri, "The Imbalanced Training Sample Problem: Under or over Sampling," 2004, pp. 806–814.
- [8] Tin Kam Ho, "Random decision forests," in Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, pp. 278–282, doi: 10.1109/ICDAR.1995.598994.
- [9] Y. Amit and D. Geman, "Shape Quantization and Recognition with Randomized Trees," Neural Computation, vol. 9, no. 7, pp. 1545–1588, Oct. 1997, doi: 10.1162/neco.1997.9.7.1545.
- [10] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [11] O. Okun and H. Priisalu, "Random Forest for Gene Expression Based Cancer Classification: Overlooked Issues," in Pattern Recognition and Image Analysis, pp. 483–490.
- [12] R. E. Sharpe et al., "Increased Cancer Detection Rate and Variations in the Recall Rate Resulting from Implementation of 3D Digital Breast Tomosynthesis into a Population-based Screening Program," Radiology, vol. 278, no. 3, pp. 698–706, Mar. 2016, doi: 10.1148/radiol.2015142036.
- [13] C. E. Metz, "Basic principles of ROC analysis," Seminars in Nuclear Medicine, vol. 8, no. 4, pp. 283–298, Oct. 1978, doi: 10.1016/S0001-2998(78)80014-2.
- [14] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation," 2013.
- [15] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer Learning with Partial Observability Applied to Cervical Cancer Screening," 2017, pp. 243–250.
- [16] A. Gupta and A. Anand, "cervical_cancer_prediction." 2020, [Online]. https://github.com/mr-sesquipedalian/cervical_cancer_prediction.git



Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Recent advancements of PCM based indirect type solar drying systems: A state of art

Mukul Sharma^a, Deepali Atheaya^a, Anil Kumar^{b,c,*}^a Department of Mechanical Engineering, Bennett University, Greater Noida 201310, India^b Department of Mechanical Engineering, Delhi Technological University, Delhi 110 042, India^c Centre for Energy and Environment, Delhi Technological University, Delhi 110 042, India

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Solar energy
Solar drying
Indirect solar dryers
Phase change materials
Energy analysis
Economic analysis

ABSTRACT

Food storage is a widely known practice performed since ancient times at the domestic and industrial levels, and the vital process of food storage is food drying. The dried food has a longer shelf life than fresh food. To obtain the dried product with good taste and quality as the original food crop, this must be dried in a controlled environment. Solar food drying is the process through which the better quality of dried food crop can be obtained with low investment and minimal deterioration chances. Among the categories of solar food dryers, an indirect type solar dryer is most advantageous. It provides better quality of dried food crop in a shorter duration, and it can be embedded with the phase change materials (PCM) to achieve maximum efficiency during utilisation. PCM material helps to make the food dryer works during off sunshine hours. This work focuses on the advancement of PCM-based indirect type solar dryers during recent years. Paraffin wax emerged as a widely used PCM material for drying applications. The PCM-based indirect solar dryer developed by Atalay and Cankurtaran showed the highest thermal efficiency and CO₂ mitigation among the various discussed drying systems.

© 2021 Elsevier Ltd. All rights reserved.

Selection and peer-review under responsibility of the scientific committee of the Technology Innovation in Mechanical Engineering-2021.

1. Introduction

Food plays a vital role in human life [1]. Without food, no one can survive in the world. In the current pandemic situation, scarcity of food was observed globally, leading to hundreds of people worldwide due to hunger. However, at some places, a large amount of food was deteriorated during the lockdown period due to the unavailability of supply chains and food preservation techniques available at the domestic level. Food crops such as mangoes and tomatoes cannot be stored for a more extended period in the raw form [2,3]. Hence, these crops should be preserved by drying.

Food drying is an energy-intensive technique [1]. It requires a large amount of electrical energy for drying food crops using conventional drying equipment. In India, the cost of electrical energy is very high as well as its generation creates a lot of air pollution,

including the emission of greenhouse gases in the environment [4–6]. Therefore, the utilization of electrical energy for food drying is not economical for farmers and domestic users, which adversely affects the environment [7]. Hence, to protect the environment by minimising food drying costs, an alternate method is essential [8].

Solar energy is the most prominent source of energy accessible in each corner of the world [9]. Therefore, solar energy in food drying can save greenhouse gas emissions and be economical for society [8,10]. The general solar drying process adopted worldwide is an open sun drying procedure in which the crop is dehydrated under direct sunshine in open environment [11]. Even though the method is free, but it also has some substantial drawbacks. The method is lengthy and reduces the quality and colour of the crop [12]. The crop can get deteriorated with dirt, moisture, and insects [4]. If this food is cooked, it can upset human life and cause illness [7]. Further shortcomings are deterioration of crop due to adverse environmental circumstances like rain, forfeiture of the crop by birdies and insects, and deterioration through micro-organism and fungal development [8].

* Corresponding author at: Department of Mechanical Engineering, Delhi Technological University, Delhi 110 042, India.

E-mail address: anilkumar76@dtu.ac.in (A. Kumar).

<https://doi.org/10.1016/j.matpr.2021.04.280>

2214-7853/© 2021 Elsevier Ltd. All rights reserved.

Selection and peer-review under responsibility of the scientific committee of the Technology Innovation in Mechanical Engineering-2021.

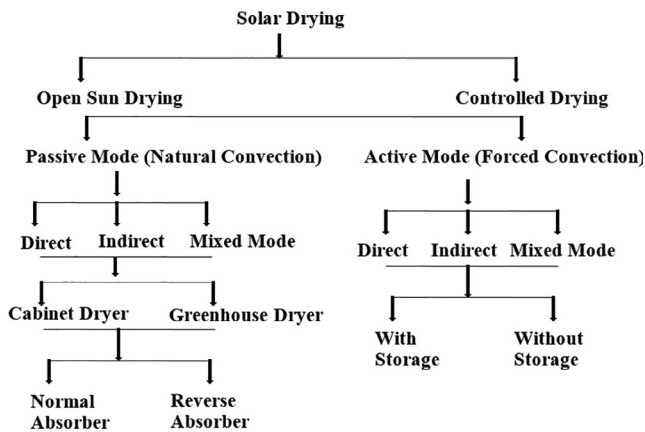


Fig 1. Categories of solar drying.

1.1. Types of solar dryers

The solar dryers harness solar radiation's energy and use that energy in crop drying [12]. On this basis, solar drying can be categorised into various types, as presented in Fig. 1.

Mainly, solar dryers are categorised into three categories [7,8]:

- Direct solar dryers
- Indirect solar dryers
- Mixed-mode type solar dryers

1.1.1. Direct solar dryers:

In the direct solar dryers, the crop is dried in inclusion at high temperatures. The food crop is kept in a cabinet where it receives incident shorter wavelength solar emission on the glass, transmitted inside the cabinet [13–15]. Some part of this incident solar radiation is captivated by the crop and the inner drying cabinet, and the other part is reflected in the atmosphere. This absorbed radiation increases the crop's temperature and the drying cabinet's internal environment, and in turn, the crop emits radiations of long-wavelength [15]. Thus, the temperature of the food crop inside the cabinet rises. These dryers are commonly known as solar cabinet dryers [3]. This working principle of a direct type solar dryer is displayed in Fig. 2.

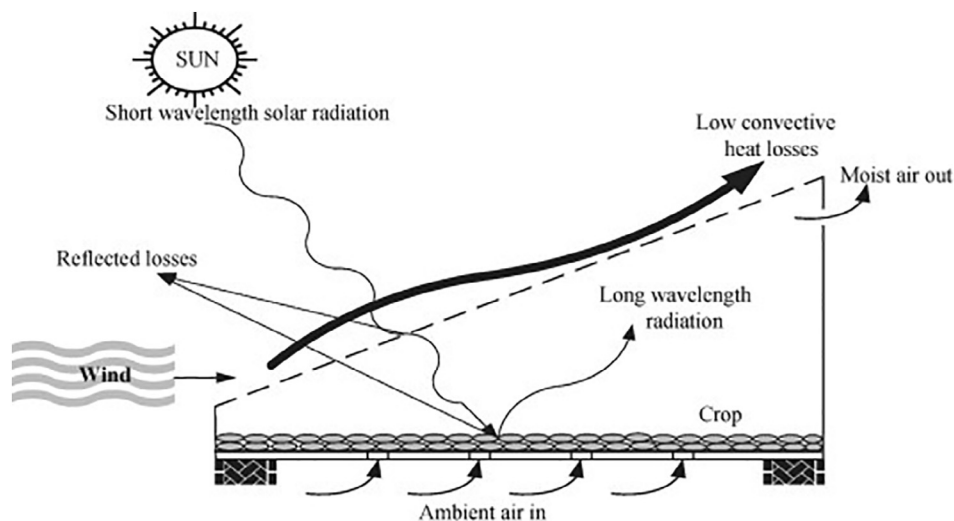


Fig 2. Working principle of direct type solar dryers [7]

Direct type solar dryers have the following limitations:

- The size of cabinet dryers is small, and it has limited usage [13].
- As a result of direct exposure of food crops to solar radiation, the cabinet dryers' dried products have faded colour, and the original properties of the dried products are lost [16].
- As moisture condensed on the cabinet's top glass sheet, the solar dryers' transmissivity gets reduced [7].

1.1.2. Indirect type solar dryers

The indirect type solar dryer consists of a solar radiation collector with a drying cabinet [7]. The solar emission falls over the solar radiation collector, and in turn, the air temperature inside the collector rises, which further passes through the dryer cabinet. The air circulation in the solar dryer can be through either by natural convection mode or forced convection mode. This air is further dumped into the atmosphere through vents or chimneys [15–17]. The crop drying in the indirect solar dryer occurs due to conductive and convective losses in the crop [3]. The working of indirect type solar dryer can be presented in Fig. 3.

The advantages of indirect types of the solar dryer are:

- The drying process can be controlled, and the drying produce has better quality than direct type solar dryers [18].
- A higher temperature inside the drying chamber can be achieved [19].
- The colour and quality of food crops are similar to the original crop sample.
- Food loss due to insects and birds is minimal.
- Heat losses can be minimised by insulating the drying cabinet.

1.1.3. Mixed-mode type solar dryers

This solar dryer type is a merger of indirect and direct solar dryers [20]. The crop drying process occurs due to the airflow through the solar radiation collector, and direct solar radiation is transmitted through the drying cabinet [21]. The food drying rate is higher than the other dryers, but product quality is not as good as indirect type solar dryers [2].

A big problem with existing solar dryers is that they can only work during the sunshine hours [22]. Due to this problem, the productivity of the solar dryer is less. Therefore, for efficient crop drying, the dryer should also work during no sunshine hours [23]. To

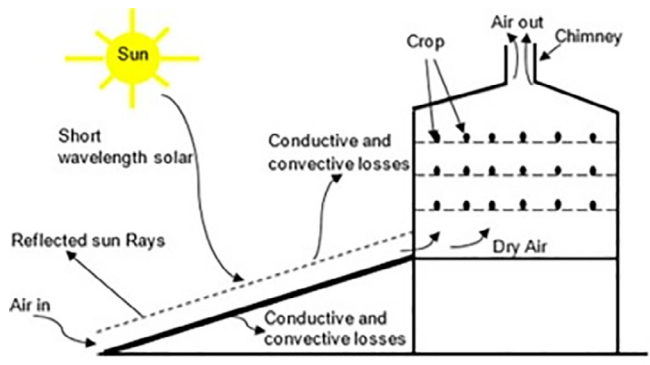


Fig 3. Working of an indirect type solar dryer [17]

achieve better solar dryer efficiency, the usage of phase change materials (PCM) was recommended. The PCM is thermal storage material that accumulates the surplus of the absorbed thermal energy in daylight hours and utilise it during no sunshine hours [23–26]. Since, the direct type of solar dryer can only work with direct solar radiations; therefore, the phase change materials cannot be embedded with this kind of solar dryer. Hence, the indirect type of solar dryers can be preferred to fulfill the objectives. In this work, recent trends of various PCM embedded indirect type solar dryers were studied thoroughly. Their various parameters that show their viability for drying various food materials were reported.

2. Review of PCM based indirect type solar dryers

Atalay and Cankurtaran (2020) fabricated and tested an industrial indirect type PCM-based solar dryer in Turkey. Paraffin wax was used for energy storage. The developed solar dryer consisted of air collectors, a drying chamber, an energy storage chamber, and centrifugal fans. The exergo-economic and exergo-environmental parameters were assessed for drying strawberries using the solar dryer. It was concluded that fans consume the most energy in the solar dryer. The cost of maximum energy destruction by fans was 0.2286 \$/h, and the exergy efficiency of the fans was 55.28%, with a minimum exergo-economic factor of 0.507.

In comparison, the drying cabinet had maximum exergy efficiency of 94.46%. The total embodied energy calculated for the system was 25283.44 kWh for an assumed average life span of 20 years. The energy payback period was reported to be 6.82 years, and the CO₂ mitigation amount was determined as 99.60 Tons [27].

Singh and Mall (2020) developed a PCM-based indirect type natural convection solar dryer for banana drying. The setup consisted of a solar collector and a drying chamber with a chimney. The PCM used for storing the energy was paraffin wax. The thermal performance was investigated for the drying arrangement. The maximum drying efficiency and the overall collector efficiency were 9.88 and 66.32%, respectively [28].

Bhardwaj et al. (2019) fabricated and tested a novel indirect solar dryer incorporated with sensible heat storage and PCM for medicinal plants drying in the western Himalayan region. The system comprised a flat plate solar collector with mixed iron scraps and copper tubing containing engine oil as sensible heat storage material and paraffin RT-42 as PCM was placed inside two containers and kept at the bottom-most area of the drying compartment. The dryer was tested under forced convection mode for drying *Valeriana Jatamansi*, and it was detected that the vapour content was decreased to 9% from the base value of 89% in 120 h. The system's average exergy and energy efficiency with SHSM and PCM was 26.10% and 0.81%, respectively [29].

Swami et al. (2018) developed a PCM-based solar dryer for fish drying in the coastal Konkan region. The dryer included a heating chamber, drying chamber, chimney, and an air blower. The dryer dimensions were 745 mm × 525 mm × 540 mm. The phase change material used in the heating chamber was paraffin wax. The dimensions of the flat plate collector were 1165 mm × 480 mm × 150 mm. The dryer works under forced convection mode. The dryer was tested for fish drying using paraffin wax (C-23) and paraffin wax (C-31) simultaneously, and it was concluded that paraffin wax (C-23) was suitable for fish drying. It reduced the drying period of fish by 75% [30].

El-Sebaai and Shalaby (2017) fabricated and investigated an indirect solar dryer with PCM for drying *Thymus* leaves. The whole experiment was conducted inside a room. The IDSD system consisted of two flat plate solar collectors installed over the roof of a room, an air blower, and a drying chamber kept inside the room. The PCM was stored in two shell and tube storage units attached at the drying chamber's bottom. Paraffin wax was selected and used as PCM for drying the selected sample. The experiment was performed with and without PCM, and it was concluded that the drying time was reduced by 50% in a system integrated with PCM compared to without PCM [22].

Khadraoui et al. (2017) fabricated and examined a PCM-based indirect type forced convection solar dryer. The dryer was verified in the no-load condition with paraffin wax as PCM was kept inside the solar energy accumulator. The solar energy accumulator's daily energy and exergy efficiency was observed as 33.9% and 8.5%, correspondingly. It was reported that the drying chamber's temperature was maintained between 4 and 16 °C during the night [23].

Aiswarya and Divya (2015) analysed a PCM-based indirect solar dryer's economic feasibility for drying cultivated products. The constructed solar dryer comprised an inclined solar air heater and a drying compartment. Paraffin wax (PCM) was kept at the bottom of the drying chamber. The dryer's capital investment was \$406.88, and the payback period was 0.578 years during the drying of crops like potato and cassava [24].

Jain and Tewari (2015) developed an indirect pass natural convective solar crop dryer embedded with PCM. The dryer components were a flat plate solar collector, natural ventilation system, drying chamber with crop trays, and packed bed PCM energy storage. It was detected that the drying chamber's temperature was

Table 1
Summary of literature review.

Literature	Year	PCM used	Parameters investigated
Atalay and Cankurtaran [27]	2020	Paraffin wax	Exergy efficiency, total embodied energy, energy payback period, CO ₂ mitigation amount
Singh and Mall [28]	2020	Paraffin wax	Maximum drying efficiency, Overall collector efficiency
Bhardwaj et al. [29]	2019	Paraffin RT-42	Performance parameters, Average exergy, and energy efficiency
Swami et al. [30]	2018	Paraffin wax	Performance parameters
El-Sebaai and Shalaby [22]	2017	Paraffin wax	Performance parameters and statistical analysis
Khadraoui et al. [23]	2017	Paraffin wax	Performance parameters, Daily energy, and exergy efficiency
Aiswarya and Divya [24]	2015	Paraffin wax	Various economic parameters such as capital investment of the dryer, payback period
Jain and Tewari [31]	2015	Paraffin wax	Thermal efficiency, Various economic parameters such as capital investment of the dryer, payback period
Shalaby and Bek [32]	2014	Paraffin wax	Maximum thermal efficiency and other performance parameters

6 °C more than the ambient during the night. The reported thermal efficiency and the dryer's payback period were 28.2% and 1.5 years, respectively [31].

Shalaby and Bek (2014) constructed a PCM-based novel indirect solar dryer. The drying system components were two solar air heaters, a drying chamber, a PCM storage unit, and a blower. The system was experimentally verified for drying of *O. basilicum* and *T. neriifolia* using paraffin wax as PCM, and it was described that the final moisture was obtained in 12 h and 18 h of drying and the extreme thermal efficiency of the arrangement was 70% [32]. The summary of the literature reviewed is given in Table 1.

3. Conclusions

Food drying is an essential process of saving food for future purposes. It helps in storing the food for a long time and minimising its wastage. The conventional food drying process uses a large amount of high-grade energy, viz. electricity can be saved using solar food drying processes, reducing greenhouse gas emissions in the atmosphere. The widely-used open sun-drying process results in low-quality dried products and some food product loss due to various environmental conditions, which can be decreased by employing solar dryers for food drying purposes. The indirect type solar dryers embedded with PCM for energy storage have various advantages. They produce a better quality of dried products with high efficiency, and these can be utilised during off sunshine hours. Recent advancements of PCM-based solar dryers were studied in this state-of-the-art review. The following conclusion were drawn from this study:

- Paraffin wax was a primarily used PCM material for solar drying applications.
- The PCM-based solar dryer developed by Aiswarya and Divya exhibited the lowest payback period of 0.578 years.
- The PCM-based solar dryer fabricated by Atalay and Cankurtaran reported the highest maximum exergy efficiency of 94.46%, and it also had the highest CO₂ mitigation of 99.6 Tons.

CRediT authorship contribution statement

Mukul Sharma: Methodology, Writing - original draft. **Deepali Atheaya:** Visualization, Investigation, Supervision. **Anil Kumar:** Conceptualization, Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

There is no direct funding for this work. Authors are highly thankful to Bennett University, Greater Noida, India, and Centre for Energy and Environment, Delhi Technological University, for providing basic infrastructure for compiling this work.

References

- [1] A. Jain, M. Sharma, A. Kumar, A. Sharma, A. Palamanit, Computational fluid dynamics simulation and energy analysis of domestic direct-type multi-shelf solar dryer, *J. Therm. Anal. Calorim.* (2019).
- [2] O. Prakash, A. Kumar, V. Laguri, Performance of modified greenhouse dryer with thermal energy storage, *Energy Rep.* 2 (2016) 155–162.
- [3] M.S. Sodha, A. Dang, P.K. Bansal, S.B. Sharman, An analytical and experimental study of open sun drying and a cabinet tyre drier, *Energy Convers. Manag.* 25 (3) (1985) 263–271.
- [4] O. Prakash, A. Kumar, Environmental analysis and mathematical modelling for tomato flakes drying in a modified greenhouse dryer under active mode, *Int. J. Food Eng.* 10 (4) (2014) 669–681.
- [5] P.S. Chauhan, A. Kumar, Thermal analysis of insulated north-wall greenhouse with solar collector under passive mode, *Int. J. Sustain. Energy* 37 (4) (2018) 325–339.
- [6] S. Vijayan, T.V. Arjunan, A. Kumar, Exergo-environmental analysis of an indirect forced convection solar dryer for drying bitter gourd slices, *Renew. Energy* 146 (2020) 2210–2223.
- [7] O. Prakash, A. Kumar, Y.I. Sharaf-Eldeen, Review on Indian solar drying status, *Curr. Sustain. Energy Reports* 3 (3–4) (2016) 113–120.
- [8] A. Kumar, R. Singh, O. Prakash, Ashutosh, Review on global solar drying status, *Agric. Eng. Int. CIGR J.* 16 (4) (2014) 161–177.
- [9] P.S. Chauhan, A. Kumar, C. Nuntadusit, J. Banout, Thermal modeling and drying kinetics of bitter gourd flakes drying in modified greenhouse dryer, *Renew. Energy* 118 (2018) 799–813.
- [10] P. S. Chauhan, A. Kumar, C. Nuntadusit, Thermo-environmental and drying kinetics of bitter gourd flakes drying under north wall insulated greenhouse dryer, *Sol. Energy*, 162 (April) 2017, 205–216, 2018.
- [11] O. Prakash, A. Kumar, Y. I. Sharaf-Eldeen, Review on Indian Solar Drying Status, *Curr. Sustain. Energy Reports*, 3 (3–4), 113–120, 2016.
- [12] A. Kumar, G.N. Tiwari, Thermal modeling and parametric study for forced convection green house drying of jaggery.pdf, *Int. J. Agric. Res.* 3 (2006) 265–279.
- [13] P.P. Singh, S. Singh, S.S. Dhaliwal, Multi-shelf domestic solar dryer, *Energy Convers. Manag.* 47 (13–14) (2006) 1799–1815.
- [14] H. Elhage, A. Herez, M. Ramadan, H. Bazzi, M. Khaled, An investigation on solar drying: a review with economic and environmental assessment, *Energy* (2018).
- [15] O. Ekechukwu, B. Norton, Review of solar-energy drying systems II: an overview of solar drying technology, *Energy Convers. Manag.* 40 (6) (1999) 615–655.
- [16] V. Shrivastava, A. Kumar, P. Baredar, Developments in Indirect solar dryer: a review, *Dev. Indirect Sol. Dry. a Rev.* 3 (4) (2014) 67–74.
- [17] A.B. Lingayat, V.P. Chandramohan, V.R.K. Raju, V. Meda, A review on indirect type solar dryers for agricultural crops – Dryer setup, its performance, energy storage and important highlights, *Appl. Energy*, vol. 258, no. October 2019, p. 114005, 2020.
- [18] A. Lingayat, V.P. Chandramohan, V.R.K. Raju, A. Kumar, Development of indirect type solar dryer and experiments for estimation of drying parameters of apple and watermelon: Indirect type solar dryer for drying apple and watermelon, *Therm. Sci. Eng. Prog.* 16 (2020), <https://doi.org/10.1016/j.tsep.2020.100477>.
- [19] P. Demissie, M. Hayelom, A. Kassaye, A. Hailesilassie, M. Gebrehiwot, M. Vanierschot, Design, development and CFD modeling of indirect solar food dryer, *Energy Procedia* 158 (2019) 1128–1134.
- [20] S. Singh, S. Kumar, New approach for thermal testing of solar dryer: development of generalized drying characteristic curve, *Sol. Energy* 86 (7) (2012) 1981–1991.
- [21] B.O. Bolajii, A.P. Olalusi, Performance evaluation of a mixed-mode solar dryer, *AU J. Technol.* 11 (4) (2008) 225–231.
- [22] A.A. El-Sebaei, S.M. Shalaby, Experimental investigation of drying thymus cut leaves in indirect solar dryer with phase change material, *J. Sol. Energy Eng. Trans. ASME* 139 (6) (2017) 1–8.
- [23] A. El Khadraoui, S. Bouadila, S. Kooli, A. Farhat, A. Guizani, Thermal behavior of indirect solar dryer: Nocturnal usage of solar air collector with PCM, *J. Clean. Prod.* 148 (2017) 37–48.
- [24] M.S. Aiswarya, C.R. Divya, Economic analysis of solar dryer with PCM for drying agriculture products, *Int. Res. J. Eng. Technol.* 02 (04) (2015) 1948–1953.
- [25] S. Verma, S. Mohapatra, S. Chowdhury, G. Dwivedi, Cooling techniques of the PV module: a review, *Mater. Today Proc.* (2021).
- [26] A.A. Bachchan, S.M.I. Nakshbandi, G. Nandan, A. Kumar Shukla, G. Dwivedi, A. Kumar Singh, Productivity enhancement of solar still with phase change materials and water-absorbing material, *Mater. Today Proc.* (2021).
- [27] H. Atalay, E. Cankurtaran, Energy, exergy, exergoeconomic and exergo-environmental analyses of a large scale solar dryer with PCM energy storage medium, *Energy*, p. 119221, 2020.
- [28] D. Singh, P. Mall, Experimental investigation of thermal performance of indirect mode solar dryer with phase change material for banana slices, *Energy Sources Part A Recover. Util. Environ. Eff.* (2020).
- [29] A. K. Bhardwaj, R. Kumar, R. Chauhan, Experimental investigation of the performance of a novel solar dryer for drying medicinal plants in Western Himalayan region, *Sol. Energy*, 177(May 2018), pp. 395–407, 2019.
- [30] V.M. Swami, A.T. Autee, Experimental analysis of solar fish dryer using phase change material, *J. Energy Storage* 20 (669) (2018) 310–315.
- [31] D. Jain, P. Tewari, Performance of indirect through pass natural convective solar crop dryer with phase change thermal energy storage, *Renew. Energy* 80 (2015) 244–250.
- [32] S.M. Shalaby, M.A. Bek, Experimental investigation of a novel indirect solar dryer implementing PCM as energy storage medium, *Energy Convers. Manag.*, (83), 1–8, 2014.

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/351108139>

HBRP Publication Page 1-11 2021. All Rights Reserved Recent Trends in Automation and Automobile Engineering

Conference Paper · April 2021

DOI: 10.5281/zenodo.4629049

CITATIONS

0

READS

21

3 authors, including:



Md ZIA Arzoo

Delhi Technological University

2 PUBLICATIONS 0 CITATIONS

[SEE PROFILE](#)



Mozamml Hassan

Delhi Technological University

1 PUBLICATION 0 CITATIONS

[SEE PROFILE](#)

Review on Spray Powder Process of Additive Manufacturing

Md Zia Arzoo¹, Mozammil Hassan^{2}, A.K.Madan³*

Department of Mechanical Engineering, Delhi Technological University

**Corresponding Author*

E-mail Id:-mozammilhassan_2k18me129@dtu.ac.in

ABSTRACT

As in Today's fast-paced growing world, all technologies are continuously advancing in every field to increase performance. One such advancement in the manufacturing sector is additive manufacturing, which is one of the recent technologies and it's expected to have a market reach of 26.68 billion dollars by 2027 growing at a rate of 14.4% according to reports and data. Additive manufacturing is the process of combining material to make objects by layer addition from the 3D CAD model. It is the opposite of subtractive manufacturing and it can produce products with greater accuracy. In this paper, we will be discussing briefly the spray powder process of additive manufacturing which includes 3D printing and direct energy deposition (DED), and we will be briefly going through the various metal and alloys which can be used in this process, working of these processes, mechanical properties of the product, various applications and its use in future.

Keywords:- Computer aided design (CAD), Additive manufacturing (AM), Direct Energy deposition (DED), 3D printing (3DP)

INTRODUCTION

Additive manufacturing (AM) is the process in which first the computer takes the information from a CAD (Computer-aided design) file and then converts it into STL (Stereolithography) file, in which the 3D CAD design in CAD software is approximated by triangles and then it is sliced layer by layer with every layer containing information to be printed. There are various types of additive manufacturing processes like liquid, molten/filament, powder, solid layer processes. The Powder bed process and spray powder process are the two types of powder processes in additive manufacturing. The Spray powder process consists of

1. Direct energy deposition (DED) and
2. 3D printing

DED definition As per ASTM (F2792):

“An additive manufacturing process in which focused thermal energy is used to fuse materials by melting as they are being deposited”

Direct energy deposition, blown powder additive manufacturing is an additive manufacturing method that uses a focused energy source such as an electron beam, plasma arc, laser to melt a material which then simultaneously will be deposited by a nozzle. It uses two different techniques: 3D cladding and 3D welding.

DED normally uses wire feed through a nozzle with plasma, laser or electron beam to create a molten pool. It is harder to deliver the wire in rougher surfaces and complex geometries which limits its application in various fields. Therefore, due to this reason powder is preferred mostly and is used frequently as the material used in the Direct energy deposition process.

It has attracted a lot of interest because of its ability to print any metal or metal-alloy device, primarily functionally gradient materials [13]. It is also used for remanufacturing and repairing parts to extend their life and decrease its

environmental effect. One of the examples of DED laser engineered net shaping (LENS) and laser deposition [3,4].

The 3D printing process involves spraying a water-based liquid binder onto a starch-based powder in order to print data from a CAD drawing. It is an MIT licensed process. The various advantages of 3D printing are less expensive machines, more cost-effective process, no resins to cure, no chemical post-processing required.

There are two major processes in 3D printing (1) material jetting 3D printing and binder jetting 3D printing. In material jetting 3D printing uses droplets which are

then poured by passing a UV light through each layer [5,6]. In the binder jetting a binder in fluid form is sprayed and imprinted onto a powder bed utilizing an inkjet head to bind the powder particles layer by layer as shown in Figure 2. The determination of binder significantly affects the mechanical strength and various properties of the manufactured part.

Great advancements has been made in 3D printing such as metal extrusion which is now used in which a product is made from the material by the mechanical head like the way inkjet printers extrude ink onto paper[7]. That's it is also known as inkjet 3D printing.

PRINTING PROCESS

Direct Energy Deposition

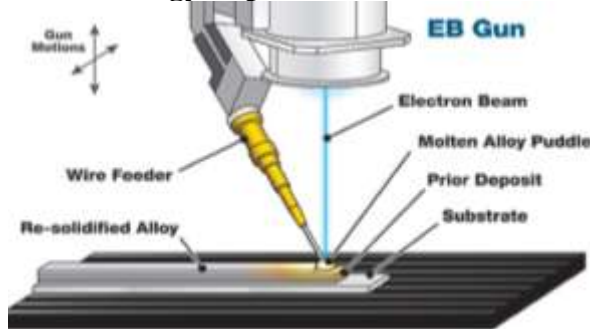


Fig.1:- Wire feed DED

In direct energy deposition, control and direction of the heat source and feedstock involve in the main process and it mainly depends on the energy and melts dynamics.

The various heat sources utilized indirect energy deposition are: electron beam, electric arc, laser and plasma. The feedstock is directed by the heat source at the place of deposition deposition that's why it is called direct energy deposition as shown in Figure 1. After this, the feedstock either powder or wire is fed into the path where heat source melts it, which cause the heat source to spray or drip into a comparative large melting pool. Meltpool is directed by controlling the motion of

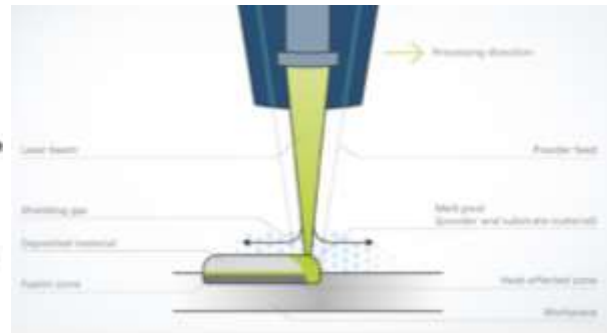


Fig.2:- Powder feed DED

feedstock and heat source around the toolpath where it freezes and forms solid metal bead. Generally, in wire-based DED, the bead size is much larger than the input feedstock. The melting mechanism is similar as that for traditional welding and have high energy requirement for melt pool and successful bond formation of the deposited material to the part.

3D Printing

The 3DP process is quite a long process and it involves transforming a digital CAD file into a solid 3D object. The process is shown in following 4 steps in Figure 3 below:

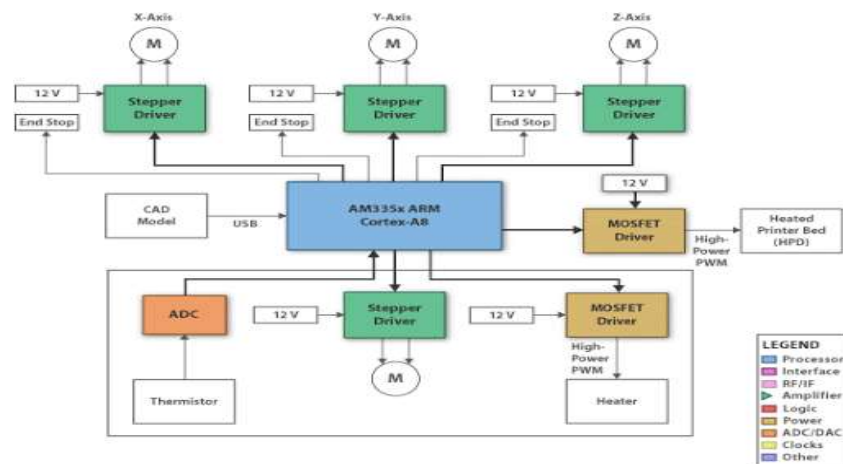


Fig.3:- 3D printing process[6]

Step-1

CAD file

In starting, a 3D CAD model is being made. This program creates a file containing all the information is then sent to the 3D printer within this process the software slices the design into thousands of horizontal layer. After this these layers are printed one on top of each other until the 3D object is being made. Various programs used for designing are AutoCAD, Solidworks, and various other programs like Google Sketchup etc.

Step-2

Motoring

The operating of 3D printing requires driving a stepper motor with high accuracy and low torque. The function of the stepper motor is to convert mechanical shaft rotation in motor from digital pulses. It provides high reliability at low costs, high torque at lower speeds. They are a special type of synchronous motors which when receives an electric impulse by its control unit rotates a certain degree. A 3d printer generally needs 4 stepper motors, 3 of them for motions in X, Y, Z directions and one for moving the bedplate.

Step-3

Processing

For independent operation and higher efficiency and flexibility an AM335x

microprocessor is used which consists of 1) a microprocessor unit 2) a Graphics accelerator subsystem for 3D graphics to support the display. It can scale the speed of operation from 600MHz to greater than 1 GHz.

Step-4

MOSFET drive

Since the stepper motor is an open-loop system, it requires a high accuracy component to print in the specific places so MOSFET drive with a bipolar stepper motor is used which allow the motor to move in both direction with fast frequency operation due to MOSFET.

MECHANICAL PROPERTIES

DED

Tensile Strength

The printed part's ductility and tensile strength are controlled by the various process parameters in DED and the material's microstructure. In various cases, a varying trend of ductility and tensile behaviour for the same material printed by direct energy deposition was seen like tensile strength of Ti-6Al-4V manufactured by DED was found quite similar to the wrought manufactured Ti-6Al-4V, with a decrease in ductile power[8]. in one other research, it was seen that the Ti-6Al-4V manufactured by DED was more tensile because it had a finer microstructure when

compared with wrought Ti-6Al-4V, but still, the ductility power was less than that of wrought Ti-6Al-4V due to a mixture of internal defects and finer microstructure[9]. and in one another research part made from DED showed tensile behaviour and anisotropic porosity in three different orientation, due to micro-structural anisotropy[10].

Hardness

Due to the variation in build direction, microhardness value may change along the build direction. The microhardness is comparatively lower in the central layer and relatively higher in the last and first layer. Thermal history in process can be the reason that heat is built at the central layer which results in lower microhardness values[11]. It was reported with the increase in the substrate thickness more hardness and finer microstructure was found because the addition of substrate mass leads to a faster heat sink. With the increase in the substrate, hardness decreased because of the decrease in cooling rate and thermal gradient which lead to lead to a coarser microstructure. [12]. post-processing of Additive Manufactured parts (like heat treatment or ageing) or selection of alloy can provide better control on the hardness, instead of varying process parameter in direct energy deposition process as reported by Zuback et al.[13].

Fatigue

The fatigue is seen mainly by the defects growth and microstructure[14]. Fatigue is one of the important factors to see the structural integrity of the direct energy deposition printed part. By determining the probable fatigue initiation site and fatigue crack growth we can estimate the fatigue life of the part processed by DED[15]. Growth of fatigue crack was mainly found in the plane with some cracks moving in the direction of tensile force. It was found that the crack growth rate was location-

dependent and also was varying along with different directions[16]. As reported by several authors fatigue results were not consistent. Such as, Ti-6Al-4V processed by LENSTM had greater fatigue life as compared with cast Ti-6Al-4V[17]. In one research it was found out that DED processed Ti-6Al-4V has alike properties to that of cast Ti-6Al-4V and heat-treated DED Ti-6Al-4V had alike properties to that of wrought Ti-6Al-4V[18]. It was seen that the fatigue life improved by hot isotactic pressing of the DED processed because it closes the pore sites inside that part.[19].

Residual Stress

Because of the occurrence of a steep thermal gradient between surrounding material and heat source residual stresses are generated during DED or any other additive manufacturing methods. By the cracking and distortion, residual stress may damage the printed parts. It was found out in the research that in the centre compressive stresses were present while at the surface tensile stresses were found. Generally, in dissimilar metals, residual stresses were found to be higher [20]. Residual stresses in additive manufacturing can be categorized into two types based on the length scale on microscale and nanoscale and the macroscale[21]. Some of the most common methods to reduce the residual stresses are: using in-situ process in which process parameter are monitoring with the feedback loop to tune them; by decreasing the thermal gradient with preheating the initial substrate; post-processing methods such as heat treatment can relieve residual stresses[22].

3D Printing

Tensile Strength

Various studies have discovered that, for all materials, 3D printed material's tensile strength is majorly dependent on the specimen's mass. The tensile strength can be estimated using a two-step process.

First, the printed material's exterior is visually tested for sub-optimal layers caused by under and over extrusion[22]. The mass of samples is then reported to determine if the substance is under extrusion. In a study on ABS, Nylon 618, HIPS, PC, T-Glass, Semiflex, Ninjaflex. Polycarbonates were found to be strongest with a tensile strength of 49Mpa. And most

flexible was found out to be Ninjaflex as it did not fracture till an extension of 800%, with a tensile strength of 12Mpa at 800% extension. Nylon was found out to be stronger than Ninjaflex and more flexible than most materials, which contribute to a good combination of strength and flexibility. [22].

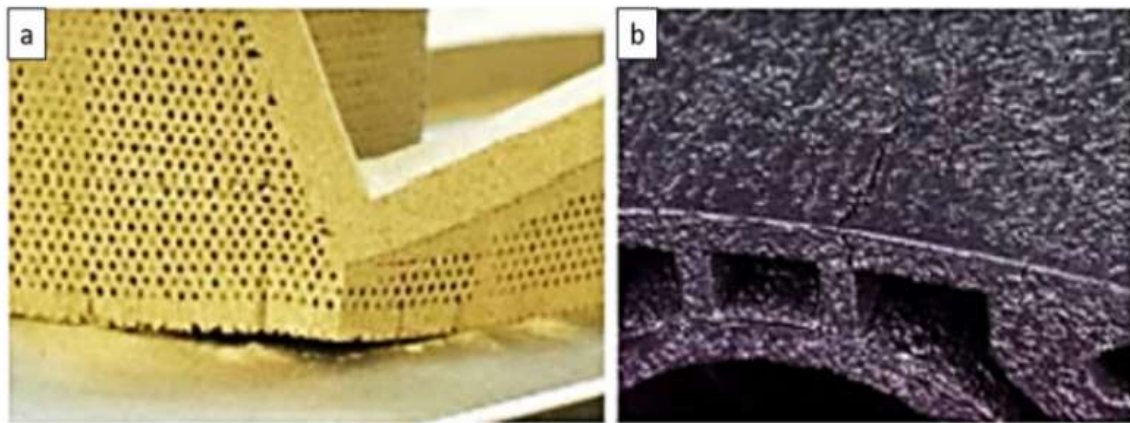


Fig.4:- The results of residual stresses in 3D printed parts[23] (a) distortion and separation from the base plate (b) crack formation

Residual Stresses

In 3D printing, materials usually go under repeated contraction and expansion from cooling and heating during the printing process. This leads to residual stresses in the printed material Which further leads to cracks, warpage, and other forms of deformation[23].

The Figure 4 shows the results of residual stresses in 3D printed parts (a) separation and distortion from the base plate (b) formation of crack For these reasons several methods are used to compensate the residual stresses some of which are simulating print scenarios to get optimal print conditions to ensure proper printing and varying post-processing treatments are used to reduce the effect residual stresses. Ultrasonic waves can be used during both the build and after the build. In which short pulses are sent of ultrasonic waves which help in detecting internal flaws[24].

Surface Finish

In comparison to traditional machining techniques, 3D printed metals have rougher surfaces due to various factors which are 3D printing has layer by layer deposition mechanism which yields rougher surfaces, metal powder is required as raw material, partially melted particles are found in 3D printed metals which leads to rougher surfaces. The surface quality of 3D printed metals can be improved with the help of machining, polishing, and various surface hardness treatments such as shot peening, high-frequency mechanical impact treatment[25].

Fatigue Behaviour

Fatigue in the case of polymeric 3D printed materials depends on various factors which are environmental conditions, physical properties, mechanical loading factors as shown in Figure 5 [26]

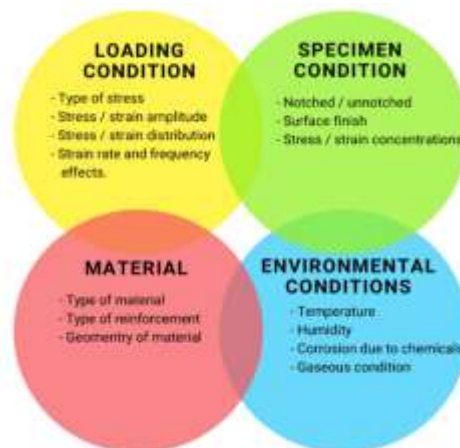


Fig.5:-Factors which effect the fatigue life of polymeric materials[26].

Fused deposition modelling based fibre composites had better bending and tensile strength than neat polymers. The fibre properties of FDM printed composite could help in higher fatigue life[26]. In the case of 3D printed metals factors which helps in better fatigue life are optimal heat input, correct building direction, better surface finish, and various heat treatment processes[25].

APPLICATIONS

DED

In Repairing components

Repairing is one of the most important operations to extend the life of parts and enhance their functionality. Repairing help in reducing environmental impact, as a lower amount of material and energy waste occurs. DED is used as a standard repairing technique in gas turbines in which Ni-based superalloy is given through coaxial powder feeder[27], and repairing of steam circuit parts at the thermal power station, by depositing Co-based alloys on them in order to keep the high temperature bearing thermal properties [28]

In construction materials

Traditional casting makes primitive structures(uniform microstructure) but with DED engineered microstructure is possible, which might have better mechanical

properties. Using DED in construction could help largely in tackling greenhouse gases as 30% of greenhouse emission is due to construction industries which could be decreased by partially adopting DED for specialized parts[29].

In cladding and welding of parts

During the welding of dissimilar material, traditional welding produces high residual stresses at the interfaces which might lead to early failure and the outcome can be dangerous. Whereas in DED the composition is the function of position, which results in a seamless transition from one joint to the other. This can decrease the residuals and increase the mechanical integrity of the joints[30]. Multi-axis cladding can be done with the help of DED through which it is possible to deliver material at any angular positions.

Various Biomedical applications

DED alters the mechanical properties of a material by varying the geometry or orientation of the printed parts. It is more beneficial to make porous implants using Direct energy deposition as compared to traditional casting[31]. It has gained priority in the dental and implant industry. Mostly Co-based alloys, Ni-Ti based alloys, Ti and its alloys, 316L stainless steel. DED is used in biomedical industries since

it can combine different materials to achieve optimal properties, and it is easier to build custom implants as per patients need. Biocompatibility tests via LENS™ on porous Ti-6Al-4V proved that it has the ability of cell growth on implants having a pore size of 200 micrometer or greater than that. [31].

3D Printing

Medical Applications

3D printing have a huge implementation in this field some of the major applications are

Organs can be printed which can be used as a substitute to real body parts such as titanium jaws, titanium pelvic etc. tissues engineering has great strides as it can print 3D blood vessels[32]. Additive manufacturing of stem cells has to lead to various possibilities in printing artificial organs and we can say an endless world of possibilities are waiting due to 3D printing[33].

The Figure 6 shows 3D printed human heart which is made by bioprinters which can print human tissues[34]

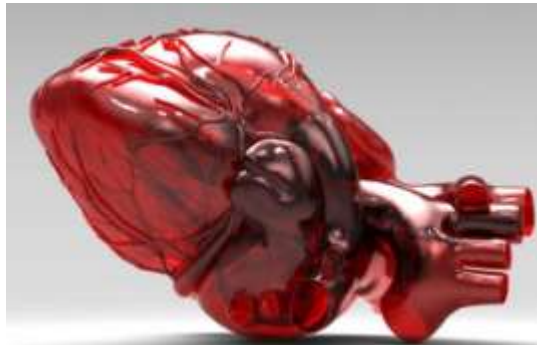


Fig.6:- 3D printed human heart which is made by bioprinters which can print human tissues[34]

Nowadays, people are using 3D printed teeth which can be customized as per patients and also hearing aid are being made using 3D printing.

Aerospace and Automotive Industry

3D printing has great potential in the aerospace industry as it can be used to create lightweight parts, improved and complex geometries. At the same time it can reduce the material which in turn will help in the reduction of fuel usage.[35]. It can be used to print automotive parts as it was done by a local motor which printed the first electric 3D printed car and they also created the first 3D printed bus named 'OLLI' which is driverless, recyclable, electric 3D printed bus. Ford has also used 3D printing to print protoengine parts and prototypes of the car also BMW uses 3D

printing to printing hand tools which are then used in automotive testing[36].

Food Industry

3D printing has a great future ahead in the food industry as it is used to print customized foods which can remove unnecessary things from food and enhance the presence of vital nutrients.[37]. therefore food can be customized as per individual preferences. And food can have higher quality and lower costs. It can help in a world where people can have a diet that doesn't enforce exercise[38].

3D Printing in Various Other Industries

- 3D printing is being used in industries like the electric and electronic industry for the manufacturing of structural electronic devices like electrodes,

electronic materials etc. It can help in providing low cost and higher time efficiency for mass production of materials[39].

- 3D printing has entered in fabric and fashion industry for the manufacturing of 3D printed shoes and jewellery. For instances, large companies such as Nike and Adidas have applied 3D printing for the mass production of athletic shoes [40]
- 3D printing technology can also be applied to the construction industry, where it can be used to construct whole buildings or individual parts. Also in China, they built ten single-story houses in a day which requires months to be complete by traditional methods. Therefore 3D printing is a faster, cheaper and safer method[41].

CONCLUSION

The above work is the utmost effort to discuss various spray powder additive manufacturing and learn about the process and how the functioning of 3D printing and DED takes place, discuss about the mechanical properties such as fatigue, residual stresses and hardness of both DED and 3D printing manufactured products. We have also tried to explain the various applications of both DED and 3D printing over various domain. So as we can infer from all this these processes will work as a revolution in future in manufacturing of various component and making the components more efficiently. Apart from manufacturing of machine component both 3D printing and DED has a great future in other fields such as construction, medical industry and fashion products and even food industry. We have tried to talk about all these domains in very precise and brief manner so that the reader will understand them. And can do further research in the

specific domain as per their interests.

REFERENCES

1. Pinkerton, A. J., Ul Haq Syed, W., & Li, L. (2006). An experimental and theoretical investigation of combined gas-and water-atomized powder deposition with a diode laser. *Journal of Laser Applications*, 18(1), 73-80.
2. Shin, Y. C., Bailey, N., Katinas, C., & Tan, W. (2018). Predictive modeling capabilities from incident powder and laser to mechanical properties for laser directed energy deposition. *Computational Mechanics*, 61(5), 617-636.
3. Dass, A., & Moridi, A. (2019). State of the art in directed energy deposition: from additive manufacturing to materials design. *Coatings*, 9(7), 418.
4. Tofail, S. A., Koumoulos, E. P., Bandyopadhyay, A., Bose, S., O'Donoghue, L., & Charitidis, C. (2018). Additive manufacturing: scientific and technological challenges, market uptake and opportunities. *Materials today*, 21(1), 22-37.
5. Lboro.ac.uk. 2021. *Binder Jetting / Additive Manufacturing Research Group / Loughborough University*. [online] Available at: <<https://www.lboro.ac.uk/research/amrg/about/the7categoriesofadditivemanufacturing/binderjetting/>> [Accessed 2 March 2021].
6. Ziaee, M., & Crane, N. B. (2019). Binder jetting: A review of process, materials, and methods. *Additive Manufacturing*, 28, 781-801.
7. General Electric "7 Things You Didn't Know About 3D Printing," mashable.com, Dec 3, 2013 [Online]. Available:<http://mashable.com/2013/12/03/3d-printingbrandspeak/>. [Accessed: June. 18, 2014].
8. Sandgren, H. R., Zhai, Y., Lados, D. A., Shade, P. A., Schuren, J. C., Groeber, M. A., ... & Gavras, A. G.

- Characterization of fatigue crack growth behavior in LENS fabricated Ti-6Al-4V using high-energy synchrotron x-ray microtomography, *Addit. Manuf.*(2016).
9. Beese, A. M., & Carroll, B. E. (2016). Review of mechanical properties of Ti-6Al-4V made by laser-based additive manufacturing using powder feedstock. *Jom*, 68(3), 724-734.
 10. Carroll, B. E., Palmer, T. A., & Beese, A. M. (2015). Anisotropic tensile behavior of Ti-6Al-4V components fabricated with directed energy deposition additive manufacturing. *Acta Materialia*, 87, 309-320.
 11. Shamsaei, N., Yadollahi, A., Bian, L., & Thompson, S. M. (2015). An overview of Direct Laser Deposition for additive manufacturing; Part II: Mechanical behavior, process parameter optimization and control. *Additive Manufacturing*, 8, 12-35.
 12. Kistler, N. A., Corbin, D. J., Nassar, A. R., Reutzel, E. W., & Beese, A. M. (2019). Effect of processing conditions on the microstructure, porosity, and mechanical properties of Ti-6Al-4V repair fabricated by directed energy deposition. *Journal of Materials Processing Technology*, 264, 172-181.
 13. Zuback, J. S., & DebRoy, T. (2018). The hardness of additively manufactured alloys. *Materials*, 11(11), 2070.
 14. Nalla, R. K., Ritchie, R. O., Boyce, B. L., Campbell, J. P., & Peters, J. O. (2002). Influence of microstructure on high-cycle fatigue of Ti-6Al-4V: Bimodal vs. lamellar structures. *Metallurgical and Materials Transactions A*, 33(3), 899-918.
 15. Lewandowski, J. J., & Seifi, M. (2016). Metal additive manufacturing: a review of mechanical properties. *Annual review of materials research*, 46, 151-186.
 16. Sandgren, H. R., Zhai, Y., Lados, D. A., Shade, P. A., Schuren, J. C., Groeber, M. A., ... & Gavras, A. G. Characterization of fatigue crack growth behavior in LENS fabricated Ti-6Al-4V using high-energy synchrotron x-ray microtomography, *Addit. Manuf.*(2016).
 17. Kobryn, P. A., & Semiutin, S. L. (2001). Mechanical properties of laser-deposited Ti-6Al-4V. In *2001 International Solid Freeform Fabrication Symposium*.
 18. Prabhu, A. W., Vincent, T., Chaudhary, A., Zhang, W., & Babu, S. S. (2015). Effect of microstructure and defects on fatigue behaviour of directed energy deposited Ti-6Al-4V. *Science and Technology of Welding and Joining*, 20(8), 659-669.
 19. Zhai, Y., Lados, D. A., Brown, E. J., & Vigilante, G. N. (2016). Fatigue crack growth behavior and microstructural mechanisms in Ti-6Al-4V manufactured by laser engineered net shaping. *International Journal of Fatigue*, 93, 51-63.
 20. Ian Gibson, I. G. (2015). Additive Manufacturing Technologies 3D Printing, Rapid Prototyping, and Direct Digital Manufacturing.
 21. Li, C., Liu, Z. Y., Fang, X. Y., & Guo, Y. B. (2018). Residual stress in metal additive manufacturing. *Procedia Cirp*, 71, 348-353.
 22. Tanikella, N. G., Wittbrodt, B., & Pearce, J. M. (2017). Tensile strength of commercial polymer materials for fused filament fabrication 3D printing. *Additive Manufacturing*, 15, 40-47.
 23. Langnau, L., 2021. *3D Printing Residual Stress: What Is It? - Make Parts Fast*. [online] Makepartsfast.com. Available at: <<https://www.makepartsfast.com/3d-printing-residual->

- stress/#:~:text=In%20some%203D%20printing%20additive,of%20deformation%20in%20an%20object> [Accessed 2 March 2021].
24. 3DPrint.com | The Voice of 3D Printing / Additive Manufacturing. 2021. *Using Ultrasonic Waves to Analyze Residual Stress in 3D Printed Metal Parts - 3DPrint.com | The Voice of 3D Printing / Additive Manufacturing*. [online] Available at: <<https://3dprint.com/269334/reviewing-am-3d-printing-residual-stress-analysis-ultrasonic-waves/>> [Accessed 2 March 2021].
 25. Mfg40.fi. 2021. [online] Available at: <<https://mfg40.fi/wp-content/uploads/2019/12/Fatigue-Properties-of-3D-Printed-Metals.pdf>> [Accessed 2 March 2021].
 26. Shanmugam, V., Das, O., Babu, K., Marimuthu, U., Veerasimman, A., Johnson, D. J., ... & Berto, F. (2020). Fatigue behaviour of FDM-3D printed polymers, polymeric composites and architected cellular materials. *International Journal of Fatigue*, 143, 106007.
 27. Bi, G., & Gasser, A. (2011). Restoration of nickel-base turbine blade knife-edges with controlled laser aided additive manufacturing. *Physics Procedia*, 12, 402-409.
 28. Díaz, E., Amado, J. M., Montero, J., Tobar, M. J., & Yáñez, A. (2012). Comparative study of Co-based alloys in repairing low Cr-Mo steel components by laser cladding. *Physics Procedia*, 39, 368-375.
 29. Buchanan, C., & Gardner, L. (2019). Metal 3D printing in construction: A review of methods, research, applications, opportunities and challenges. *Engineering Structures*, 180, 332-348.
 30. Hofmann, D. C., Roberts, S., Otis, R., Kolodziejska, J., Dillon, R. P., Suh, J. O., ... & Borgonia, J. P. (2014). Developing gradient metal alloys through radial deposition additive manufacturing. *Scientific reports*, 4(1), 1-8.
 31. Xue, W., Krishna, B. V., Bandyopadhyay, A., & Bose, S. (2007). Processing and biocompatibility evaluation of laser processed porous titanium. *Acta biomaterialia*, 3(6), 1007-1018.
 32. Shannon Stapleton “‘Critical challenge’: Doctors can now 3D-print blood vessels,” rt.com, June 01, 2014 [Online]. Available: <http://rt.com/news/162848-3d-print-blood-vessels/>. [Accessed: June. 18, 2014].
 33. “3D Printing in Medicine: How Technology Will Save Your Life,” August 13, 2013, [Online]. Available: <http://www.cgtrader.com/blog/3d-printing-in-medicine-how-technology-will-save-your-life/>. [Accessed: June.18, 2014].
 34. The Daily Telegraph, “3D printing of human organs,”timeslive.co.za, 2 September, 2013 [Online]. Available:
 35. Joshi, S. C., & Sheikh, A. A. (2015). 3D printing in aerospace and its long-term sustainability. *Virtual and Physical Prototyping*, 10(4), 175-185.
 36. M. Petch, “Audi gives update on use of SLM metal 3D printing for the automotive industry,” 3D Printing Industry, 2018. [Online]. Available: <https://3dprintingindustry.com/news/audi-gives-update-use-slm-metal-3d-printing-automotive-industry-129376/>. [Accessed 2019].
 37. Dankar, I., Pujolà, M., El Omar, F., Sepulcre, F., & Haddarah, A. (2018). Impact of mechanical and microstructural properties of potato puree-food additive complexes on extrusion-based 3D printing. *Food and Bioprocess Technology*, 11(11), 2021-2031.

38. Liu, L., Meng, Y., Dai, X., Chen, K., & Zhu, Y. (2019). 3D printing complex egg white protein objects: properties and optimization. *Food and Bioprocess Technology*, 12(2), 267-279.
39. Lee, J., Kim, H. C., Choi, J. W., & Lee, I. H. (2017). A review on 3D printed smart devices for 4D printing. *International Journal of Precision Engineering and Manufacturing-Green Technology*, 4(3), 373-383.
40. S. Horaczek, "Nike hacked a 3D printer to make its new shoe for elite marathon runners," Popular Sciences, 2018. [Online]. Available: <https://www.popsci.com/nike-3d-printed-sneakers>. [Accessed 2019].
41. Mellisa Goldin "Chinese Company Builds Houses Quickly With 3D Printing," mashable.com, April 29 2014, [Online]. Available: <http://mashable.com/2014/04/28/3d-printinghouses-china/>. [Accessed: June. 18, 2014].

Cite as

Md Zia Arzoo, Mozammil Hassan, & A.K.Madan. (2021). Review on Spray Powder Process of Additive Manufacturing. Recent Trends in Automation and Automobile Engineering, 4(1), 1–11.
<http://doi.org/10.5281/zenodo.4629049>

Sensory Vision Substitution using Tactile Stimulation

Pavitra Gandhi
Information Technology
Delhi Technological University
Delhi, India
pavitra.gandhi96@gmail.com

Anamika Chauhan
Information Technology
Delhi Technological University
Delhi, India
anamika@dce.ac.in

Abstract—When a person loses vision, he/she ordinarily does not lose the capability to see; they get deprived of the ability to relay the sensory signals to the brain which is evident in congenitally blind people. It has been said that brain is a highly flexible task-machine which means that even with input from alternate senses the regions of brain can retain and conserve their ability to visualize which is the plasticity property of human brain. This is where sensory substitution comes into the picture. Sensory vision substitution has come a long way since it first came into existence. This concept was realized into a device called sensory substitution device (SSD) which produced exciting experimental results and even assisted in further understanding how human brain functions. In this review, various mechanisms designed for visual restoration with its technical and biological functioning will be discussed. Towards the end, how effective the models are at solving the problem with further scope of improvements will be elaborated.

Keywords—brain plasticity, electrotactile, neuromodulation, somatosensory, vibrotactile

I. INTRODUCTION

In this review we have shed light on how we can retrieve sensory perception by exploiting sensory substitution to assist the visually impaired. Section 2 introduces, how human brain perceives vision, its properties and its related disorders and impairments. Section 3 briefly explains what SSDs are their role in sensory substitution and also lists down different types of stimulators being fabricated and tested and further development for better functionality and accuracy. Section 4 gives a detailed explanation about the design and functioning of tactile vision substitution system (TVSS) and also presents some of the other TVSS-based designed systems having better performance. Section 5 and 6 states some applications and concludes this review and provides some opinions on future scope.

II. VISION AND PROBLEMS IT FACES

A. Functioning of Visual Perception

Visual perception can be defined as the ability to construe the surrounding environment using the incident light reflected by the objects in the environment. We can divide human visual perception into two parts:

- First part constitutes scene acquisition (capturing light) where the light rays are focused on retina which converts light energy into electrical energy patterns which are further processed and transmitted via visual pathway to visual cortex in central nervous system which is evident in Fig. 1.

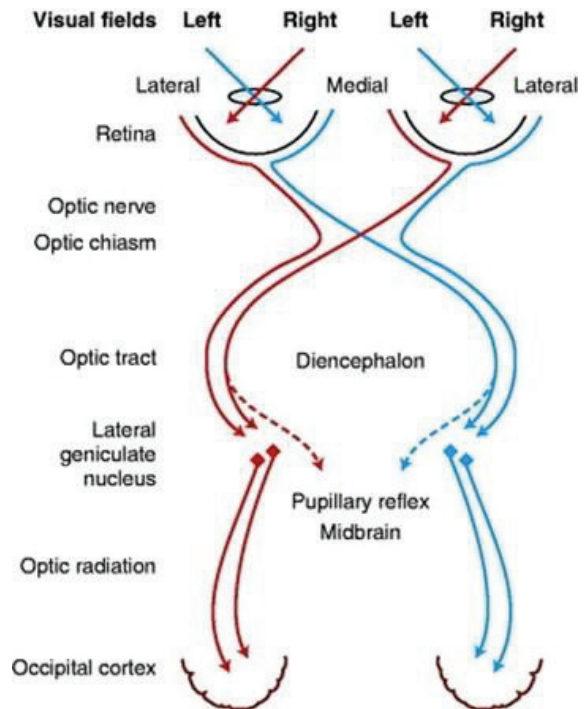


Fig. 1. Visual Perception - how light from external source reaches the brain via visual pathway. [23]

- Second part is the visual perception which is a knotty integration of various parameters which are light sense, form and contrast sense, as well as sense of color. The receptive field organization of the retina and visual cortex are used to encode this information about a visual image.

Apart from the above-mentioned parameters another important parameter is depth. Keeping the eyes as well as the head maintained in a fixed position and one eye closed, what one sees is the area comprised of visual field for that eye. If the second eye is opened, the area it perceives is almost the same as what was visible with one eye. Hence, it can be said that a large part of the visual fields of both the eyes converges. However, there is a small part of area that is only visible from one eye. Even though, almost same areas are visualized by both the eyes, the small difference in vision becomes the cause of the depth perception of the environment. This happens because of the dissimilarity in the relative position of the target object present within the area of the two eyes, as these objects are viewed from marginally different angles which holds significant importance for stereoscopic (depth) vision.

B. Target Problem

Vision plays a critical role in assisting humans for existence. Hence, it becomes next to impossible to survive without vision. Vision is very much taken for granted but when it deteriorates, one struggles while walking, reading, working and while participating in daily crucial activities for living. It's been consistently stated that damage to vision severely impacts quality of life. As per WHO reports [1], globally, more than 2.2 billion people face vision impairment or worse (blindness), of whom more than 1 billion of them face vision impairment, which if taken care of, could have been averted or are yet to be tackled. 80% of visual impairment is either preventable or curable with treatment which includes quality eye care and rehabilitation. Severe conditions, however, may require surgical treatment and their cost increases radically. In vision impairment either the eye which processes and transmits the spatially encoded pulse is damaged or the vision processing pathways and visual cortex is compromised. The scope of this survey is limited to the problem where the person has just lost the capability to relay image from the eye (retina) and the vision processing pathways are still intact [2].

C. How Plasticity is basis of Visual Rehabilitation?

Brain plasticity [3], is the capability of the brain to reform continuously all through the person's life. The adult brain is not completely hard-wired. The neural connections keep changing throughout the lifetime. Sensory substitution devices are in existence because of "instrumental sensory plasticity" which is nothing but the ability of the brain to reform and reorganize. This can occur when there is: (a) a functional demand, (b) some sensor technology to achieve that demand, and (c) training and psycho social factors that keep up with that particular functional demand.

Brain plasticity alternatively can be defined by classifying into 2 categories: structural plasticity where brain's physical structure and its functional plasticity is changed by experiences or memories. In this case the functions of the brain shifts from to undamaged area from damaged area. With an intention to exploit both structural and functional plasticity, to overcome sensory loss specifically vision loss, renowned neuroscientist Dr. Paul Bach-y-Rita, also well-known as "father of sensory substitution", started working on sensory substitution system, producing exciting results. Since then he has published many of his researches which will be summarized in this paper. Tactile vision substitution system (TVSS), termed by him, is based on the principle of brain plasticity in which one sensory modality is used to retrain another sensory modality which has lost its sensation, for example, vision.

III. SENSORY SUBSTITUTION DEVICE (SSD)

A. Can the use of SSDs be thought-out as "Seeing"?

SSDs have been studied significantly for vision impaired since a long time in laboratories. This survey further elaborates how sensory substitution model functions with its practical implementation. Sensory substitution is grounded on the ideology that the brain holds capacity to figure out sensory data even though the source is not a natural channel. Information from touch receptors is transduced into an electric action potential to the visual cortex which is otherwise the job of photoreceptors in the eye.



Fig. 2. A particular type of tactile tiles used in parking lots and streets. [4]

A very relevant and enlightening but a very common example is tactile paving [4] as shown in Fig. 2. They are used on footpaths, stairs and train stations for visually impaired, moreover blind, and they are used as ground surface indicator. Tactile warnings alert the visually impaired if he/she is approaching streets or any hazardous surface or any grade changes. This is done by providing a unique surface pattern of pruned domes or cones which can either be detected by long cane or person's underfoot. Hence, we can see that the visually impaired indirectly visualizes environment around him/her.

Another non-invasive brain-machine interface was Braille reading, a SSD. Here the blind uses the sense of touch, by moving the hand from left to right along each line designed using raised dots representing letters of the alphabet and punctuation marks too, to fetch information for the sense of vision. The user actually visualizes the information using touch. This technique was originally developed by Barbier and with the rise in innovation and technology, it was later developed to automatic text-to-braille converters for example Optacon. Studies demonstrated positive results with respect to feasibility. To see this in action, fMRI was used, which is a noninvasive test to measure minute changes in flow of blood because brain activity and hence finds out the sections of the brain which are triggered during perception of sense. In blind people, it was found that after receiving tactile information itself, along with somatosensory cortex some visual cortex is even activated as they try to see objects.

We can categorize SSDs based on the type of feedback system which are of two types mainly: vibrotactile and electrotactile feedback systems.

B. Vibrotactile Based Feedback Systems

Vibrotactile stimulators trigger action potentials by using pressure and the characteristics of the mechanoreceptors of the skin which are a necessity for the vibrotactile sensory perception. The vibrotactile perception depends mainly on two types of corpuscles namely Pacini and Meissner's corpuscles which are fast adapting receptors and Merkel discs being slow adapting ones. Depending on the skin surface, several characteristics of vibratory perception in the periphery were found. A detailed explanation on sensory physiology for touch is given by Kurt A. Kaczmarek in [5]. He described the characteristics of different tactile receptors of human skin. From a technical point of view, the vibration sinusoid has two different significant characteristics (firing frequency and amplitude), which hold diverse properties and will generate distinct features in the vibro-tactile perception [6].

Based on these, Optacon was designed which was one of the widely known devices based on vibrotactile based feedback systems. It is an electro-mechanical device which helps the blind to read printed matter which cannot be reproduced into Braille [7]. The Optacon has capabilities which are not offered by any other device. It provides the ability to visualize a printed page or whatever actually appears on a computer consisting of drawings, fonts, and special text designs. The blind user places their index finger onto a tactile array of 24-by-6 matrix of tiny metal rods, being one part of its electronics unit [8]. The user then scans the camera module across the printed line and an image formed, which has the size of a letter space, is transmitted to the main electronics unit via a connecting cable. In the electrode, rods that correspond to black parts of the image vibrate, thus forming a tactile image of the letter visualized by the camera module. The camera module when moved by the blind across the printed line creates a sensation of the image formed by tactile perception of printed letters under user's finger by a matrix of rods.

C. Electrotactile Based Feedback Systems

Electrotactile stimulator-based feedback systems are the whole focus of this review and the concepts governing them are a bit complex. But the basic idea is, by activation of sense of touch by stimulating skin with electric current, non-tactile information is relayed to the brain. Practically, this means that inputs from a device like a camera is fed into a group of electrodes called an array, which then provide small, regulated and painless current at precise locations on the skin based on how the signal is modulated and encoded. A more technical explanation is they are array of electrodes used to initiate the action potentials in superficial nerve endings by giving electrical stimulation to the skin. Due to electric discharge there can be sensations of burns, itching, pain and pressure.

In order to apply non-tactile vision signal which is an encoded pulse, there are a set of electrode-skin interface parameters that are needed to be taken care of which are (a) stimulation voltage, (b) current, and waveform (c) size and material of the electrodes and (d) the location, thickness, and

hydration of the skin. Lower impedance or resistivity, better conductivity and higher sensitivity to touch of the skin are also among the prime factors to be taken into consideration for better results. He tested his designed electrotactile stimulator/electrodes, on different skin interfaces of the human body and performed a comparative study to realize which one resulted the best outcome. He designed and tested his electrodes on different anatomical regions like the fingers, the back, the abdomen in the front, the forehead, and the tongue [9].

His test results made evident that dry skin is an insulator and thus offers high impedance when the skin is stimulated directly, so there is a need of high voltage current for stimulation. Normally the touch sensations are relayed to the parietal lobe while the visual sensations are relayed to occipital lobe. Electrical stimulation of the skin excites the afferent nerve fibers which take the encoded signals to the parietal lobe.

Stimulation mechanism and skin-electrodes characteristics: The current in neurons flows when its receptors are stimulated and an action potential is generated. The current (action potential) flows by the exchange of ions across the axonal plasma membrane. Hence a stimulator (electrode) is used to stimulate the receptors present on the neuron. Many investigators suggest that the afferent fibers are directly stimulated by the stimulator [10]. Pfeiffer in his review suggested that receptors of the skin are stimulated directly if small electrodes of the size of 1 mm^2 are used [11].

If performance of the system was measured, a compromise with comfort was observed. For an instance, the electrotactile stimulation in presence of perspiration or an electrolyte conductor between electrode and skin is most comfortable. The presence of electrolyte is a must as it bridges the flow of electrons in the electrode and the stimulation of receptors on neurons and also it stabilizes the electrode-skin adhesion.

The behavior of current distribution under the electrodes at the microscopic level was observed and it was found that

TABLE I. THRESHOLDS OF DIFFERENT PARAMETERS OF ELECTROTACTILE SENSATION AND PAIN/SENSATION RATIOS [5]

Electrode type/material	Body location	Electrode area (mm^2)	Waveform	Frequency (Hz)	Sensation current (mA)	Sensation Charge (nC)	P/S
Silver Coaxial	Abdomen	15.9	M-	60/200	20	40	8
					0.1	70	
Gold, silver coaxial gelled	Abdomen	11	M-/B	Single pulse	5.0	312	?
					1.5	1500	
			M+	Single Pulse	6.1	381	
					2.5	2500	
Silver Square	Wrist	49	M	Single pulse	2.7	270	?
					1	1000	
Stainless Steel Coaxial gelled	Trunk	8.42	M	Best frequency 1-100 Hz	1.5	150	1.6
	Fingertip	8.42	M		6	600	
Stainless Steel/ Aluminum gelled	Abdomen	0.785	M	50	0.4	100	6.25
Steel electrode pair	Fingertip	0.0078	M	Best frequency 1-200 Hz	(a) 0.2	100	1.5
					(b) 1.0	500	
Coaxial	Forearm, back, abdomen	7.07	?	25	17	17	8.4
					2.5	250	

^a M stands for monophasic, + or - indicated if known; B stands for biphasic (c), (d) 0.79 and 6.35-mm electrode spacing respectively. P/S is pain/sensation current ratio.

for any electrode type the conductive path through the skin is non-uniform. Saunders [12] and Grimnes [13] showed that current takes small regions of low resistance, for example small breaks in the epithelium (1-6/mm² of skin area), sebaceous glands and sweat ducts, as its path to flow. Under metal electrodes of size > 100 mm² (categorized as large electrodes) on dry skin, it was found that resistance of one of the conductive paths of the skin, drops suddenly from time to time hence much of the current of the electrode faces a shunt through that pathway [14]. This results into a high current density which leaves a red spot mark on the skin surface. The user also feels a sudden sharp sting which is the likely to occur because of electrodes which are negatively-pulsed. Kurt A. Kaczmarek has further discussed in some detail in [5]. So far most of the sensory substitution systems have considered metals to design electrodes; the most common metals being (a) gold, (b) platinum, (c) silver, and (d) stainless steel. It is also necessary to consider comfort and safety when it comes to practical implementation of TVSS.

There are various factors which define the comfort and pain sensation levels because of electrotactile stimulator. There can be mild discomfort to intolerable pain and it varies with training and psychological condition of the test subjects. Experienced and trained subjects can bear as much as twice the exposure of stimulation levels as naïve subjects. Threshold of Pain/Threshold of Sensation that is P/S ratio proves to be a useful attribute. Table I shows how P/S ratio varies as a function of (a) electrode material, (b) size and (c) placement and (d) even different attributes of stimulation waveform. It also summarizes the test results of various investigators. [5]

IV. TACTILE VISION SUBSTITUTION SYSTEM (TVSS)

Bach-y-Rita designed TVSS as a visual prosthetic system which is non-surgical and non-invasive method. It takes input from a camera and converts visual information to gentle electrical signals which are given on tongue. TVSS was also used to study brain plasticity.

A. Why “Tongue” was chosen for Sensory Perception and Associated Complications?

After many trials with different electrotactile interface areas an ideal interface for tongue was developed for sensory perception. Tongue demonstrated significantly better performance over other anatomical regions that were tested. It was found that only 3% of voltage about 5-15V was required and comparatively lesser current of about 0.4-2.0 mA than other places like fingertip. The tongue has a higher representation in the cortex of brain in comparison to fingertips, forehead, back, abdomen, etc. In other words, the upper surface of tongue is very sensitive to sense of touch, both in terms of (a) spatial acuity and (b) pressure sensitivity.

The tongue is present inside the mouth; and it doesn't have the top keratinized layer of dead cells which is usually present on the skin of other areas of body. The saliva present in the mouth and on the surface of tongue act as a conductor which results in better conduction of current and thus better signal transmission than at other sites. The cutaneous receptors (the somatosensory receptors) lie close to the tongue surface. These properties make tongue a better option for tactile stimulation.

A recent research [15] on tongue-based electrotactile brain-machine interface revealed that an electrotactile system was



Fig. 3. A broad perspective of tactile vision-substitution system. [16]

specifically designed to study the properties of the human tongue. This system found that there is non-uniform perception of sensation intensity and varying electro-tactile stimulation (ETS) dynamic range of the tongue in both posterior-anterior and lateral-medial directions. The reason behind this irregularity was: difference in type and density of tactile receptors and differential innervations of the tongue. Having established this fact, a spatial map had to be developed of the electrotactile percept magnitude across the area stimulated. With the help of this map, it can be ensured that using an algorithm system designed specifically to nullify the different patterns in local touch sensitivity, any ETS pattern that is presented across the tongue maybe appropriately and unvaryingly perceived.

B. Block diagram-based design of TVSS

TVSS, point by point, plots an optical image from the outside world onto the skin surface, to provide a perception of vision through the skin. With training, a totally blind user becomes capable of developing perceptual skills just like how the input from eyes is used generally for the same. Dr. Paul Bach-y-Rita's results of all his experiments made evident how blind test subjects could develop and visualize various perceptual concepts like perspective, looming, and parallax.

To explain the working model of TVSS designed, we will describe it using a block diagram as shown in Fig. 3. It has four main blocks: 1) an optical sensor (i.e. a digital camera), 2) a data processor for image processing, and 3) a cluster of electrodes [16].

For the optical sensor, any digital or television camera could be used along with a digital frame grabber which captures individual still frames from both an analog or digital video stream. A digital fingerprint is formed by either of the device used of the information present in the image, fetched from the camera, which is actually a numerical format of the image. A number according to the level of brightness of a pixel (elements) in the image captured, having finite number of pixels, was assigned. Even though usage of multiple numbers for different levels of brightness was possible but the system reviewed here used just two (black and white) levels. A sequential raster-scan format was used to store the image data. In raster scan format the top row points/elements/pixels are stored first, then the next lower row points, and so on until all rows are stored.

Coming to the second block, the image processor alters and formats the image such that we can get a highly efficient tactile representation for the brain to perceive. It can be seen as a data processor or a computer program that performs several image processing techniques such as edge detection and edge enhancement to outline the objects that are useful for improving image perception. To consecutively present the small sections of the captured image time division

multiplexing (TDM) was used. This also led to an increase in the effective overall spatial resolution.

The processed image was further sent to the transducer, a cluster of tactile stimulators or electrode matrix, in the form of encoded electronic pulses. The encoded pulse is nothing but a representation of the camera image in the form of amplitude, frequency and pulse duration which are the basic pulse attributes. If the image is multidimensional, then it can be represented based on variations in pulse voltage and current, duration of pulse, intervals between two pulses and various other parameters. The set of electrodes presents the encoded pulse to the blind user as a array of optical pixel points on the surface of skin. Each of the electrodes are triggered with a current directly proportional to the corresponding pixel brightness/intensity of the image so as to generate a spatial distribution of pixels of the image on the skin.

C. Electrotactile Stimulators Designs and their Test Result

A wide range of designs for an electrotactile interfaces of stimulators were designed and fabricated and tested, over the last few decades, for different body surfaces. The test results then helped in tweaking the designs based on which tongue was chosen as an ideal organ for sensory perception and an array of tactile stimulation electrodes was designed for the tongue. Collins had proposed a portable design having 64 electrodes in an array [17]. Vuillerme et al. [18] developed and proposed a stimulation matrix of 6×6 electrodes for the tongue for the purpose of balance improvement and proprioception. Further studies of electrode array comprised of a 7×7 array of 1.8×1.8 cm, that is a 49-point array of electrodes with a diameter of 8.89 mm stainless-steel electrode “pins” with flat-topped structure as shown in Fig. 4. Each of the electrodes were surrounded by an air gap insulator of 2.36 mm diameter. A flat stainless-steel plate, which is co-planar with the electrode pins, was kept in order for the current to have the return path. This whole setup of electrodes was arranged on a square grid having interelectrode spacing of 2.54 mm [19].

With this design of set of electrodes, TVSS was tested on 5 test subjects –A) 3 Males and B) 2 Females, all of them being sighted. They were subjected to 12 different tactile patterns that were 2-D approximations of a)square, b)circles and c)vertex-up equilateral triangles of size of 4×4 , 5×5 , 6×6 , and 7×7 arrays of electrodes to which the subject just acknowledged to shape of each of the figures. They inform the experimenter that they felt like a buzzing or tingling sensation on the tongue. The subjects were just asked to identify the shapes they were subjected to. As a result, a performance of 79.8 percent was found in recognition of the shapes, across all the four sizes.

Thanh Huong NGUYEN et al. [20] developed a device having stimulator array of 33 copper electrode pins of 2 mm diameter, arranged on a round grid having diameter of 4 cm and 1 mm of interelectrode spacing, as shown in Fig. 5. (a). The goal of designing this type of electrode cluster was to implement orientation navigation in 8 different directions: (a) straight forward, (b) backward, (c) left, (d) right, (e) right and forward, (f) left and forward, (g) right and backward, and (h) left and backward. This device was tested on 4 test subjects and after a short time of training of around thirty to forty-five minutes, all of them could recognize the four primary directions correctly. They even concluded that the edge of the tongue was more effective than the inner parts based on

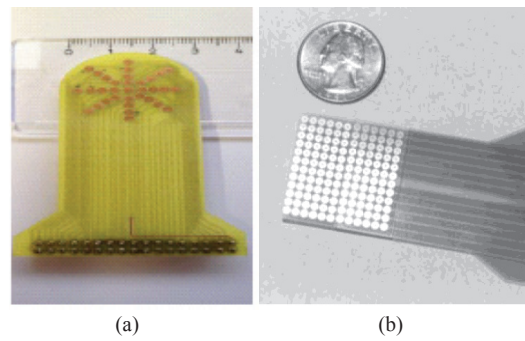


Fig. 4. Different designs of electrodes (a) Structure of round matrix [21]. (b) 144-point electrode array of TDU [22].

the feedback from the test subjects. For a detailed description on control circuits and signal generation refer [20].

Tongue Display Unit (TDU): In 2003, Dr. Paul Bach-y-Rita proposed another functional electrotactile interface having matrix size of 12×12 , that is 144-point array as shown in Fig. 5-(b). With this a device, which they named as Tongue Display Unit (TDU), was designed having flexible and thin matrix of stimulators connected via a thin cable, both having 100 μ m of thickness. They were fabricated using a polyester material – Mylar, on top which gold-plated copper circular electrodes, in the form of rectangular matrix, were deposited using photolithographic process. The center to center distance between two electrodes is 2.32 mm with each electrode being 1.55 mm in diameter and these copper-based electrodes are gold plated so as to make them biocompatible, that is to reduce electrochemical reactions at the tongue. Hence, the overall dimensions turn out to be 27×27 mm [21].

The design of TDU got its motivation from TVSS. Based on how strong or weak the tactile sensation on the tongue is, different brightness and darkness levels of the captured camera image are defined. A weak sensation corresponds to dark areas of the image. The encoded pulse includes 40- μ s pulses transmitted consecutively to each of the 144 electrodes in the pattern generated as per the image. At a rate of 50 Hz a burst of three pulses each are transmitted, having 200 Hz pulse rate within the burst. This configuration produced strong and comfortable electrotactile percepts.

So, TDU can be described as a programmable electronic device that can generate encoded pulses, which are dc-balanced, to pass through 144 channels connected to the electrode interface for stimulating the anterior and upper surface(dorsal)of tongue as shown in figure 7. It can be used

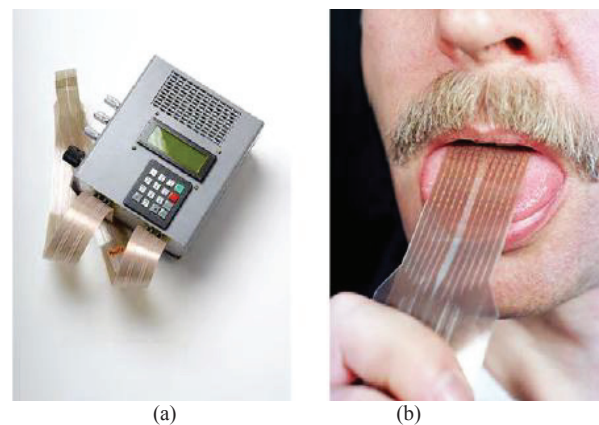


Fig. 5. (a) TDU device. (b) 144-point electrode array of TDU. [22]

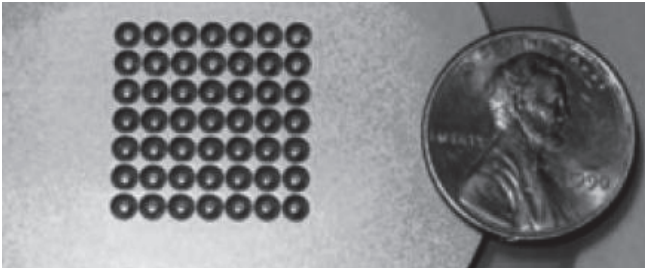


Fig. 6. A 7 x 7 stainless-steel array of electrodes shown in comparison with a U.S. penny for size comparison. [19]

as an autonomous device or it can be controlled by an external system, as per user needs. It has three modes of operations that the user can select: Remote, Standalone and Update Pattern. TDU consists of pre-programmed patterns (around 53) in its permanent memory. Depending on what mode the user selects, these patterns are selected. An in-depth information about the design overview, architecture and operation of TDU can be found in review paper by K.A. Kaczmarek [22].

TDU was tested on various blind users and as part of the learning process, it can take from 2-10 hours of user participation. During the training period, the test subjects informed that they got sensations of soda bubbles fizzing on their tongue. In the initial few minutes of usage, the user can understand the locations (up, down, left and right) of stimulation in space, hence, also the movement directions. User could then identify and get hold of the objects close by, by practicing for less than an hour. For objects far away, they could point their location and even estimate how far they are. After the users were trained further, they could even identify numbers and letters and in case the user the moving in a particular scenario, he/she can pinpoint the landmark around. A few more hours of training helped the user experience a low-resolution vision which once they used to have. Visually impaired users were even able to accomplish tasks such as reading text, tasks which include hand and eye coordination, for instance: catching a moving ball, avoiding obstacle while walking. The user could retain such development for hours to weeks and even more, even after no use.

V. APPLICATION

Tongue based electrotactile substitution system, that is TDU, has been under examination for a wide range of practical applications in neurorehabilitation and sensory substitution. More than a hundred budding applications in the domain of sensory substitution systems, brain-machine interfaces, and applications in neurorehabilitation have been identified; out of which the significant ones are listed below. For a specific application of electrotactile sensory substitution, based on tongue and basic principles of TDU, state-of-the-art BrainPort™ Vision Pro device was designed and manufactured by Wicab, Inc., located in Middleton, USA, for commercial purpose. It is a vision aid headset with tongue stimulating array of 394 electrodes connected to it also called as lollipop. TDU also inspired the researchers in producing balance substitution device which they named as BrainPort™ Balance Pro by Wicab, Inc.

VI. CONCLUSION AND FUTURE SCOPE

An explicatory research on tactile vision substitution systems (TVSS) was done. TVSS turned out to be a potential system for neurorehabilitation in case of vision related

sensory impairments in humans. It can be concluded that in order to take a step forward to this important goal, structured training programs must be created under proper supervision of investigators. The researches can also incorporate self-training option, independent of the investigator. Even though, we can see successful advancement towards our goal but it should be noted that in the real world the user which can benefit from this research cannot completely able to use these devices yet since they are not yet fully mature for full-fledged use. The TDU served as a tool to study sensory perception using tactile stimulation on the tongue. It also proved as a successful device in further exploring brain plasticity and other untouched parts of brain functions.

REFERENCES

- [1] W. H. Organisation, "World report on vision," October 2019. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment>.
- [2] J. K. A. N. O'Regan, "A sensorimotor account of vision and visual consciousness," in *Behavioral and brain sciences*, 2001.
- [3] C. M. a. C. V. A. Coppens, "Changes in occipital cortex activity in early blind humans using a sensory substitution device," *Brain research*, vol. 826, no. 1, pp. 128-134, 1999.
- [4] Wikipedia, "Tactile paving," [Online]. Available: https://en.wikipedia.org/wiki/Tactile_paving.
- [5] K. A. J. G. W. P. B.-y.-R. a. W. J. T. Kaczmarek, "Electrotactile and vibrotactile displays for sensory substitution systems," *IEEE transactions on biomedical engineering* 38, vol. 1, pp. 1-16, 1991.
- [6] C. B. E.-V. S. A.-A. H. S.-M. a. M. A. H.-C. Malamud-Kessler, "Physiology of vibration sense," *Revista Mexicana de Neurociencia*, vol. 15, no. 3, pp. 163-170, 2014.
- [7] L. H. G. a. H. E. Taylor, "The Optacon: A Valuable Device for Blind Persons," *Journal of Visual Impairment & Blindness*, vol. 68, no. 2, pp. 49-56, 1974.
- [8] Wikipedia, "Optacon," [Online]. Available: <https://en.wikipedia.org/wiki/Optacon>.
- [9] P. A. U. L. BACH - Y - RITA, "Tactile sensory substitution studies," *Annals of the New York Academy of Sciences* 1013, vol. 1, pp. 83-91, 2004.
- [10] R. a. P. D. L. Butikofer, "Electrocutaneous nerve stimulation-I: Model and experiment," *IEEE transactions on biomedical engineering* 6, pp. 526-531, 1978.
- [11] E. A. Pfeiffer, "Electrical stimulation of sensory nerves with skin electrodes for research, diagnosis, communication and behavioral conditioning: A survey," *Medical and biological engineering* 6, vol. 6, pp. 637-651, 1968.
- [12] F. A. G. F. A. Saunders, "Electrocutaneous displays," in *Conference Cutaneous Communication System Devices*, 1973.
- [13] S. Grimmes, "Pathways of ionic flow through human skin in vivo," *Acta dermato-venereologica* 64, vol. 2, pp. 93-98, 1984.
- [14] R. H. Gibson, "Electrical stimulation of pain and touch," *The Skin Senses*, pp. 223-260, 1968.
- [15] M. E. J. G. B. a. Y. P. D. Tyler, "Spatial mapping of electrotactile sensation threshold and intensity range on the human tongue: Initial results," 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 559-562, 2009.
- [16] K. P. B.-Y.-R. W. J. T. a. J. G. W. Kaczmarek, "A tactile vision-substitution system for the blind: computer-controlled partial image sequencing," *IEEE transactions on biomedical engineering* 8, pp. 602-608, 1985.
- [17] C. C. Collins, "A portable seeing aid prototype," *J. Biomed. Systems* 5, pp. 3-10, 1971.
- [18] N. N. P. O. C. A. F. Y. P. a. J. D. Vuillerme, "A wireless embedded tongue tactile biofeedback system for balance control," *Pervasive and Mobile Computing* 5, vol. 3, pp. 268-275, 2009.
- [19] P. K. A. K. M. E. T. a. J. G.-L. Bach-y-Rita, "Form perception with a 49-point electrotactile stimulus array on the tongue: a technical note," *Journal of rehabilitation research and development* 35, pp. 427-430, 1998.

- [20] T. H. T. H. N. T. L. L. T. T. H. T. N. V. a. T. P. V. Nguyen, "A wireless assistive device for visually-impaired persons using tongue electrotactile system," In 2013 International Conference on Advanced Technologies for Communications (ATC 2013), pp. 586-591, 2013.
- [21] P. M. E. T. a. K. A. K. Bach-y-Rita, "Seeing with the brain," International journal of human-computer interaction 15, vol. 2, pp. 285-295, 2003.
- [22] K. A. Kaczmarek, "The tongue display unit (TDU) for electrotactile spatiotemporal pattern presentation," Scientia Iranica 18, vol. 6, pp. 1476-1485, 2011.
- [23] "Visual System," Veterian Key, 2016. [Online]. Available: <https://veteriankey.com/visual-system/>.

Simulation and Comparative Study of Various Maximum Power Point Tracking Techniques

Abhishek Singh ,Sambhav Khatri , Sumit Kumar Gola
Delhi Technological University

Abstract –PV cell work at its full efficiency when it is operating at its maximum power point. The energy derived from the solar-cell is used to power the load. A DC-DC converter is used as an interface between the cell and the load. A boost converter is implemented to get boosted output voltage. Perturbation and Observation technique is a basic technique which is used to track the MPP. Fuzzy Logic Control (FLC), Artificial Neural Network (ANN), Particle Swarm Optimization (PSO) and Flower Pollination Algorithms (FPA) are some modern techniques which can to track the MPP more efficiently. This literature distinguishes different MPPT techniques.

Introduction –

The model of a basic PV cell is taken in this literature. The constant irradiation value of (500W/m²) and temperature of (25°C) is taken. It is vital to consider MPPT as it is needed for supplying maximum power to the load under varying environmental conditions.

Numerous analysts and industry delegates from all over the world have built up many MPPT techniques. Some main algorithms like perturb and observe (P&O) technique, fuzzy logic control, Artificial Neural Network, Particle Swarm Optimization and Flower Pollination Algorithms are implemented in this study with detailed explanation.

This paper focuses on comparing and looking at different MPPT strategies like Perturb and Observe,

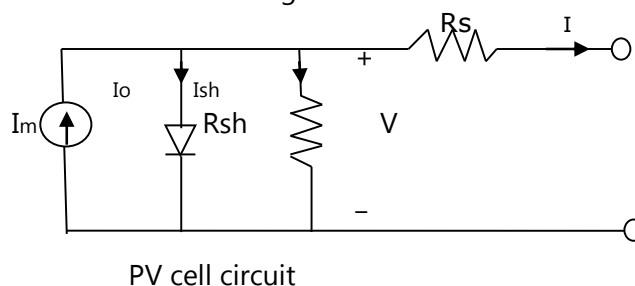
Fuzzy Logic Control (FLC) , Artificial Neural Network (ANN), Particle Swarm Optimization (PSO) and Flower Pollination Algorithms (FPA). For simulation tasks and modeling of dc-dc converter and for detailed comparison, different MPPT methods are implemented.

System Description –

The whole model can be classified into:-

- (a) Electrical force producing Solar PV framework
- (b) DC/DC support converter
- (c) MPPT procedures.

The PV cell can be represented by its basic electrical circuit shown in the figure below.



Here,

I_o - diode leakage current

I_m - module current

I_{sh} - shunt current

V - output voltage

I - output current

According to Kirchhoff's law:

$$I = I_m - I_o - I_{sh} \quad \dots \text{Equation (1)}$$

I_m is given by

$$I_m = I_{pv} - I_o N_p \{ \exp(V + R_s(N_s/N_p)/1) / (V_t N_s) \} - 1$$

...Equation (2)

N_p - Number of modules connected in parallel

N_s - Number of modules connected in series

a - Ideality factor

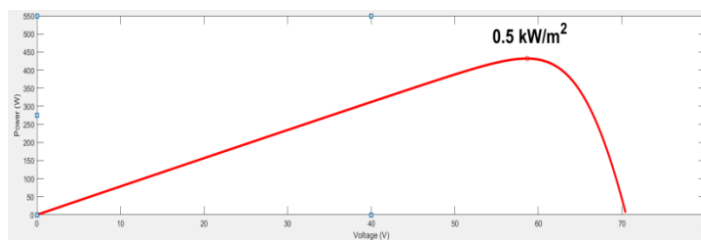
I_{pv} - PV current

I_o - Reverse leakage current

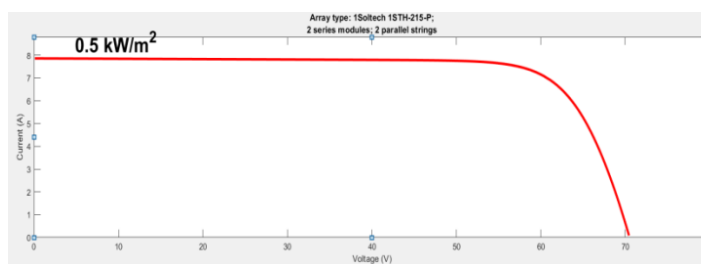
V_t - Thermal voltage

The expressions of PV cell are simulated in MATLAB SIMULINK using 1 SOLTECH-STH-215-P module. Diagram shows the I-V and P-V graphs of the pv cell under constant irradiance and constant temperature.

For irradiance of 500 Watt per meter square the ideal P_{max} is 431.9 Watt and V_{mpp} is 58 Volts.



P-V Graph



I-V Graph

The PV array is connected to the boost converter to get a high boosted voltage. The calculated ratings

DESCRIPTION	RATINGS
SWITCHING FREQUENCY	25000Hz
VOLTAGE RIPPLE	5%
CURRENT RIPPLE	3%
CAPACITOR	0.1mF

INDUCTOR	2 mH
RESISTIVE LOAD	20 Ω

MPPT CONTROL STRATEGIES –

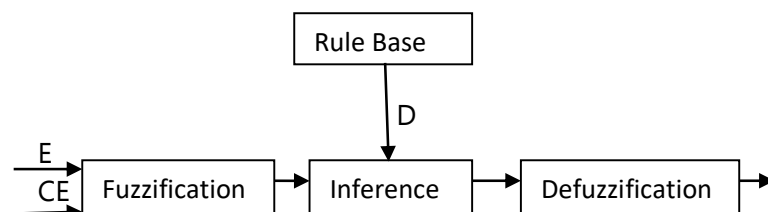
(a.) PERTURB AND OBSERVE

P&O is a simple and easy-to-design MPPT process. It only uses one voltage sensor unit to sense the voltage value of the PV array, so the cost of completion is low and the process is simple. The MPPT algorithm has a low time complexity, but it does not stop disquieting on both directions of the MPP. The current power and voltage values are compared to the previous values in this comparison and thus duty cycle is calculated. Perturbation and Observation principle is simply as follows:

- 1) If the power and voltage changes are both positive, the duty cycle should decrease.
- 2) If the difference between the power (ΔP) and voltage changes (ΔV) is negative, the duty cycle can increase.

(b.) FUZZY LOGIC CONTROL

Recently, FLC is introduced for maximum power point tracking in the PV system. The different processes of a fuzzy logic controller are as shown in the Figure below. FLC controllers are very advantageous as well as robust.



The two inputs i.e. change of error (CE) and error (E) are defined as,

$$E(K) = \frac{P_{pv}(K) - P_{pv}(K-1)}{V_{pv}(K) - V_{pv}(K-1)} \quad \dots \text{Equation (3)}$$

$$CE(K) = E(K) - E(K-1) \quad \dots \text{Equation (4)}$$

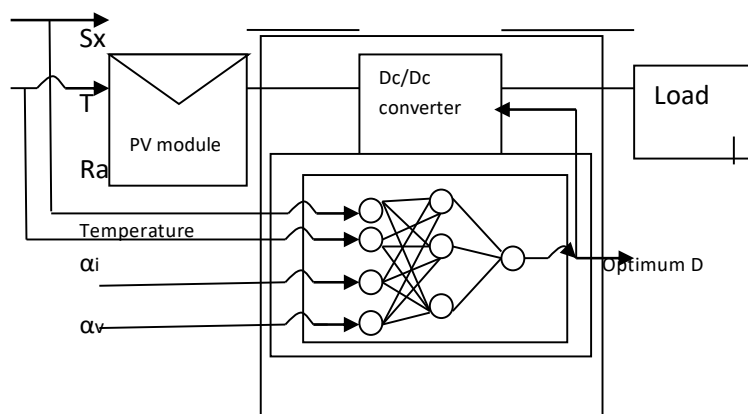
The Instantaneous power of PV array is P_{pv} . Fuzzy inference is processed using Mamdani's method. Defuzzification is done to find out the duty cycle.

Dutycycle output is processed by using center of gravity method. The fuzzy rule base matrix which is used in this literature is given as follows:

(E , CE) NB		NS	Z	PS	PB
NB	PB	PS	NS	NS	NB
NS	PS	PS	NS	PB	NB
Z	NB	NB	NS	PS	PB
PS	NS	NS	PB	NB	PS
PB	NS	NS	PB	PB	PB

(c.) ARTIFICIAL NEURAL NETWORK

The ANN is trained using NNTOOL BOX in Matlab. Data set of Temperature and irradiation varying between 25-50°C and 200-1200W/m² respectively is given as input to neural network and their corresponding V_{mpp} values are obtained from P-V array plots.

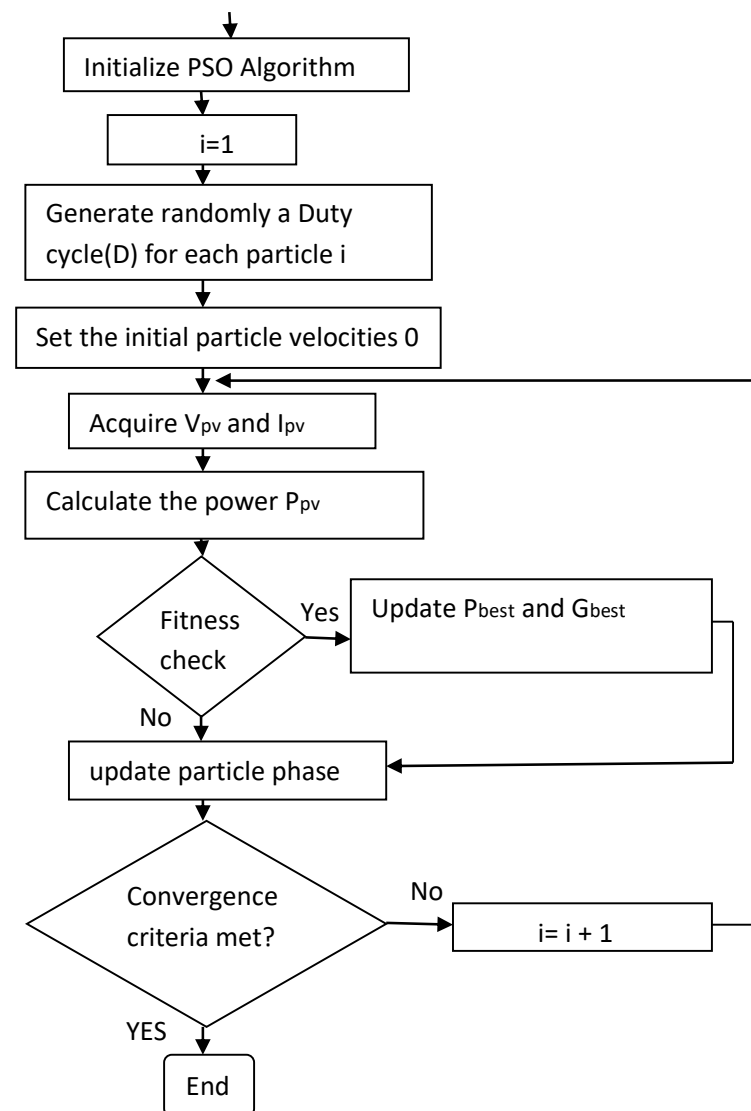


(d.) PARTICLE SWARM OPTIMISATION

The Particle Swarm Optimization explores the Search-Space, and can be utilized to decide the

segments **BEGIN** gs needed to streamline a particular Objective Function. The operation starts with a random selection, proceeds with a search for ideal solutions through prior iterations, and assesses the solution quality through the wellness/fitness. The PSO controller is appropriate for the deduction of the worldwide ideal. It is a basic algorithm, and has a high following precision.

The principle of the PSO algorithm is shown in the following flowchart.



(e.) FLOWER POLLINATION ALGORITHM

All of the initial pollens must be spread throughout the entire duty cycle spectrum, and the entire power-duty cycle curve (P-D curve) must be searched. For the global quest, the best pollen number m selection is critical. The probability of finding the global optimal solution increases as the number m is increased, yet the convergence time increases.

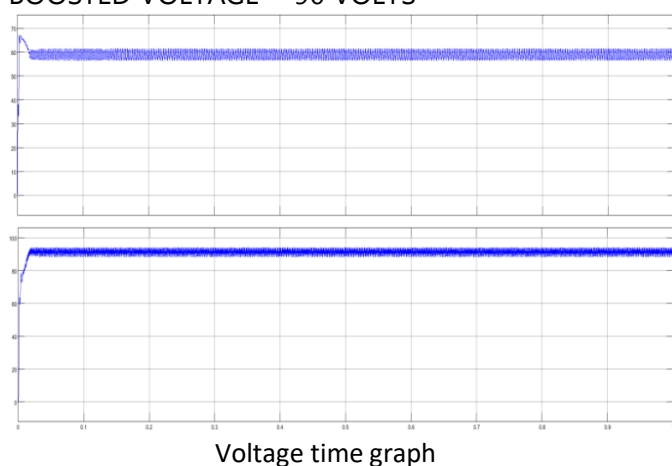
The flower pollination algorithm has been shown to be efficient in finding global optimal solutions in a short amount of time. It's been commonly used to solve nonlinear optimization problems in recent years. In comparison to other approaches, the FP algorithm is easy to modify and has less parameters. The conversion probability parameter can be used to enforce dynamic conversion between global and local search, and thus the balance between global and local search is well solved. It also outperforms PSO in terms of convergence speed.

SIMULATION RESULTS AND COMPARISONS –

(a.)P&O ALGORITHM :

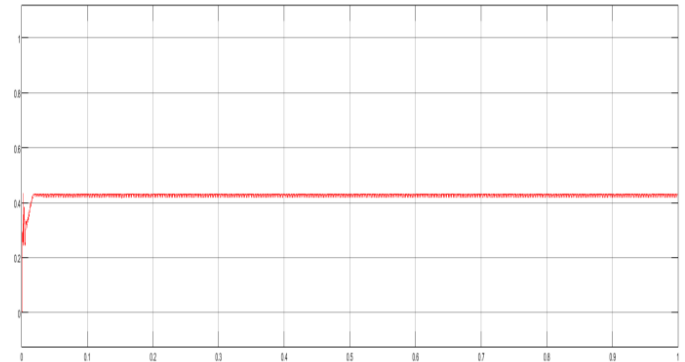
INPUT VOLTAGE =60 VOLTS

BOOSTED VOLTAGE = 90 VOLTS



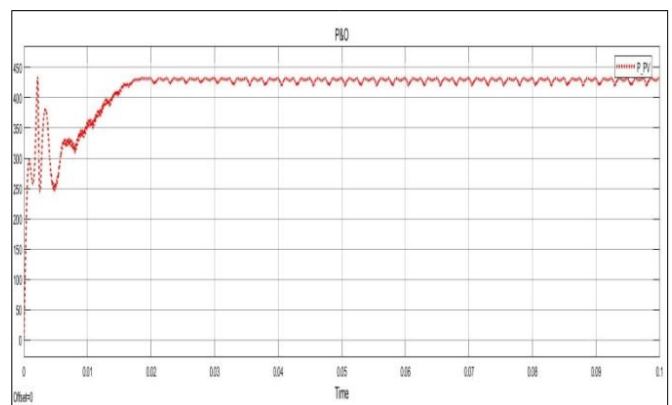
Voltage time graph

Maximum Power =424.78 Watt(Run for 1 Second)



Power time graph

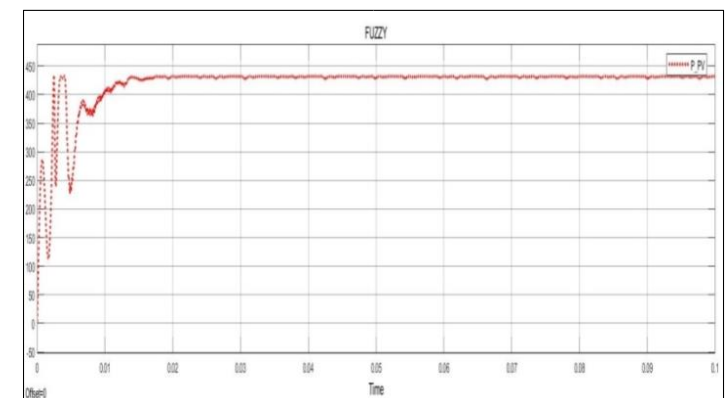
TRANSIENT PERIOD of 0.1 second (Slow Time response)

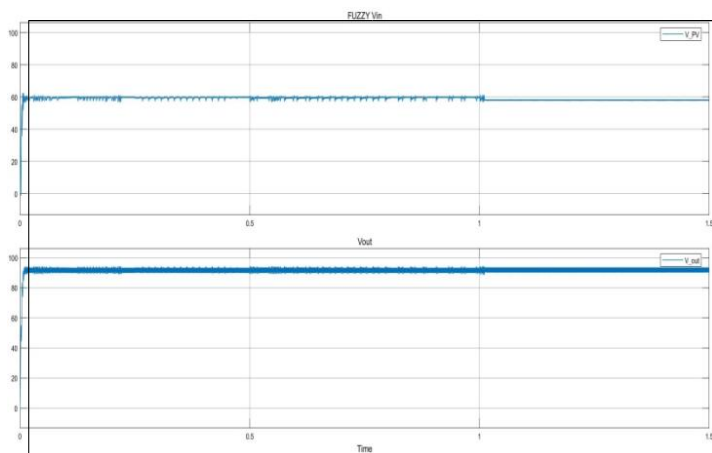


(b.)FUZZY LOGIC CONTROL

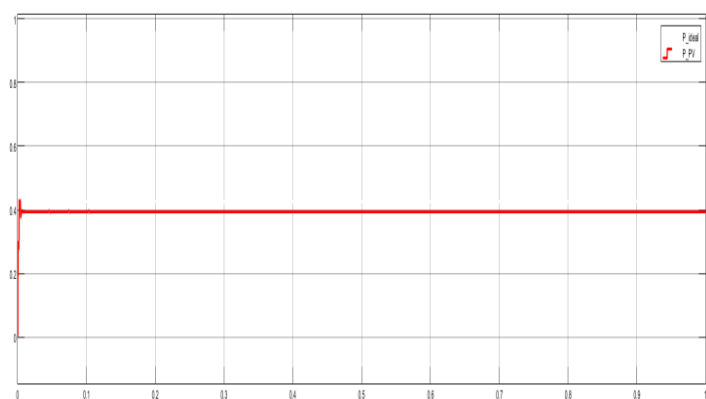
INPUT VOLTAGE =60 VOLTS

BOOSTED VOLTAGE = 90 VOLTS



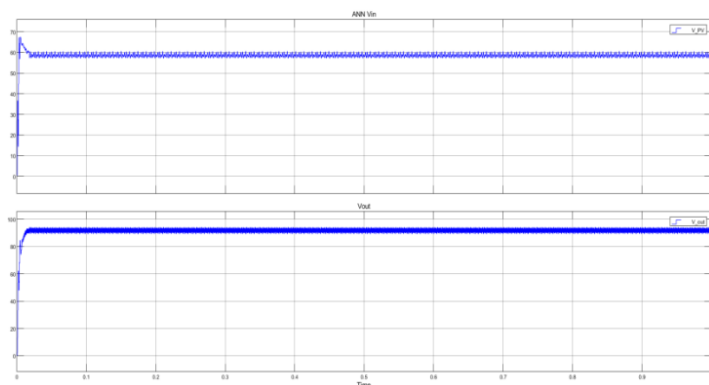


Maximum Power = 428.6 Watt (Run for 1 Second)



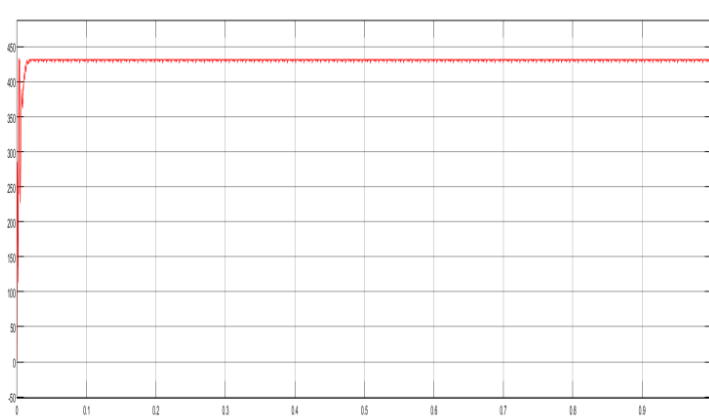
TRANSIENT PERIOD of 0.1 second (Moderate Time response)

(c.) ANN
INPUT VOLTAGE = 60 VOLTS
BOOSTED VOLTAGE = 90 VOLTS



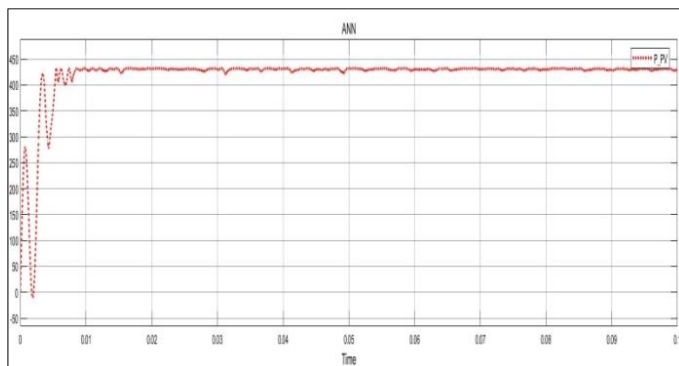
Voltage time graph

Maximum Power = 430.8Watt (Run for 1 Second)

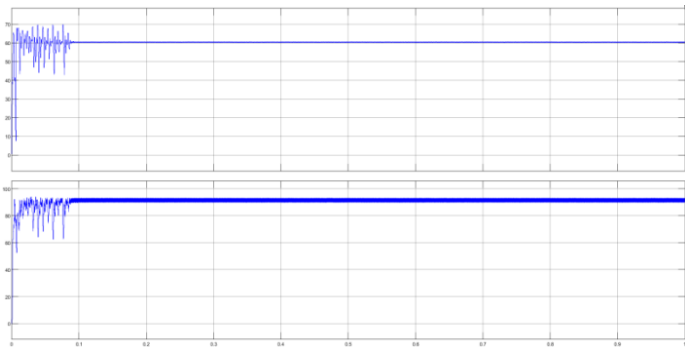


Power time graph

TRANSIENT PERIOD of 0.1 second (Fast Time response)

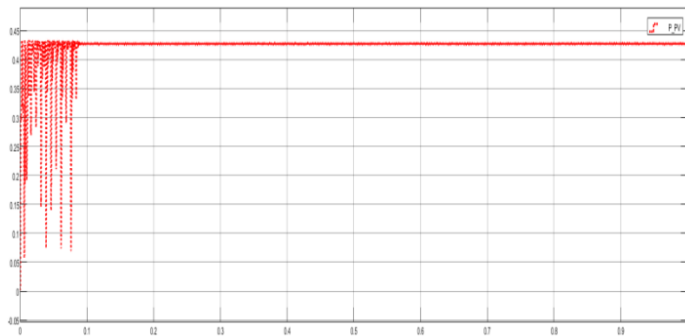


(d.) PSO ALGORITHM
INPUT VOLTAGE = 60 VOLTS
BOOSTED VOLTAGE = 90 VOLTS



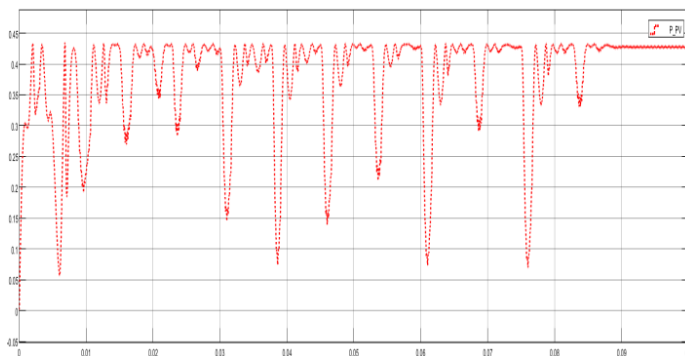
Voltage time graph

Maximum Power = 426.69 Watt (Run for 1 Second)

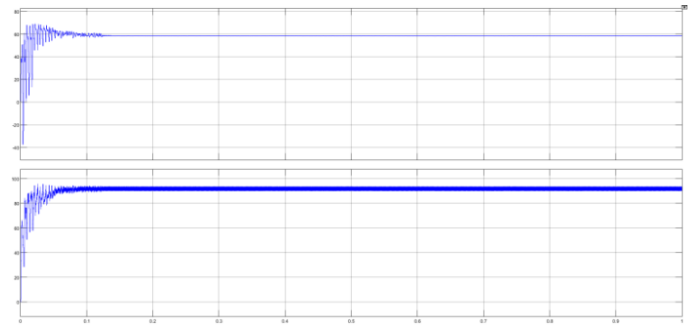


Power time graph

TRANSIENT PERIOD of 0.1 second (Slowest Time response)

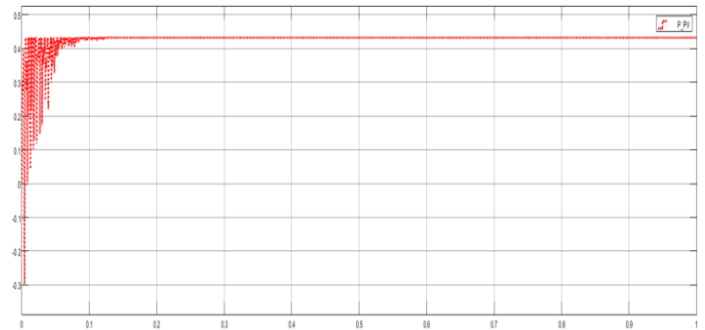


(e.) FLOWER POLLINATION ALGORITHM
INPUT VOLTAGE = 60 VOLTS
BOOSTED VOLTAGE = 90 VOLTS



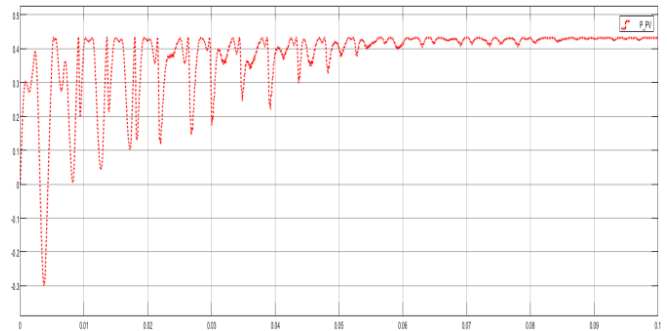
Voltage time graph

Maximum Power = 427.23 Watt (Run for 1 Second)



Power time graph

TRANSIENT PERIOD of 0.1 second (Slow Time response)



	TRANSIENT RESPONSE	COMPLEXITY	POWER CONSUMPTION
P&O	SLOW	SIMPLE	LESS EFFICIENT
FLC	FAST	COMPLEX	HIGH EFFICIENT
ANN	VERY FAST	VERY COMPLEX	VERY HIGH EFFICIENT
PSO	SLOWEST	COMPLEX METAHEURISTIC TECHNIQUE	EFFICIENT
FPA	SLOW	COMPLEX METAHEURISTIC TECHNIQUE	EFFICIENT

CONCLUSION

This paper shows the detailed comparison of different maximum power point tracking controllers. It can be seen from the graph that the time response in reaching maximum power point is fastest in the ANN controller and time response in the FLC controller is moderate while P&O controller has slow time response. PSO controller has the slowest time response. All the controllers are successfully tracking ideal maximum power i.e. 431.9Watt.

ANN is the most efficient technique achieving power point followed by fuzzy controller and then P&O algorithm. P&O is a simple technique while Fuzzy and ANN are two very complex modern techniques. PSO and FPA are two complex meta-heuristic techniques. FPA is better than PSO in convergence speed.

Hence ANN has most power consumption and is the most efficient algorithm.

REFERENCES

1. Mohamed A. Eltawil, Z. Z. (2013). MPPT techniques for photovoltaic applications. Renewable and Sustainable Energy Reviews.
2. M. Nkambule, A. Hasan and A. Ali, "Proportional study of Perturb & Observe and FLC MPPT Algorithm for photovoltaic system under changing weather conditions," 2019 IEEE 10th GCC Conference & Exhibition (GCC), Kuwait, Kuwait, 2019, pp. 1-6
3. MdSamiulHaqueSunny, AbuNaimRakib Ahmed and MdKamrulHasan, "Design and simulation of maximum power point tracking of photovoltaic system using ANN", 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2016
4. S Narendiran, S. K. (2016). Fuzzy logic controller based maximum power point tracking for PV system. 3rd International Conference on Electrical Energy Systems (ICEES). Chennai, India: IEEE.
5. K. Ishaque, Z. Salam, M. Amjad, and S. Mekhilef, "An improved particle swarm optimization (PSO) based MPPT for PV with reduced steady state oscillation," IEEE Trans. Power Electron. 27, 3627–3638 (2012).
6. Yang, X., & Karamanoglu, M. (2014). Multi-objective flower algorithm for optimization. ProcediaComputSci, 18, 861–868.
7. V. Phimmason, Y. Kondo, T. Kamejima, and M. Miyatake, "Evaluation of extracted energy from PV with PSO-based MPPT against various types of solar irradiation changes," in 2010 International Conference on Electrical Machines and Systems (ICEMS), 2010, pp. 487–492.

Solving Community Detection in Social Networks: A comprehensive study

Prashant Kumar
Computer Science Department
Delhi Technological University
Delhi, India
prash.kumar047@gmail.com

Raghav Jain
Computer Science Department
Delhi Technological University
Delhi, India
raghavjain106@gmail.com

Shivam Chaudhary
Computer Science Department
Delhi Technological University
Delhi, India
shiva.chaudhary2001@gmail.com

Sanjay Kumar
Computer Science Department
Delhi Technological University
Delhi, India
sanjay.kumar@dtu.ac.in

Abstract—In today's World, social media platforms such as facebook, instagram, linkedin connects various users forming a social network graph. In these social media graphs, detecting communities is a very essential task as communities helps us in grouping users showing similar behaviour and in this way the social network can be divided into different clusters of nodes with same behaviour. This community information can help us take useful decisions and extract important information about users in a particular community. In last few years scientists and researchers from different fields are trying to solve this problem using various methods and techniques. The proposed research work has conducted an extensive and exhaustive survey on different methods applied to the problem of community detection. This paper summarizes and compares all those techniques by classifying them into four broad domains - matrix factorization, random walk, deep learning and spectral methods. With so much of work going in the field in last few years, this survey paper helps researchers in getting started in the field of social networks and detecting communities within these networks.

Index Terms—Community Detection, Social Network Analysis, Deep Learning, Matrix Factorization, Random Walk, Spectral Clustering, Label Propagation

I. INTRODUCTION

A network can be defined in a number of ways by computer science researchers, mathematicians, statisticians or physicists. But to visualize any network, the definition from graph theory will be used. According to which there are some users and any interaction between them leads to an edge creation. The interaction between different users/nodes is a result of characteristics such as friendship, relations between family, business relations etc. A network is not the same as a graph because a network can contain much more information or data about users or their interaction than a graph. A network can also be dynamic in nature which means continuously changing the nature of graph structure by addition or deletion of nodes or edges.

Communities are found everywhere from simple graph dataset to real world human interactions. Detecting these communities is an important problem as it helps us to gather

a lot of information about a cluster. A community can be visualized as a cluster/collection of vertices/users with more interaction between them than other members of the graph structure as shown in Fig1. Users within the same community will exhibit similar behavior and functions. In real life it can be compared with a group of friends or people having the same background or interests. Detecting communities in a graph can also be considered as a graph partitioning problem which comes under the category of NP-hard problems.

Community detection has a variety of applications in a number of different fields. Some of these applications are detecting communities in networks of criminal organizations, analysing and study of groups which are susceptible to an epidemic disease, also used by companies for dividing market into smaller groups and clusters for advertisement targeting, suggesting products or friends on social media platforms and recommendation system, link prediction and influence maximization. This survey paper summarizes and compares the

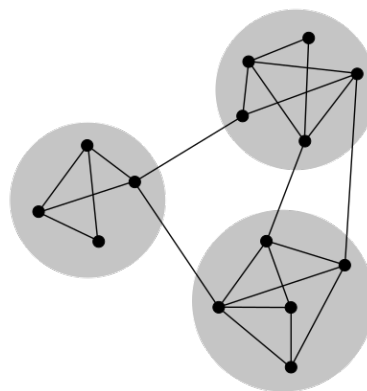


Fig. 1. Community structure [24]

recent work of researchers and scientists working in the field of community detection in graphs in the last 5-6 years. This survey paper consists of 6 sections which are divided as follow.

Section 2 defines the community detection problem. Section 3 consists of work done in recent years in the area of community detection. Section 4 lists the information about famous datasets that are publicly available. Section 5 summarizes the various methods categorized by us. Section 6 discusses the future scope and challenges in this domain and concludes this survey paper.

This research work has conducted a thorough survey of papers on community detection that were published on top international conferences in the area of artificial intelligence, deep learning and data mining such as NIPS, AAAI, ICLR, KDD etc. Also, this survey includes articles, which were published in top reputed journals. This paper will serve as a platform for researchers and scientists to follow trends in the field of social network analysis and community detection.

II. PROBLEM FORMULATION

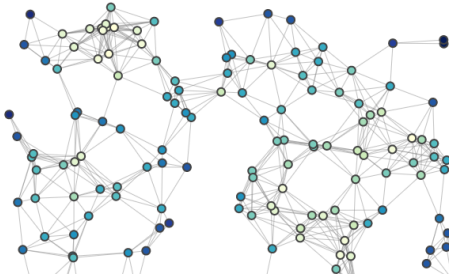


Fig. 2. Graph Structure [27]

In this Paper, a network is considered as a graph structure that is used to represent complex and real life interactions and relationships as shown in Fig2. According to graph theory, a weighted graph is defined as $G = (V, E, W)$ where V and E are the set of vertices and edges. W represents the weights between those edges. On a similar note, an unweighted graph is defined as $G = (V, E)$. Here weight is considered as 1 and thus it can be ignored in the definition. A community is defined as a cluster or subgraph within the graph that has more connections within that cluster than the vertices outside that cluster. A subgraph C_i is considered to be a community of graph G if $deg^-(v_i) > deg^+(v_i)$ where $deg^-(v_i)$ refers to indegree of node v_i and $deg^+(v_i)$ refers to outdegree of node v_i and $v_i \subset C_i$. The aim of community detection algorithms is to detect communities C in graph G where $C = \{C_1, C_2, \dots, C_k\}$ refers to the set of k different communities within graph G .

III. RELATED WORK

This section covers the recent progress by researchers and scientists to solve the problem of community detection. After reading research papers and articles of last few years from top conferences and journals, the approaches to solve community detection can be divided into these subcategories - matrix factorization, random walk, deep learning and spectral methods as shown in Fig3. We will go in more detail about

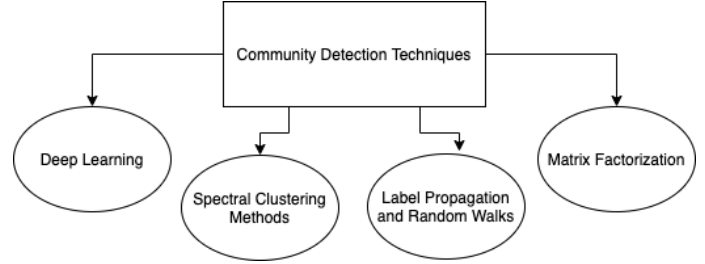


Fig. 3. Community Detection Techniques

work done in each domain in each subsection summarizing various methods and algorithms used for community detection.

A. Deep Learning

The research in the area of deep learning is growing exponentially for the last few years. There is no denying the fact that deep learning has been used to solve various problems in computer science and other fields such as protein folding, capturing black hole images, drug discovery and sports analytics. Neural networks are very powerful tool which can be used to approximate any mathematical function according to universal approximation theorem and has also shown strong representation power. The main contribution of deep learning in solving community detection has been discussed below.

Autoencoders are a very popular and powerful neural network architecture which has very strong representation powers. It consists of two neural networks. One neural network is encoder which encodes our input into lower dimension representations. Other neural networks called decoder tries to reconstruct original data from this lower level representations. It is trained by minimizing the error which is a function of original input and reconstructed output. It is unsupervised technique. The encoder and decoder can be any neural network like simple Artificial neural network, convolution neural network or LSTM. This encoding of input into lower representations is the main reason why it has been used in detecting communities in a graph. Yang et al. [1], proposes a deep learning architecture by stacking autoencoders in series. They feed the input modularity matrix into first autoencoder and try to obtain the lower dimensional representation from this encoder by minimizing the reconstructive error loss and feed it into next auto encoder. After a series of similar steps, they finally applied the k-means clustering technique to the encoded output from last autoencoder to detect communities. Dhillon et al. [2] proposes a similar architecture but rather than connecting autoencoders in series and training separately each autoencoder, they stacked them parallelly and trained all autoencoders simultaneously.

Generative adversarial networks are another powerful deep learning architecture which consists of two neural networks. One is called discriminator and the other one is known as generator. GANs can be considered as a minmax game where generator tries to generate data from real data set and discriminator tries to distinguish between real data and data

generated by generator. In this way both tries to minimize their loss by competing with each other trying to fool the other model. GANs and adversarial machine learning has shown successful results in the task of graph embedding and representations. Yuting Jia et al. [3] proposes a new architecture CommunityGAN that solves one of the problem arising in traditional methods that is not able to detect overlapping communities. Most of the previous techniques believed that one node belongs to only one community that is why they are not able to work on overlapping communities. But CommunityGAN is able to solve this problem. Other than this, the embedding generated by CommunityGAN is able to represent the relation between nodes and communities showing their membership power.

Deep network embedding techniques refers to mapping higher dimensional data which in case is graph into a lower level representations in such a way that the data preserves its original structure and information. After obtaining the lower level representation techniques any clustering technique can be applied to obtain communities. Sanjay Kumar et al. [4] used SDNE embedding framework for obtaining the low level representation of input graphs and after that used K Means algorithm optimised with gravitational search algorithm to obtain communities. Sandro Cavallari et al. [5] introduced a new framework where they learn community embedding instead of individual node embedding. They form a cyclic structure of node embedding, community detection and community embedding in which node embedding gives communities, and better communities will help generate better community embedding, and better community embedding will optimize node embedding.

Graph Neural Networks and Graph Convolutional Networks are a new class of neural networks that works on graphs and Non-Euclidean domains. Idea of GCNs is based on convolution neural networks. CNNs usually operate on images by capturing the surrounding information of a pixel of image. On a similar note, here convolution framework of GCNs tries to capture the surrounding information of a node or edge. There has been rapid research development in domain of graph neural networks. Zhengdao Chen et al. [6] propose a new graph neural network called Line graph neural network that solves the community detection problem in a supervised manner and it uses a non backtracking operator which is defined on the edge adjacency list.

B. Label Propagation and Random Walks

Label propagation is a semi-supervised algorithm Zhu and Ghahramani [7] that initially assigns labels to a small subset of the data and then as the algorithm proceeds the labels are propagated to all the unlabeled data points in the space. Random walks is a stochastic process in which an object randomly moves through a mathematical space or structures like a network of connected nodes which can then be used to gain information on the hidden structures (like communities) in the given space.

Krylov Subspace Approximation HE et al. [8] is a technique in which local community is detected by finding a linear sparse coding on the Krylov subspace i.e. the local approximation of spectral subspace. There is a local sampling which uses a seed node to find a comparatively small subgraph G_s in a given graph G . Based on different random short walk diffusion and local community detection by finding sparse relaxed indicator vector lying in local spectral subspace we find the subordinative probability of the corresponding nodes. To get the “Local Spectral subspace” random short walks for probability diffusion from seed set is used instead of eigenvalue decomposition

Multiple Community Detection Hollocoou et al. [9] technique uses seed nodes to extract communities from the given graph structure. For each seed node a score is calculated to obtain a local community around the seed nodes. The scores are used to define an embedding which maps all the nodes to a vector of appropriate dimension which can be fed to a clustering algorithm DBSCAN M. Ester et al. [10] to obtain K clusters of nodes. Clusters with lower threshold than a specified value are considered the direction to be moving forward, the algorithm moves forward by picking a new seed in each of them. Repeating the steps over multiple iterations till no new seeds are formed.

Targeting influential users in a network and then propagating flow of information originating from those nodes with certain probability can help use identify closely grouped nodes. Community detection by simulating information flow Venkatesaramani and Vorobeychik [11] uses this technique to identify community structures in a given graph network. In this technique “Alpha detection” i.e. identifying users that are likely to be the source of information in a network. After these alpha nodes are identified they are treated as the source from where information will be propagating through the network. We end up with X communities where X is the number of alpha nodes.

Random walk is a stochastic process describing a path that consists of random successive steps on some mathematical space. This method can also be used on graph data structure. The basic idea is that we have a walker that randomly explores a network, so a node having high visiting probability is considered to be near the central node of a community and can be considered to be part of that cluster. A method utilizing this approach, Multi-Walker Chain Bian et al. [12] is proposed where a group of K walkers is used. These walkers explore the network one by one for multiple iterations updating the visiting probability of each node. To identify the communities the top L nodes the the largest mean scores is selected and the conductance value of the subgraph introduced by the nodes is calculated. Node set with the smallest value of conductance is returned as a community.

Diffusion methods can be used to detect community structures in a network. In this type of approach a conceptual dye is injected on a particular node of multiple nodes in a network and watch the spread of the dye diffused over time steps across the edges of the network. The manner in which

the dye diffuses provides us with hidden information on the structure of a graph. A Nonlinear Diffusion Ibrahim and Gleich [13] method for community detection is proposed where a semi-supervised technique is used, meaning that the diffusion adjusts by using feedback from the results. This method is repeated for a specified number of iterations or until there are no significant changes in the diffusion method.

C. Matrix Factorization

Matrix factorization (decomposition) is a collaborative filtering algorithm that works by decomposing the input matrix into two lower dimensional matrices making it easier to infer and calculate information from them. For example if we take number 10 then it can be factored into two parts, 2 and 5. Two most widely used methods are LU matrix decomposition and QR matrix matrix decomposition.

Embedding is a process of representing a high dimensional non-linear structure like a graph into multiple 1d vectors each of same dimension preserving as much information possible from the original structure. The algorithm proposed in Skrlj et al. [14] traverses the embedding space and for each embedding tries multiple values of k (number of clusters). The method SCD uses two-step approach in finding the optimal value of k and further use it to identify communities in the network. It does that based on a Silhouette score - $SilhouetteGlobal(p, k)$ where p is the parameters passed to the embedding technique and k is the number of cluster to obtain in the network.

Hierarchical knowledge graphs can provide us with relevant real-world information about the clustering structure in a network, they are not always explicit in a network but can be useful in finding communities. We can decompose such HKGs to provide contextual information. This proposed approach Bhatt et al. [15] uses this knowledge graph to enhance the detection of communities by using the graph and the HKGs as input and finds community labels as well as the context from the HKG.

Normally graph embedding techniques and community detection are done separately. The proposed vGraph method Sun et al. [16] tries to solve both the problems by learning the embedding and detecting communities simultaneously by introducing a concept of node and community embedding also assuming that every node can be a part of multiple communities. Using this approach the representation of node can take advantage from the information gained by detection of node communities and vice versa.

Non negative matrix factorisation (NMF) based methods factorize the adjacency matrix of a graph and converts it into two non-negative factor matrices. Now each column in the factor matrix can be analyzed as an inclination of a node to belong in different communities and the other can be used to identify mappings between the original network and the community membership. This proposed method Ye et al. [17] uses Deep Autoencoder like NMF for community detection which uses auto encoders to learn the mappings between the factor matrices. The components of the autoencoder guide each other in the learning phase obtaining an ideal community

membership of nodes. This proposed method Adaptive Affinity Learning for Accurate Community Detection, Ye et al. [18] is another that uses Non negative matrix factorisation (NMF). Using a technique to adaptively learn an affinity matrix and capture the essential equivalence between the nodes leading to an improved community detection. This method embeds each node into a low-dimensional vector using transformation matrix, preserving the community structure. Using a mutual guiding system makes the model more accurate in detecting the similarity between the nodes.

D. Spectral Clustering

Spectral clustering is one of the oldest methods to detect clusters in graph dataset or real world non graph dataset. This clustering technique is inspired from graph theory. A graph can be represented in many different forms such as degree matrix, adjacency matrix or graph laplacian matrix. Spectral clustering clusters nodes on the basis of information gathered from eigenvalue of these matrices. The main contributions of spectral clustering in detecting communities are discussed below.

Fang Hu et al. [19] proposed a novel algorithm called node2vec-SC which consists of two phases. Node2vec is used to learn node embeddings of each node in the graph. Then spectral clustering technique is used to detect communities after calculating and finding eigenvalues and eigenvectors of similarity matrices, degree matrices and normalized laplacian matrices. This algorithm is also equally feasible for real world datasets.

Xiang Li et al. [20] discusses how spectral clustering can work on Heterogeneous Information Networks. Heterogeneous Information Networks (HINs) are the networks which are used to model real world interactions where every edge or link refers to a different type of interactions and relations between different types of vertices. Meta paths are a method of representing and modelling relations between different nodes in knowledge graphs or HINs. They propose a spectral clustering method where they form clusters by forming a similarity matrix with the help of meta paths rather than random walks.

Spectral clustering being one of the oldest techniques to detect communities has one disadvantage that it is not scalable to large graph datasets and real world datasets because of its high computational complexity due to the formation of similarity matrix, graph laplacian and calculating eigenvalues and eigenvectors of this similarity matrix. Lingfei Wu et al. [21] uses random binning features to speed up the process of formation of similarity matrix and calculation of eigenvalues and eigenvectors as they help in faster convergence. They also used single value decomposition(SVD) factorization method to calculate the eigenvalues faster.

Detecting overlapping communities is one the major challenges faced by researchers in the field of social network analysis. Yuan Zhang et al. [22] proposes an extended version of spectral clustering to solve the problem of overlapping communities. They have used K median technique for clustering

eigenvalues of similarity matrix rather than using K means technique. This version of spectral clustering works well till communities are not largely overlapped.

Yixuan Li et al. [23] proposes a new spectral clustering method to detect communities in networks of large sizes. They propose two major changes in traditional spectral clustering. One of them is to initialize a random walk from seed nodes to detect nodes that may be present in target communities to reduce the number of computations of eigenvalues and eigenvectors. Other is to replace k-means algorithm in spectral space and find sparse vectors in to detect communities.

All these reviewed papers are mentioned in Table1 with all the datasets and techniques that were used. This table also lists all the different evaluation metrics that were being used by different researchers to assess their techniques such as NMI score, F1 score both of which are most popular evaluation metrics.

IV. DATASET

Stanford Network Analysis Project (SNAP) provides a collection of large network datasets with more than 50 networks. Each having millions of nodes and edges. They include networks spanning a wide range of real world applications like social network, road network, citation networks, web graphs, communication networks etc. The majority of papers that we surveyed used networks from this dataset only like Amazon networks, where the nodes are representative of the products and the edges are linked with commonly co-purchased products. Other networks include ego-Facebook, ego-Gplus, ego-Twitter, DBLP etc.

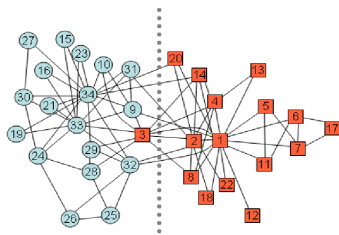


Fig. 4. Karate Club Network [25]

Apart from SNAP datasets, other datasets that are publicly available and frequently used are Dolphins dataset which consists of a graph of dolphin interactions, football dataset which consists of a network of different football games visualized as a graph, Les miserables dataset which consists of a graph made from characters based on novel Les miserables, Zachary Karate dataset which contains a graph made from relationship of around 34 karate students of a particular karate club shown in Fig3, Polblogs which is network of different left and right political bloggers and Polbooks which is a network of United states political books.

Lancichinetti Fortunato Radicchi(LFR) generated networks were also used throughout the papers we surveyed. a sample of LFR network is shown in Fig5. This algorithm generates artificial graph networks that are closely similar to the real world

TABLE I
THE TABLE CONTAINS INFORMATION SUCH AS DATASET USED, TECHNIQUES USED AND EVALUATION METRIC USED IN DIFFERENT PAPERS WE SURVEYED

Publication	Dataset used	Technique Used	Evaluation Metric
Fang Hu et al. [19]	Karate, Dolphin, Football, Les Miserables networks	Node2Vec embedding with spectral clustering	NMI,AMI,FMI, ARI scores
Xiang Li et al. [20]	DBLP, Yelp, Freebase	Spectral clustering with meta paths	NMI, Purity, Rand index
Lingfei Wu et al. [21]	Pendigits, letter, mnist, ijcnn1, acoustic, ijcnn1, codrna, covtypemult, poker	Spectral clustering with SVD and random binning features	NMI score, Rand index, F-measure, accuracy
Yuan Zhang et al. [22]	SNAP ego-networks	Spectral clustering optimized with K medians	Extended normalized variation of information
Yixuan Li et al. [23]	Amazon, youtube, dblp,orkut	Spectral clustering and random walk to find seed nodes	F1 score
HE et al. [8]	SNAP, DBLP, Amazon, Youtube, Orkut	Linear Sparse coding on krylov subspace	F1 score
Hollocou et al. [9]	SNAP	Seed nodes to extract communities from the given graph structure	F1 Score
Venkatesaramani and Vorobeychik [11]	Karate Club network, SNAP	“Alpha detection” to identify nodes likely to be the source of information in a network	Conductance
Bian et al. [12]	SNAP - Amazon, Youtube, Orkut, LiveJournal	K multiple random walker chain	F1 Score and consistency
Ibrahim and Gleich [13]	SNAP, LFR synthetic graph	Semi-supervised nonlinear diffusion	F1 Score, Conductance
Skrlić et al. [14]	E-mail network	Embedding space traversal based on Silhouette score	NMI score

Ye et al. [17]	Email, Wiki, Cora, Citeseer, Pubmed	Non negative matrix factorisation (NMF), Deep Autoencoder like NMF	NMI, Adjusted Rand (ARI) Index
Ye et al. [18]	Polbooks, Football, PoliticsIE, Polblogs, Olympics, PoliticsUK, EmailEU	NFM, Adaptive learning of Affinity matrix	NMI scores, ACC
Yang et al. [1]	Karate, Dolphins, polblogs, polbooks, football, friendship6, friendship7, cora	Stacked autoencoders	NMI score
Yuting Jia et al. [3]	LiveJournal, youtube, Orkut, dblp, amazon	Generative adversarial networks	F1 score
Sanjay Kumar et al. [4]	Karate, football, polblogs, polbooks, word networks	SDNE embedding technique with k means and gravitational search algorithm	NMI score
Sandro Cavallari et al. [5]	Blogcatalog, DBLP, wikipedia, karate network, flickr	Deep community embeddings	Conductance, NMI, micro F1, macro F1 scores
Zhengdao Chen et al. [6]	Amazon, youtube, dblp	LINE GNNs	Overlap score
Bhatt et al. [15]	G+ ego network, Twitter, DBLP, Reddit	Hierarchical knowledge graphs(HKG) to enhance the detection of communities	F-Measure, Jaccard measure
Sun et al. [16]	Citeseer, Cora, Cornell, Texas, Facebook, Youtube, Amazon, Dblp, Coauthor-CS	Simultaneous embedding and community detection with shared information	F1 Score and Jaccard similarity

networks. It has an advantage of accounting for heterogeneity in node degree distribution and community sizes.

V. DISCUSSION

Spectral clustering techniques are also used by various researchers to detect clusters or communities in graphs and networks. The advantage it has over other clustering techniques like k means is that spectral clustering does not presume any information about clusters or communities such k-means where it assumes that all the data points will be clustered around some centroids. The disadvantage it has shown is that it is very computationally expensive for large real world datasets because before clustering it has to make similarity matrix and calculate eigenvectors. Various researchers have used different

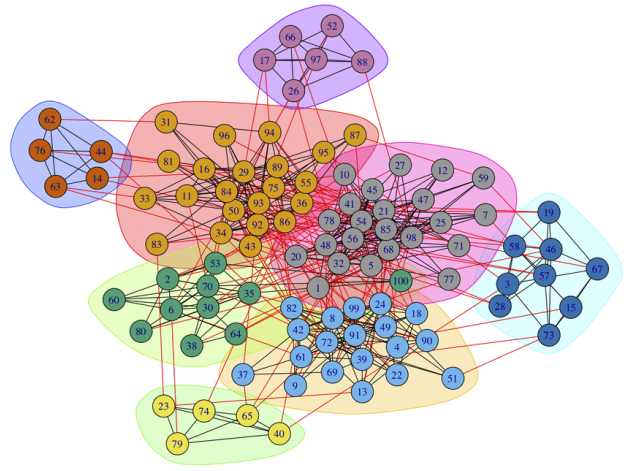


Fig. 5. LFR Network [26]

techniques to speed up this process by using techniques such as SVD and random binning features.

Theoretically Neural networks can approximate any mathematical functions and possess great representation powers making it a very strong tool to work on graphs and networks. Neural networks architectures such as autoencoders, deep embedding techniques generates low level non linear embeddings which will be able to better represent the data in lower vector space and also able to preserve the non linear features of the data.

Matrix factorization techniques are able to decompose the matrix into smaller matrices which can improve the ability to detect the hidden features and ways the nodes are connected. One of the decomposed matrices can be used to identify one feature in the network and the others for a different feature and then use them in parallel to get better/accurate community detection results. These decomposed matrices are also able to preserve as much information as possible so we don't have to worry about significant information loss as the result of decomposition.

Label propagation and random walk techniques have been proved to be very effective in detecting communities as they are able to go inside structure and propagate information at the node level. Label propagation is able to diffuse information from one node to the neighbouring nodes and by analyzing the pattern of the diffusion of the labels we can gain important insight on the existing community structure in the network. Similarly random walk algorithms walk on the network space from node to node learning information about the neighbours and thus aiding in the community detection process.

VI. CONCLUSION AND FUTURE WORK

Today we live in a world where being connected is the norm, and forming communities is a natural flow that happens when people stay connected. These communities can vary in sizes and forms and contain intricate information on how these systems interact with each other and the factors that bring them

to from different communities. Analyzing this structure helps us understand the surrounding environment and demystifies the relationship amongst different nodes, giving us insight on the social phenomena.

In our paper, we conducted a thorough survey of the latest papers in the field of community detection and analysed state of the art techniques and methods applied by different researchers and scientists to solve community detection. We also categorized these papers into four subcategories on the basis of algorithms used to solve the problem of community detection. Apart from algorithms and methods, we also analyze the popular and publicly available datasets and listed them in our paper.

There has been a rise in the research of social network analysis in the last few years and it is only going to increase in the future. There has been lots of opportunities and challenges in the problem of community detection. In last few years a lot of research has been conducted in graph representation learning and graph neural networks and it will be interesting to explore the intersection of graph neural networks and community detection. There are also lot of challenges that need to be tackled such as scalability on real world networks, detection of overlapping communities and also make algorithms robust to changing nature of real world graphs. These are some of the major problems that needs to be solved and can pave a way to new research directions.

REFERENCES

- [1] Yang, Liang Cao, Xiaochun He, Dongxiao Wang, Chuan Wang, Xiao Zhang, Weixiong. (2016). Modularity based community detection with deep learning.
- [2] Dhilber, M. Surampudi, Durga. (2019). Community Detection in Social Networks Using Deep Learning. 10.1007/978-3-030-36987-3_15.
- [3] Jia, Yuting Zhang, Qinqin Zhang, Weinan Wang, Xinbing. (2019). CommunityGAN: Community Detection with Generative Adversarial Nets.
- [4] Kumar, Sanjay Panda, B Aggarwal, Deepanshu. (2020). Community detection in complex networks using network embedding and gravitational search algorithm. *Journal of Intelligent Information Systems*. 1-22. 10.1007/s10844-020-00625-6.
- [5] Cavallari, Sandro Zheng, Vincent Cai, Hongyun Chang, Kevin Cambria, Erik. (2017). Learning Community Embedding with Community Detection and Node Embedding on Graphs. 377-386. 10.1145/3132847.3132925.
- [6] Chen, Z., Li, L., Bruna, J. (2019). Supervised Community Detection with Line Graph Neural Networks. *arXiv: Machine Learning*.
- [7] Zhu, Xiaojin Ghahramani, Zoubin. (2003). Learning from Labeled and Unlabeled Data with Label Propagation.
- [8] He, Kun Shi, Pan Bindel, David Hopcroft, John. (2019). Krylov Subspace Approximation for Local Community Detection in Large Networks. *ACM Transactions on Knowledge Discovery from Data*. 13. 52. 10.1145/3340708.
- [9] Hollocou, Alexandre Bonald, Thomas Lelarge, Marc. (2018). Multiple Local Community Detection. *ACM SIGMETRICS Performance Evaluation Review*. 45. 76-83. 10.1145/3199524.3199537.
- [10] Martin Ester Hans-Peter Kriegel Jiirg Sander Xiaowei Xu.(1996) A Density-Based Algorithm for Discovering Clusters. Institute for Computer Science, University of Munich in *Large Spatial Databases with Noise*
- [11] Venkatesaramani, Rajagopal Vorobeychik, Yevgeniy. (2018). Community Detection by Information Flow Simulation.
- [12] Bian, Yuchen Ni, Jingchao Cheng, Wei Zhang, Xiang. (2017). Many Heads are Better than One: Local Community Detection by the Multi-walker Chain. 21-30. 10.1109/ICDM.2017.11.
- [13] Ibrahim, Rania Gleich, David. (2019). Nonlinear Diffusion for Community Detection and Semi-Supervised Learning. *WWW '19: The World Wide Web Conference*. 739-750. 10.1145/3308558.3313483.
- [14] Škrlj, Blaž Kralj, Jan Lavrac, Nada. (2020). Embedding-based Silhouette community detection. *Machine Learning*. 109. 10.1007/s10994-020-05882-8.
- [15] Bhatt, Shreyansh Padhee, Swati Sheth, Amit Chen, Keke Shalin, Valerie Doran, Derek Minnery, Brandon. (2019). Knowledge Graph Enhanced Community Detection and Characterization. *WSDM '19: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 51-59. 10.1145/3289600.3291031.
- [16] Sun, Fan-Yun Qu, Meng Hoffmann, Jordan Huang, Chin-Wei Tang, Jian. (2019). vGraph: A Generative Model for Joint Community Detection and Node Representation Learning.
- [17] Ye, Fanghua Chen, Chuan Zheng, Zibin. (2018). Deep Autoencoder-like Nonnegative Matrix Factorization for Community Detection. 1393-1402. 10.1145/3269206.3271697.
- [18] Ye, Fanghua Li, Shenghui Lin, Zhiwei Chen, Chuan Zheng, Zibin. (2018). Adaptive Affinity Learning for Accurate Community Detection. 1374-1379. 10.1109/ICDM.2018.00188.
- [19] F. Hu, J. Liu, L. Li et al., Community detection in complex networks using Node2vec with spectral clustering, *Physica A* (2019), doi: <https://doi.org/10.1016/j.physa.2019.123633>.
- [20] Li, Xiang Kao, Ben Ren, Zhaochun Yin, Dawei. (2019). Spectral Clustering in Heterogeneous Information Networks.
- [21] Wu, Lingfei Chen, Pin-Yu Yen, Ian Xu, Fangli Xia, Yinglong Aggarwal, Charu. (2018). Scalable Spectral Clustering Using Random Binning Features.
- [22] Zhang, Y., Levina, E., Zhu, J. (2020). Detecting Overlapping Communities in Networks Using Spectral Methods. *SIAM J. Math. Data Sci.*, 2, 265-283.
- [23] Li, Y. He, Kun Bindel, David Hopcroft, J.E.. (2015). Uncovering the small community structure in large networks: a local spectral approach. *Proceedings of the 24th international conference on world wide web*. 658-668.
- [24] https://en.wikipedia.org/wiki/File:Network_Community_Structure.svg
- [25] https://www.researchgate.net/figure/Zacharys-karate-club-network-Members-of-the-communities-resulting-after-the-split-are_fig2_225168779
- [26] https://www.researchgate.net/figure/Sample-LFR-benchmark-graph-of-size-n-100-with-parameter-values-set-to-M-axDeg-30_fig8_325719324
- [27] <https://towardsdatascience.com/a-tale-of-two-convolutions-differing-design-paradigms-for-graph-neural-networks-8dadffa5b4b0>

STOCK PRICE ESTIMATION BASED ON HISTORICAL INFORMATION AND TEXTUAL SENTIMENT ANALYSIS

Abhay Gupta^{*1}, Aman Gupta^{*2}, Anshul Chaudhary^{*3}, Rajesh Kumar Yadav^{*4}

^{*1}Department of Computer Science & Engineering, Delhi Technological University, New Delhi, India.

^{*4}Assistant Professor Department of Computer Science & Engineering, Delhi Technological University, New Delhi, India.

ABSTRACT

Stock Market Price estimating has been a subject of interest among experts and scientists for quite a while. Stock costs are difficult to foresee due to their unstable nature, which relies upon an assortment of political and financial variables, change of administration, speculator feelings, and numerous different things. Foreseeing costs dependent on verifiable or literary information alone has ended up being lacking. A fruitful assessment of future share costs may capitulate a critical benefit. A stock trade hypothesis is the obligation to select future assessments of an entity's share or other currency-related items traded on a market. But there are methods and technologies that are supposed to allow us to get future price details. However, not a single good forecasting model has succeeded in beating the market trend further. As per the timetable information custom, theory is regularly founded on previous verifiable information and market patterns, authentic connection information and hypothesis can be determined.

Keywords: Stock Price Estimating, Financial Blogs, Sentiment Analysis, Prophet.

I. INTRODUCTION

In this precedented times, when people losing jobs, exhausting savings and struggling to cope with the economic challenges there are facing. Speculation is a protected choice for getting their future yet one doesn't have a clue where to contribute and the amount to contribute on the grounds that one couldn't say whether there will be a benefit or a misfortune on one's venture. This raises a fascinating issue on the grounds that a great many people for the most part wind up putting resources into any of the financial exchange areas. The answer for this issue permits us, find out about financial exchange choices and assisting with settling on it more precise choices. Stock market is most volatile and dynamic marketing system and to address this problem of chaotic and dynamic stock market we not only minimize our prediction based on technical factors^[18] rather considering, analysing and pre-processing the fundamental factors such as customer ratings, brand value, financial news sentiment, etc. of companies.

Stock market forecasts^[7] are an endeavour to anticipate what's to come, the estimation of an organization's stock or other monetary instrument available to be purchased to trade. stock exchange forecasts are moreover alluring tests. As indicated by a compelling business sector theory, stock costs ought to follow an arbitrary travel example and thusly ought not be the case can be anticipated with in excess of 50% precision.

Appropriate stock anticipating^[8] can prompt extraordinary advantages for both the dealer and the purchaser. Frequently, it is expressed that estimates are more turbulent than irregular, which implies that they can be anticipated via cautiously investigating securities exchange history. We in this way need a framework that can foresee the market esteem near the apparent worth, consequently expanding precision. Introduction of essential highlights like client suppositions, monetary sites, news, and so on nearby stock anticipating has pulled in numerous analysts as a result of its proficient and exact estimations.

In our project, we attempt to improve the exactness of stock worth gauges by get-together a ton of time game plan data. Apart from technical factors, we will also take into account fundamental factors like brand value, customer sentiments, ratings, performance graph, etc. to reflect upon the dynamic changes in the stock market and eventually provide stable and accurate results.

An essential goal of this task is to share the scholarly comprehension of more exact and exact financial exchange expectation.

To develop an accurate automated system for precise stock value prediction to have higher return on

investment and thus reducing risk and making greater profit.

To address the problem of chaotic and dynamic stock market we not only minimize our prediction based on technical factors rather considering, analysing and pre-processing the fundamental factors such as customer ratings, brand value, financial news sentiment, etc. of companies.

II. RELATED WORKS

Different procedures in the stock assumption space^[16] can be orchestrated into two social affairs. The essential social event joins tallies that attempt to improve the introduction of figure by redesigning the supposition models, an assortment of contraptions has been utilized, including Support Vector Machine^[4], LSTM, etc., while the sub optimal of assessments bases on improving the highlights subject to which the supposition that is in the primary get-together of the calculations that emphasis on the guess models.

STOCK PRICE PREDICTION BASED ON ANN

In 2011, a connection between the introduction of ANN plus, SVM was finished. Every count has its technique for learning models and subsequently expecting. Artificial Neural Network (ANN) is a notable and later procedure which similarly combine particular examination for making estimates in financial business areas. ANN^[5] incorporates a bunch of edge capacities. These capacities prepared on recorded information subsequent to associating each other with versatile loads and they are utilized to make future expectations. ANN fuses a lot of edge limits. These limits arranged on recorded data ensuing to partner each other with flexible burdens and they are used to make future assumptions. ANN can consider as a computation or a mathematical model which is animated by the utilitarian or essential ascribes of natural neural associations. These neural associations are made so it can remove plans from riotous data. ANN^[4] first train a system using a huge illustration of data known as getting ready stage then it familiarizes the association with the data which was avoided from the arrangement stage, this stage known as endorsement or assumption stage.

STOCK PRICE PREDICTION BASED ON LSTM

LSTM is that where the data having a spot with the past state drives forward. The main purpose behind using this model in protections trade figure is that the assessments depend on a ton of data and are all around dependent on the drawn-out history of the market. So, LSTM^[3] guides jumble up by giving a manual for the RNN through holding information for additional organized stages making the measure more careful. In this way approving itself as by and large more solid wandered from different strategies.

III. PROPOSED METHOD

Our approach is to examine the historical data of stocks of different companies and using that dataset to train our proposed model. After obtaining the processable data we select relevant features from this large dataset that may impact the prediction. It's an optimisation step and has a lot of importance as a good feature engineering leads to lesser time and space complexity of the project and more manageable system^[14].

Also, our model expects feature vectors derived from the processed data as input to get trained and make predictions. To yield excellent results, a set of technical as well as fundamental features could be used so that our prediction could get an overall exposure to all expectations that could affect the price of the stock.

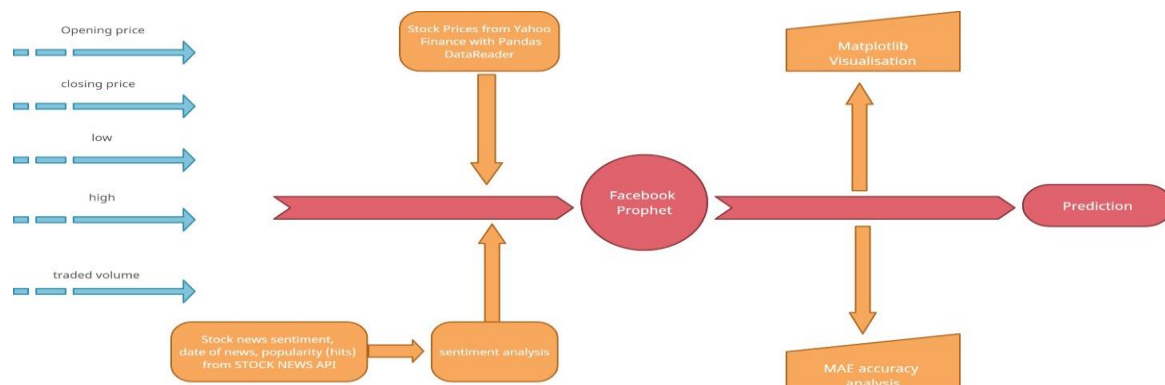


Figure 1 : Model Flow Chart

PRE-NECESSITIES

- Timeline testing can be done on a PC / workplace or on servers.
- To carry out the mentioned work, we will utilize the Jupyter Notebook. In the event that the product can now be imported, you need to launch and configure the Jupyter Notebook Python 3 pad.
- Operating on larger databases has increased memory and will need a system with any 2GB of working primary storage capacity to execute computations.

News Information:

Presently, official information sources to get authentic monetary news are restricted. Most sources like News API^[2] just permit admittance to the most recent couple of long stretches of information free of charge and it would cost roughly ₹70000 to assemble all the information important for this venture. I had the option to discover a news Programming interface, The Stock News API, giving news to stocks at similarly lesser cost and is a basic HTTPREST Programming interface that empowers us to get the most recent stock news from different uplifting news sources from web. We used the API to get news, date of publish of the news, sentiment of the news^[2] and its popularity of companies in the stock market.

Numerical Measures:

Opening Value - The underlying expense is the expense at which the security is first sold opening of the exchange upon the appearance of trading. The beginning exchanging cost of any stock is its own day by day opening cost. Cost open key imprint for the business movement of the day, particularly for those intrigued estimating transient outcomes like informal investors.

Closing Value - The end cost of the stock is the standard seat used to follow its exhibition over the long haul. The end cost is considered as the last an incentive at which the stock sold during the ordinary exchanging period day.

Low Price - The lowest price is the least selling cost in stock at the most recent day exchanging. Today's minimum is the lowest internal security trade Price.

High Price - Higher stock price means higher stock trading price. Today's top price is where the stock is sold at highest during trading day course. Top prices are often greater than closing or starting prices of stocks.

Volume - Volume alludes to the quantity of offers exchanged a given time span. A stock's volume alludes to the quantity of offers that are sold, or exchanged, throughout a specific timeframe (normally every day).

Stock Split - A stock split^[10] is the point at which an organization separates the current portions of its stock into numerous new offers to help the stock's liquidity.

A stock split or stock separation expands the quantity of offers in an organization. A stock split causes a decline of market cost of individual offers, not causing a difference in absolute market capitalization of the organization.

Dividend - It is a movement of advantages by an undertaking to its financial backers. Exactly when an association secures an advantage or overabundance, it can pay a degree of the advantage as a benefit to financial backers.

Here, including fundamental features is the tricky part as to get the sentiment of the stakeholders is not an easy task to execute. Albeit certain other basic highlights are effectively accessible in the library yfinance itself, for example, net revenues, profit, revenue on valuation, additional information for deciding an organization's capability for subsequent development, incomes and basic worth.

Our implementation includes sentiment analysis of the stakeholders and the news related to the stock to get hold of the market and make predictions to higher accuracies even when external factors affect the stock price.

It is seen that generally ARIMA^[1] is best utilized for expectation yet it isn't most appropriate for non-straight information designs. It gives best outcomes for time strategy which has a couple season(s) of authentic information and solid occasional impacts. Incredibly, the center fundamental Information Science group of Facebook organization distributed a fascinating new strategy as of late called by name Prophet, connects with information fashioners and analysts to test or perform surveying in Python at scale. Prophet^[1] is really a strategy to perform determining of time arrangement information mostly dependent on added substance model dissimilar to non-straight patterns that in everyday works with day by day, week after week, and yearly too for

irregularity, in addition to occasions.

Thus, we are using fbprophet^[1] library which gives a Prophet which is a method for determining measurement information maintained an added substance model where non-direct examples are fit with step by step, after quite a while after week, and yearly abnormality, notwithstanding occasion impacts. Prophet gives precise outcomes even if there should be an occurrence of missing data and movements in the pattern, and usually handles inconsistencies well. It works best with measurements that have strong occasional effects and different various times of recorded data.

Prophet is really an added substance framework that has the accompanying:

$$\mathbf{m}(t) = \mathbf{n}(t) + \mathbf{o}(t) + \mathbf{p}(t) + \epsilon \quad (1)$$

- $\mathbf{n}(t)$ represents the trend(s), which errands long stretch decay or extension in data. Prophet has two example related model(s), a piecewise straight model and a drenching improvement model, which depends upon the sort of guess.
- $\mathbf{o}(t)$ represents Fourier approach with inconsistency, that chooses how the information will be affected considering season related factor(s) like the season
- $\mathbf{p}(t)$ represents the special times of year effect^[11] or enormous occasions which profoundly impacts business time arrangement information (e.g., The day after Thanksgiving, New Item Dispatch, Superbowl and so forth)
- ϵ addresses a mistake term which is final.

We generate Visualisations^[6] of the predictions in the form of graphs and data frames to get a better insight of the predictions and analysis. This step is not a necessary one but is important in practical usage.

Performance Analysis using MAE:

This is the final step in the sequence where the evaluation of the model^[12] takes place which is very helpful in:

- Determining the risk factor involved in using the prophet and
- Analysing our model in comparison with other models based on technical analysis only or any other algorithm.

In the case of our model, we are using Mean Absolute Error as a performance metric to determine the accuracy of the model.

Mean absolute error is calculated as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |x_i - x| \quad (2)$$

where x represents the actual price and x_i represents the anticipated cost and n represents the quantity of questions or expectations made.

IV. RESULT AND ANALYSIS

We trained our model using technical factors such as opening and closing values, previous day prices etc. and sentiment analysis of fundamental properties to capture dynamic nature of stocks. Here, we were able to make predictions for the share values of Google for a span of 30 days starting from 16th October 2020 till 27th November 2020. The result is depicted in the below graph:

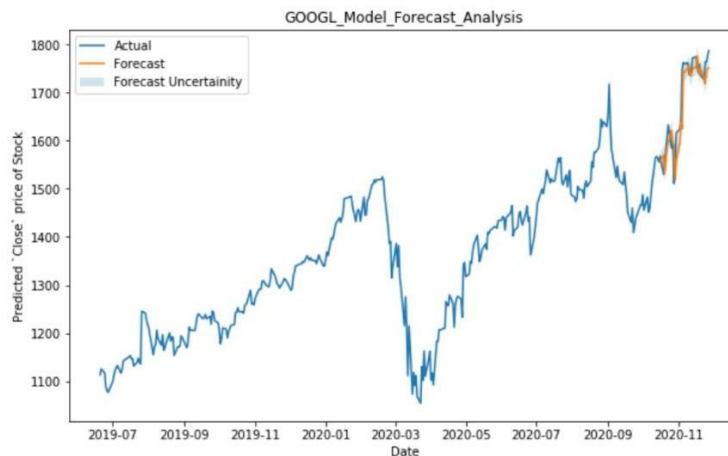


Figure 2 : Predicted closing price vs Dates mentioned

Below table depicts the forecasted values of all the factors we took into consideration for making speculations. Although time period was from 16th October 2020 to 27th November 2020 as mentioned above but below figure displays result for last ten days:

```
In [112]: forecast.tail(10)
```

```
Out[112]:
```

	ds	trend	yhat_lower	yhat_upper	trend_lower	trend_upper	Close_lag_1	Close_lag_1_lower	Close_lag_1_upper	Close_lag_2	...	weekly	w
20	2020-11-13	679.423509	1731.560805	1767.647013	679.423509	679.427412	393.493299	393.493299	393.493299	62.947036	...	63.242683	
21	2020-11-16	679.495258	1733.556780	1772.087405	679.495258	679.499727	404.683666	404.683666	404.683666	62.680167	...	63.558914	
22	2020-11-17	679.519174	1758.103089	1794.449862	679.519174	679.525006	405.356462	405.356462	405.356462	64.461708	...	63.885502	
23	2020-11-18	679.543091	1739.741495	1776.091396	679.543091	679.553962	400.654544	400.654544	400.654544	64.568819	...	64.546251	
24	2020-11-19	679.567007	1719.298726	1754.389763	679.567007	679.578265	392.664693	392.664693	392.664693	63.820259	...	63.891305	
25	2020-11-20	679.590924	1730.305532	1765.629516	679.590924	679.603487	399.479981	399.479981	399.479981	62.548250	...	63.242683	
26	2020-11-23	679.662673	1714.747425	1749.802974	679.662673	679.683326	391.045434	391.045434	391.045434	63.633265	...	63.558914	
27	2020-11-24	679.686589	1699.744034	1736.771137	679.686589	679.710099	387.692913	387.692913	387.692913	62.290459	...	63.885502	
28	2020-11-25	679.710505	1730.260097	1763.216656	679.710505	679.738286	401.505979	401.505979	401.505979	61.756728	...	64.546251	
29	2020-11-27	679.758338	1733.074671	1768.332612	679.758338	679.790455	401.593396	401.593396	401.593396	63.955810	...	63.242683	

10 rows x 94 columns

Figure 3 : Predicted values of all the factors considered

V. CONCLUSION AND FUTURE WORK

The mean absolute error for our model is 27% initially, that is, we were able to forecast the share values for Google for said period with an accuracy of 73%. But later by including fundamental features such as profit margins, return on equity, earnings, analysis of stock values in news, etc.; we further increased the accuracy of our model to 82%, that is, reducing the absolute error to 18%.

We intended to train our model on both technical factors as well as basic indicators like financial analysis in news, year of establishment, turnover, public sentiments etc.

Our work can be additionally broadened utilizing deep learning for predicting stock values for multi-national organization yet this may fuse huge space and time complexity than our proposed approach.

VI. REFERENCES

- [1] M. B. Patel and S. R. Yalamalle, "Stock price prediction using artificial neural network," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 3, no. 6, pp. 13 755–13 762, 2014.
- [2] E. J. De Fortuny, T. De Smedt, D. Martens, and W. Daelemans, "Evaluating and understanding text-based stock price prediction models," *Information Processing & Management*, vol. 50, no. 2, pp. 426–441, 2014.
- [3] A. A. Adebiyi, A. O. Adewumi, and C. K. Ayo, "Comparison of arima and artificial neural networks models for stock price prediction," *Journal of Applied Mathematics*, vol. 2014, 2014.
- [4] K. Kohara, T. Ishikawa, Y. Fukuhara, and Y. Nakamura, "Stock price prediction using prior knowledge and neural networks," *Intelligent Systems in Accounting, Finance & Management*, vol. 6, no. 1, pp. 11–22, 1997.
- [5] Y. E. Cakra and B. D. Trisedya, "Stock price prediction using linear regression based on sentiment analysis," in *2015 international conference on advanced computer science and information systems (ICACSIS). IEEE*, 2015, pp. 147–154.
- [6] S. Deng, T. Mitsubuchi, K. Shioda, T. Shimada, and A. Sakurai, "Combining technical analysis with sentiment analysis for stock price prediction," in *2011 IEEE ninth international conference on dependable, autonomic and secure computing. IEEE*, 2011, pp. 800–807.
- [7] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," *Knowledge-Based Systems*, vol. 69, pp. 14–23, 2014.
- [8] Z. Jin, Y. Yang, and Y. Liu, "Stock closing price prediction based on sentiment analysis and lstm," *Neural Computing and Applications*, pp. 1–17, 2019.
- [9] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia, and D. C. Anastasiu, "Stock price prediction using news sentiment analysis," in *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (Big Data Service). IEEE*, 2019, pp. 205–208.
- [10] M. Lam, "Neural network techniques for financial performance prediction: integrating fundamental and technical analysis," *Decision support systems*, vol. 37, no. 4, pp. 567–581, 2004.
- [11] A. A. Adebiyi, C. K. Ayo, M. O. Adebiyi, and S. O. Otokiti, "Stock price prediction using neural network with hybridized market indicators," *Journal of Emerging Trends in Computing and Information Sciences*, vol. 3, no. 1, pp. 1–9, 2012.
- [12] C. R. Madhuri, M. Chinta, and V. P. Kumar, "Stock market prediction for time-series forecasting using prophet upon arima," in *2020 7th International Conference on Smart Structures and Systems (ICSSS). IEEE*, 2020, pp. 1–5.
- [13] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia, and D. C. Anastasiu, "Stock price prediction using news sentiment analysis," in *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (Big Data Service). IEEE*, 2019, pp. 205–208.
- [14] I. Parmar, N. Agarwal, S. Saxena, R. Arora, S. Gupta, H. Dhiman, and L. Chouhan, "Stock market prediction using machine learning," in *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC). IEEE*, 2018, pp. 574–576.
- [15] B. Qian and K. Rasheed, "Stock market prediction with multiple classifiers," *Applied Intelligence*, vol. 26, no. 1, pp. 25–33, 2007.
- [16] L. Sayavong, Z. Wu, and S. Chalita, "Research on stock price prediction method based on convolutional neural network," in *2019 International Conference on Virtual Reality and Intelligent Systems (ICVRIS). IEEE*, 2019, pp. 173–176.

Strongly coupled plasma effect on excitation energies of O-like ions and photoionization of F-like ions

R Sharma² and A Goyal^{1*}

¹Department of Physics, Shyam Lal College, University of Delhi, Delhi 110032, India

²Department of Applied Physics, Delhi Technological, Delhi 110042, India

Received: 16 August 2020 / Accepted: 05 April 2021

Abstract: The main goal of the present study is to analyze and estimate the plasma screening effect on excitation energies O-like ions and photoionization process of F-like ions under the influence of strongly coupled plasma. We have employed an ion sphere model (ISM) in flexible atomic code (FAC) to study and analyze plasma effect in atomic structure of O-like ions and photoionization of F-like ions. We have presented shift in excitation energies of lowest 10 levels of Fe XIX, Co XX, Ni XXI, Cu XXII and Zn XXIII at electron densities ranges 10^{22} – 10^{24} cm⁻³ under plasma environment. We have also compared our transition wavelengths for Fe XIX with wavelengths observed in ENEA laser facility and calculated from Hebrew University Lawrence Livermore Atomic Code (HULLAC) code at electron density 10^{21} cm⁻³. The lowering in ionization potentials and effect on photoionization cross sections for ground and first excited states of Fe XVIII, Co XIX, Ni XX, Cu XXI and Zn XXII also studied at electron densities ranges 10^{22} – 10^{24} cm⁻³ at five photo-electron energies ranges 100–500 eV.

Keywords: Energy levels; Radiative data; Spectroscopic parameters; Transition wavelength

1. Introduction

From last few years, the theoretical and experimental research on the study of plasma screening effect on atomic properties of highly charged ions immersed in hot and dense plasma has been increased remarkably [1–15]. Due to potential applications in plasma spectroscopy, inertial confinement fusion and astrophysics, atomic physics in hot and dense plasma has become an emerging and developing field. The screening effects help in the examination and investigation of radiation emitted from plasma and physical phenomenon such as continuum lowering and pressure ionization [16]. These effects play a significant role in the determination of fruitful details about change in spectral line shifts, line broadening and profiles of ingrained atoms or ions. The hot and dense plasma effects on atomic properties have also great interest in the diagnosis of plasmas, opacities as well as in the study of quantum field

properties. Atomic processes such as photoionization (PI), electron impact excitation (EIE), radiative recombination (RR), etc., of atoms/ions also exhibit prominent variation under the effect of plasma environment [17, 18]. Under the influence of hot dense plasma environment, the plasma electrons and ions modify atomic potential thereby affecting atomic processes of ions/atoms. The study of photoionization is useful in the computation of charge state population and radiative properties of plasma.

Depending on the value of coupling constant, the plasma environment can be divided into two categories, weakly and strongly coupled plasma (SCP). For weakly coupled plasma (WCP), the screening effect can be introduced by Debye model. While for strongly coupled plasma, ion sphere (IS) model has been adopted. Both models have been employed several times in the past for different atomic systems embedded in plasma [19–40]. Apart from these two models, several other models such as polarized-correlation sphere model [41], hybrid model [42], nonlinear Debye–Hückel model [43] and Stewart–Pyatt model [44] have also been applied for describing plasma screening effect for different ranges of temperature and density. Further, for the study of properties of hot and dense

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s12648-021-02106-0>.

*Corresponding author, E-mail: arun.goyal.du@gmail.com

plasma, hypernetted-chain approximation and density functional theory has been applied by Ichimaru [45] and Dharma-Wardana et al. [46]. The SCP plasma has been identified laser-produced plasma, inertial confinement fusion, etc. Most of the atomic physicists apply ion sphere model for the investigation of SCP. Polarizability and hyper polarizability of H and He-like ions in SCP have been studied by Sen et al. [47]. Basu et al. [48] have presented dynamic polarizability using IS model. Bhattacharya et al. [49] and Das et al. [23] have calculated plasma screening effect of SCP on transition energies, oscillator strengths, etc., for H-like and Li-like ions resp. M. Das [50] have studied effect of SCP on ground state photoionizations for H-like and Li-like ions by solving radial equation with shooting method approach. Li et al. [51] have presented the variation of exchange energy shifts with temperature and density for He-like Al. by applying ions sphere model. Li et al. [52] have investigated the effect of hot and dense plasma on excitation energies and oscillator strengths of Be-like ions. The energies of doubly excited states of He atom and Li^+ have been computed by Saha et al. [22] under the influence of weakly coupled plasma by Debye model. Recently, Das et al. [53] have studied the variation of excitation energies and ionization potential with Debye screening length and IS radius for Al atom and its ions under the influence of weakly and strongly coupled plasma. In the past, a detailed study of weakly coupled plasma on photoionization is available in the literature [54–59], while only Jung et al. [60] and M. Das [50] have studied photoionization process under SCP.

In previous studies, atomic physicists have implemented different methods for solving Schrödinger equation to determine wave functions and atomic parameters of atoms/ions embedded in strongly coupled plasmas. Das et al. [23] have used Fock-space multi-reference coupled cluster (FS-MRCC) level of theory for the calculation of excitation energies of Li-like ions. M. Das [50] have solved Schrödinger equation by employing Shooting method approach and Runge–Kutta method. Li et al. [51] have applied self-consistent field method for the determination of wave functions and electron density for bound and free electrons. While Li et al. [52] have modified multi-configuration Dirac Fock (MCDF) method in GRASP2 [61] by adding plasma screening effect in one-electron potential. Saha et al. [22] have utilized trial wave functions in the Hylleraas basis set to introduce correlation effects in Ritz variational method. Recently, Das et al. [53] paper has adopted relativistic cluster coupled (RCC) method based on perturbation theory.

In various astronomical objects such as sun, stars and galaxies and laboratory plasmas, the spectra of O-like ions of iron period have been observed [62–80]. The super solar abundances for oxygen and other elements from Chandra

spectrum of gas flowing out of the active galactic nucleus have been observed by Fields et al. [81]. In this observation, they determined that the abundance of oxygen is eight times solar. Due to the large range of applications of these ions for the analysis of spectra from astrophysical sources and modeling and diagnosis of different type of plasmas, the study of atomic processes of these ions under the influence of plasma is also needed. Garcia et al. [82] have also provided atomic data for the modeling of spectra of K lines of oxygen of astrophysical photoionized plasmas. Recently, Deprince et al. [83] have studied plasma environment effect on energy levels, radiative data and auger widths for oxygen ions from neutral oxygen to O VII. To analyze plasma effects, they have implemented Debye–Hückel potential in multi-configuration Dirac–Fock (MCDF) method. In the past, Khattak et al. [84] have determined a line shift for Ti XXI at electron density greater than 10^{24} cm^{-3} from laser-produced plasma which was again produced by Belkhiri et al. [85] using an ion sphere model. But till date, high density plasma embedding with ion sphere model (ISM) for density greater than equal to 10^{22} cm^{-1} has not been taken into account for study of atomic parameters of O-like ions and photoionization of F-like ions. So the main goal of the present work is to analyze the plasma environment effect on O-like ions and F-like ions of iron period by introducing ion sphere model potential.

2. Theoretical approach

In the literature, plasma environment effect using ISM has been studied by various methods, codes and approaches. So, we will discuss here only in brief. For highly charged ions embedded in strongly coupled plasma, it is assumed that electrons inside ion sphere interact strongly with embedded ions. In our calculations, we have used modified flexible atomic code (FAC) [86] for studying plasma environment effect. The effective potential for strongly coupled plasma embedded atomic system is given by [36],

$$V_{\text{eff}}(r_i) = -\frac{Z}{r_i} + \frac{(Z - N)}{2R} \left[3 - \left(\frac{r_i}{R} \right)^2 \right] \quad (1)$$

where Z , N and R denotes nuclear charge, N is no. of bound electrons and R is ion sphere radius. This ion sphere radius can be easily determined from the following relation:

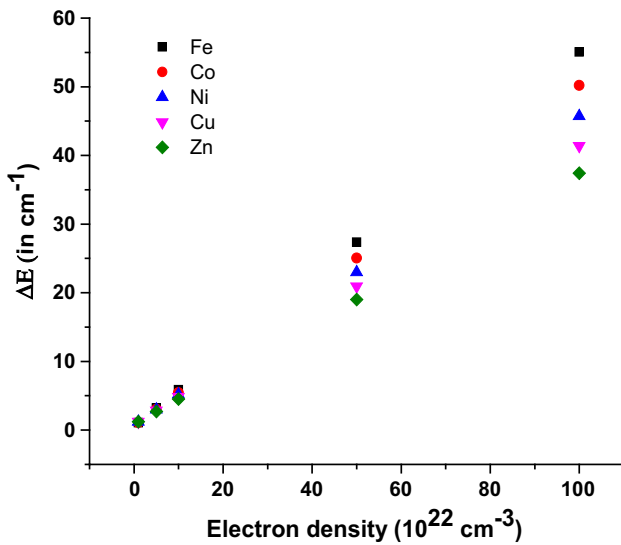
$$R = \left(\frac{3}{4\pi n} \right)^{1/3} \quad (2)$$

Table 1 Difference in excitation energies (in cm^{-1}) with and without plasma for lowest 10 levels of O-like ions with electron density (in cm^{-3})

Fe								
Level No	Configuration	2 J	Parity	1E + 22	5E + 22	1E + 23	5E + 23	1E + 24
1	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	4	+	0.000	0.000	0.000	0.000	0.000
2	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	0	+	1.089	3.217	5.872	27.355	55.060
3	$2s_{1/2}^2 2p_{1/2} 2p_{3/2}^3$	2	+	1.215	3.713	6.814	31.566	62.408
4	$2s_{1/2}^2 2p_{1/2} 2p_{3/2}^3$	4	+	1.262	5.036	9.742	47.382	94.404
5	$2s_{1/2}^2 2p_{3/2}^4$	0	+	2.813	11.068	21.397	105.149	213.300
6	$2s_{1/2} 2p_{1/2}^2 2p_{3/2}^3$	4	−	8.840	48.167	97.407	492.004	987.637
7	$2s_{1/2} 2p_{1/2}^2 2p_{3/2}^3$	2	−	8.267	46.143	93.584	473.888	951.818
8	$2s_{1/2} 2p_{1/2} 2p_{3/2}^4$	0	−	7.662	44.005	89.543	454.588	913.289
9	$2s_{1/2} 2p_{1/2} 2p_{3/2}^4$	2	−	8.404	46.473	94.189	478.147	965.174
10	$2p_{1/2}^2 2p_{3/2}^4$	0	+	18.962	102.061	206.195	1044.837	2111.185
Co								
Level No	Configuration	2 J	Parity	1E + 22	5E + 22	1E + 23	5E + 23	1E + 24
1	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	4	+	0.000	0.000	0.000	0.000	0.000
2	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	0	+	1.074	3.029	5.449	25.056	50.204
3	$2s_{1/2}^2 2p_{1/2} 2p_{3/2}^3$	2	+	1.276	3.816	6.938	31.976	63.225
4	$2s_{1/2}^2 2p_{1/2} 2p_{3/2}^3$	4	+	1.253	4.880	9.378	45.441	90.520
5	$2s_{1/2}^2 2p_{3/2}^4$	0	+	2.712	10.309	19.773	96.297	194.024
6	$2s_{1/2} 2p_{1/2}^2 2p_{3/2}^3$	4	−	8.098	44.562	90.196	455.773	914.257
7	$2s_{1/2} 2p_{1/2}^2 2p_{3/2}^3$	2	−	7.533	42.618	86.551	438.505	880.083
8	$2s_{1/2} 2p_{1/2} 2p_{3/2}^4$	0	−	6.880	40.368	82.333	418.478	840.151
9	$2s_{1/2} 2p_{1/2} 2p_{3/2}^4$	2	−	7.557	42.465	86.212	437.578	881.188
10	$2p_{1/2}^2 2p_{3/2}^4$	0	+	17.147	93.334	188.725	955.705	1925.719
Ni								
Level No	Configuration	2 J	Parity	1E + 22	5E + 22	1E + 23	5E + 23	1E + 24
1	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	4	+	0.000	0.000	0.000	0.000	0.000
2	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	0	+	1.205	2.993	5.201	22.991	45.716
3	$2s_{1/2}^2 2p_{1/2} 2p_{3/2}^3$	2	+	1.462	4.038	7.206	32.493	64.088
4	$2s_{1/2}^2 2p_{1/2} 2p_{3/2}^3$	4	+	1.310	4.817	9.166	43.952	87.453
5	$2s_{1/2}^2 2p_{3/2}^4$	0	+	2.850	9.966	18.821	90.096	180.491
6	$2s_{1/2} 2p_{1/2}^2 2p_{3/2}^3$	4	−	7.259	41.118	83.502	423.035	848.353
7	$2s_{1/2} 2p_{1/2}^2 2p_{3/2}^3$	2	−	6.686	39.231	79.994	406.646	815.938
8	$2s_{1/2} 2p_{1/2} 2p_{3/2}^4$	0	−	5.944	36.827	75.533	385.781	774.433
9	$2s_{1/2} 2p_{1/2} 2p_{3/2}^4$	2	−	6.606	38.690	78.881	401.605	807.736
10	$2p_{1/2}^2 2p_{3/2}^4$	0	+	15.276	85.389	173.166	878.107	1766.482
Cu								
Level No	Configuration	2 J	Parity	1E + 22	5E + 22	1E + 23	5E + 23	1E + 24
1	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	4	+	0.000	0.000	0.000	0.000	0.000

Table 1 continued

Cu								
Level No	Configuration	2 J	Parity	1E + 22	5E + 22	1E + 23	5E + 23	1E + 24
2	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	0	+	1.226	2.854	4.843	20.926	41.389
3	$2s_{1/2}^2 2p_{1/2} 2p_{3/2}^3$	2	+	1.577	4.180	7.338	32.899	64.780
4	$2s_{1/2}^2 2p_{1/2} 2p_{3/2}^3$	4	+	1.353	4.759	8.946	42.740	84.944
5	$2s_{1/2}^2 2p_{3/2}^4$	0	+	2.916	9.677	18.038	85.518	170.650
6	$2s_{1/2} 2p_{1/2}^2 2p_{3/2}^3$	4	−	6.323	37.819	77.292	393.233	788.766
7	$2s_{1/2} 2p_{1/2}^2 2p_{3/2}^3$	2	−	5.743	35.996	73.969	377.756	758.205
8	$2s_{1/2} 2p_{1/2} 2p_{3/2}^4$	0	−	4.912	33.440	69.307	356.011	715.047
9	$2s_{1/2} 2p_{1/2} 2p_{3/2}^4$	2	−	5.557	35.093	72.194	369.295	742.534
10	$2p_{1/2}^2 2p_{3/2}^4$	0	+	13.244	77.961	159.052	809.510	1627.368
Zn								
Level No	Configuration	2 J	Parity	1E + 22	5E + 22	1E + 23	5E + 23	1E + 24
1	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	4	+	0.000	0.000	0.000	0.000	0.000
2	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	0	+	1.228	2.713	4.516	19.021	37.427
3	$2s_{1/2}^2 2p_{1/2} 2p_{3/2}^3$	2	+	1.626	4.252	7.461	33.220	65.390
4	$2s_{1/2}^2 2p_{1/2} 2p_{3/2}^3$	4	+	1.341	4.659	8.762	41.703	82.872
5	$2s_{1/2}^2 2p_{3/2}^4$	0	+	2.927	9.426	17.468	82.119	163.460
6	$2s_{1/2} 2p_{1/2}^2 2p_{3/2}^3$	4	−	5.831	35.254	72.073	366.512	735.130
7	$2s_{1/2} 2p_{1/2}^2 2p_{3/2}^3$	2	−	5.291	33.569	68.968	352.051	706.546
8	$2s_{1/2} 2p_{1/2} 2p_{3/2}^4$	0	−	4.412	30.891	64.088	329.411	661.637
9	$2s_{1/2} 2p_{1/2} 2p_{3/2}^4$	2	−	5.009	32.351	66.589	340.638	684.592
10	$2p_{1/2}^2 2p_{3/2}^4$	0	+	12.147	72.202	147.365	749.382	1505.514

**Fig. 1** Variation of shift in excitation energies for level No. 2 of O-like ions in the presence of plasma environment as a function of electron density

where n is electron density. Here, the effect of plasma temperature is also considered in determining electron density for the calculation of ion sphere radius [16].

For N electron systems, the Dirac Hamiltonian in ISM is given by

$$H = \sum_{i=1}^N [c \vec{\alpha} \cdot \vec{p} + (\beta - 1)c^2 + V_{eff}(r_i)] + \sum_{i<j} \frac{1}{r_{ij}} \quad (3)$$

where $\vec{\alpha}$ and β are the Dirac matrices and c is the speed of light.

3. Results and discussion

3.1. Excitation energies

In present calculation, we have calculated excitation energies for lowest 10 levels of O-like Fe to Zn with and

Table 2 Comparison of transition wavelengths (in Å) for Fe XIX using Ion sphere model (ISM) in FAC with theoretically calculated and experimentally observed wavelengths at electron density 10^{21} cm^{-3}

S. No	Transitions				ISM	HULLAC [77]	Exp. [77]	ΔE_1	ΔE_2
	Upper level		Lower level						
	Configuration	J	Configuration	J					
1	$2s^2 2p_{1/2} 2p_{3/2}^2 3p_{3/2}$	2	$2s 2p_{1/2} 2p_{3/2}^4$	1	16.9485	16.9887	16.936	0.0125	0.0527
2	$2s^2 2p_{1/2} 2p_{3/2}^2 3s$	2	$2s^2 2p_{1/2} 2p_{3/2}^3$	2	15.1249	14.9968			
3	$2s^2 2p_{1/2} 2p_{3/2}^2 3s$	3	$2s^2 2p_{1/2}^2 2p_{3/2}^2$	2	14.6826	14.6687	14.694	0.0114	0.0253
4	$2s^2 2p_{1/2} 2p_{3/2}^2 3d_{5/2}$	3	$2s^2 2p_{1/2}^2 2p_{3/2}^2$	2	13.8005	13.7904	13.798	0.0025	0.0076
5	$2s^2 2p_{1/2} 2p_{3/2}^2 3d_{5/2}$	1	$2s^2 2p_{1/2} 2p_{3/2}^3$	2	13.7897	13.7893	13.778	0.0117	0.0113
6	$2s^2 2p_{1/2} 2p_{3/2}^2 3d_{5/2}$	2	$2s^2 2p_{1/2} 2p_{3/2}^3$	1	13.7190	13.7012	13.716	0.003	0.0148
7	$2s^2 2p_{1/2} 2p_{3/2}^2 3d_{5/2}$	1	$2s^2 2p_{1/2} 2p_{3/2}^3$	1	13.6713	13.6569			
8	$2s 2p_{3/2}^4 3d_{3/2}$	2	$2s 2p_{1/2} 2p_{3/2}^4$	1	13.6540	13.6404			
9	$2s 2p_{1/2} 2p_{3/2}^3 3d_{5/2}$	3	$2s 2p_{1/2}^2 2p_{3/2}^3$	2	13.5873	13.5727			
10	$2s^2 2p_{1/2} 2p_{3/2}^2 3d_{5/2}$	3	$2s^2 2p_{1/2}^2 2p_{3/2}^2$	2	13.5249	13.5127			
11	$2s^2 2p_{1/2} 2p_{3/2}^2 3d_{3/2}$	2	$2s^2 2p_{1/2}^2 2p_{3/2}^2$	2	13.5096	13.4996			
12	$2s^2 2p_{3/2}^3 3d_{3/2}$	3	$2s^2 2p_{1/2} 2p_{3/2}^3$	2	13.4975	13.4863			
13	$2s^2 2p_{3/2}^3 3d_{3/2}$	2	$2s^2 2p_{1/2} 2p_{3/2}^3$	2	13.4951	13.4853			
14	$2s^2 2p_{1/2} 2p_{3/2}^2 3d_{5/2}$	2	$2s^2 2p_{1/2} 2p_{3/2}^3$	1	13.4972	13.4849			
15	$2s^2 2p_{1/2} 2p_{3/2}^2 3d_{3/2}$	1	$2s^2 2p_{1/2}^2 2p_{3/2}^2$	2	13.4724	13.4596			
16	$2s 2p_{1/2}^2 2p_{1/2}^2 3d_{5/2}$	2	$2s 2p_{1/2}^2 2p_{3/2}^3$	1	13.4339	13.4233			
17	$2s^2 2p_{3/2}^3 3d_{3/2}$	1	$2s^2 2p_{1/2} 2p_{3/2}^3$	1	13.4166	13.4040			
18	$2s^2 2p_{1/2} 2p_{3/2}^2 3d_{3/2}$	3	$2s^2 2p_{1/2}^2 2p_{3/2}^2$	1	13.6400	13.6434			

without plasma by using FAC. In Table 1, we have listed difference in excitation energies calculated in the absence and presence of plasma environment, as a function of electron density. In Table 1, last five columns represent the electron density in cm^{-3} . From Table 1, we see that the shift in energy levels for 2–5 levels of $2s^2 2p^4$ and 6–9 levels of $2s 2p^5$ is almost close to each other while shifting for level number 10 of $2p^6$ is very large for electron density 10^{22} – 10^{24} cm^{-3} for all O-like ions. This shows that transition energies for levels of $2s^2 2p^4$ are blue shifted w.r.t that for $2s 2p^5$ while transition energies for levels of $2s 2p^5$ are blue shifted w.r.t that for $2p^6$ due to the electron screening and quantum confinement. So, our study shows that higher angular momentum quantum number levels for same principal quantum number have greater quantum confinement effect. In Table 1, we have provided the position of levels in the absence of plasma under the column “level no.”. We have also predicted that in the presence of plasma environment, the position of 6–9 levels changes as the shift in excitation energies for 6–9 levels is in the following order for O-like Fe and O-like Co

$$\Delta E_6 > \Delta E_9 > \Delta E_7 > \Delta E_8 \quad (4)$$

where ΔE represents shift in energy. While for other O-like ions, the order is

$$\Delta E_6 > \Delta E_7 > \Delta E_9 > \Delta E_8 \quad (5)$$

So, for other ions, only level numbers 8 and 9 are interchanged due to plasma environment.

From Table 1 and Fig. 1, our study also shows that with increasing electron density, the effect of electron screening increases the difference in excitation energies drastically for each O-like ion. It is also observed that the difference in excitation energies for all O-like ions is significant at high electron density 10^{24} cm^{-3} , while at other electron densities, the gap is not so large.

As we see the plasma environment effect on other O-like ions with increasing nuclear charge, a similar trend is observed but shift in excitation energies is decreasing for a particular electron density greater than 10^{22} cm^{-3} . This decrease is clearly visible in Fig. 1 for electron density 10^{24} cm^{-3} , and the gap between two consecutive ions is almost same. This implies that increase in nuclear charge

Table 3 Ionization potential (in eV) with plasma for photoionization of ground and first excited states of F-like ions to lowest 5 levels of O-like ions with electron density (in cm^{-3})

Fe											
S. No	Upper level	2 J	P	Lower level	2 J	P	1E + 22	5E + 22	1E + 23	5E + 23	1E + 24
1	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	4	+	1.3039E + 03	1.2673E + 03	1.2444E + 03	1.1659E + 03	1.1169E + 03
2	$2p_{1/2}^2 2p_{3/2}^4$	1	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	4	+	1.2912E + 03	1.2546E + 03	1.2317E + 03	1.1532E + 03	1.1042E + 03
3	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	0	+	1.3130E + 03	1.2764E + 03	1.2536E + 03	1.1750E + 03	1.1260E + 03
4	$2p_{1/2}^2 2p_{3/2}^4$	1	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	0	+	1.3003E + 03	1.2637E + 03	1.2409E + 03	1.1623E + 03	1.1133E + 03
5	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	2	+	1.3149E + 03	1.2783E + 03	1.2555E + 03	1.1769E + 03	1.1279E + 03
6	$2p_{1/2}^2 2p_{3/2}^4$	1	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	2	+	1.3022E + 03	1.2656E + 03	1.2428E + 03	1.1642E + 03	1.1152E + 03
7	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	4	+	1.3247E + 03	1.2882E + 03	1.2653E + 03	1.1868E + 03	1.1378E + 03
8	$2p_{1/2}^2 2p_{3/2}^4$	1	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	4	+	1.3120E + 03	1.2755E + 03	1.2526E + 03	1.1741E + 03	1.1251E + 03
9	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2 2p_{3/2}^4$	0	+	1.3434E + 03	1.3068E + 03	1.2839E + 03	1.2054E + 03	1.1564E + 03
10	$2p_{1/2}^2 2p_{3/2}^4$	1	—	$2s_{1/2}^2 2p_{3/2}^2$	0	+	1.3307E + 03	1.2941E + 03	1.2712E + 03	1.1927E + 03	1.1437E + 03
Co											
S. No	Upper level	2 J	P	Lower level	2 J	P	1E + 22	5E + 22	1E + 23	5E + 23	1E + 24
1	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	4	+	1.4424E + 03	1.4045E + 03	1.3808E + 03	1.2993E + 03	1.2485E + 03
2	$2p_{1/2}^2 2p_{3/2}^4$	1	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	4	+	1.4273E + 03	1.3894E + 03	1.3657E + 03	1.2842E + 03	1.2334E + 03
3	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	0	+	1.4526E + 03	1.4147E + 03	1.3909E + 03	1.3095E + 03	1.2586E + 03
4	$2p_{1/2}^2 2p_{3/2}^4$	1	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	0	+	1.4375E + 03	1.3996E + 03	1.3759E + 03	1.2944E + 03	1.2435E + 03
5	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	2	+	1.4557E + 03	1.4178E + 03	1.3941E + 03	1.3126E + 03	1.2617E + 03
6	$2p_{1/2}^2 2p_{3/2}^4$	1	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	2	+	1.4006E + 03	1.4027E + 03	1.3790E + 03	1.2975E + 03	1.2466E + 03
7	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	4	+	1.4658E + 03	1.4279E + 03	1.4042E + 03	1.3227E + 03	1.2719E + 03
8	$2p_{1/2}^2 2p_{3/2}^4$	1	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	4	+	1.4507E + 03	1.4128E + 03	1.3891E + 03	1.3076E + 03	1.2568E + 03
9	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2 2p_{3/2}^4$	0	+	1.4866E + 03	1.4487E + 03	1.4250E + 03	1.3436E + 03	1.2927E + 03
10	$2p_{1/2}^2 2p_{3/2}^4$	1	—	$2s_{1/2}^2 2p_{3/2}^2$	0	+	1.4715E + 03	1.4337E + 03	1.4099E + 03	1.3285E + 03	1.2776E + 03
Ni											
S. No	Upper level	2 J	P	Lower level	2 J	P	1E + 22	5E + 22	1E + 23	5E + 23	1E + 24
1	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	4	+	1.5878E + 03	1.5486E + 03	1.5240E + 03	1.4397E + 03	1.3807E + 03
2	$2p_{1/2}^2 2p_{3/2}^4$	1	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	4	+	1.5700E + 03	1.5307E + 03	1.5062E + 03	1.4219E + 03	1.3692E + 03
3	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2 2p_{1/2}^2 2p_{3/2}^2$	0	+	1.5990E + 03	1.5598E + 03	1.5352E + 03	1.4509E + 03	1.3983E + 03

Table 3 continued

Ni											
S. No	Upper level	2 J	P	Lower level	2 J	P	1E + 22	5E + 22	1E + 23	5E + 23	1E + 24
4	$2p_{1/2}^1 2p_{3/2}^4$	1	—	$2s_{1/2}^2, 2p_{1/2}^2, 2p_{3/2}^2$	0	+	1.5812E + 03	1.5420E + 03	1.5174E + 03	1.4331E + 03	1.3804E + 03
5	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2, 2p_{1/2}^1, 2p_{3/2}^2$	2	+	1.6036E + 03	1.5644E + 03	1.5399E + 03	1.4555E + 03	1.4029E + 03
6	$2p_{1/2}^1 2p_{3/2}^4$	1	—	$2s_{1/2}^2, 2p_{1/2}^1, 2p_{3/2}^3$	2	+	1.5858E + 03	1.5466E + 03	1.5221E + 03	1.4377E + 03	1.3851E + 03
7	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2, 2p_{1/2}^2, 2p_{3/2}^2$	4	+	1.6141E + 03	1.5748E + 03	1.5503E + 03	1.4660E + 03	1.4133E + 03
8	$2p_{1/2}^1 2p_{3/2}^4$	1	—	$2s_{1/2}^2, 2p_{1/2}^1, 2p_{3/2}^3$	4	+	1.5962E + 03	1.5570E + 03	1.5325E + 03	1.4481E + 03	1.3955E + 03
9	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2, 2p_{3/2}^4$	0	+	1.6375E + 03	1.5983E + 03	1.5737E + 03	1.4894E + 03	1.4367E + 03
10	$2p_{1/2}^1 2p_{3/2}^4$	1	—	$2s_{1/2}^2, 2p_{3/2}^4$	0	+	1.6197E + 03	1.5804E + 03	1.5559E + 03	1.4716E + 03	1.4189E + 03
Cu											
S. No	Upper level	2 J	P	Lower level	2 J	P	1E + 22	5E + 22	1E + 23	5E + 23	1E + 24
1	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2, 2p_{1/2}^2, 2p_{3/2}^2$	4	+	1.7401E + 03	1.6995E + 03	1.6742E + 03	1.5871E + 03	1.5326E + 03
2	$2p_{1/2}^1 2p_{3/2}^4$	1	—	$2s_{1/2}^2, 2p_{1/2}^2, 2p_{3/2}^2$	4	+	1.7192E + 03	1.6786E + 03	1.6533E + 03	1.5661E + 03	1.5117E + 03
3	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2, 2p_{1/2}^2, 2p_{3/2}^2$	0	+	1.7523E + 03	1.7118E + 03	1.6865E + 03	1.5993E + 03	1.5449E + 03
4	$2p_{1/2}^1 2p_{3/2}^4$	1	—	$2s_{1/2}^2, 2p_{1/2}^2, 2p_{3/2}^2$	0	+	1.7314E + 03	1.6909E + 03	1.6655E + 03	1.5784E + 03	1.5240E + 03
5	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2, 2p_{1/2}^1, 2p_{3/2}^3$	2	+	1.7588E + 03	1.7183E + 03	1.6930E + 03	1.6058E + 03	1.5514E + 03
6	$2p_{1/2}^1 2p_{3/2}^4$	1	—	$2s_{1/2}^2, 2p_{1/2}^1, 2p_{3/2}^3$	2	+	1.7379E + 03	1.6974E + 03	1.6721E + 03	1.5849E + 03	1.5305E + 03
7	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2, 2p_{1/2}^2, 2p_{3/2}^2$	4	+	1.7695E + 03	1.7290E + 03	1.7037E + 03	1.6165E + 03	1.5621E + 03
8	$2p_{1/2}^1 2p_{3/2}^4$	1	—	$2s_{1/2}^2, 2p_{1/2}^1, 2p_{3/2}^3$	4	+	1.7486E + 03	1.7081E + 03	1.6828E + 03	1.5956E + 03	1.5412E + 03
9	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2, 2p_{3/2}^4$	0	+	1.7959E + 03	1.7554E + 03	1.7301E + 03	1.6429E + 03	1.5885E + 03
10	$2p_{1/2}^1 2p_{3/2}^4$	1	—	$2s_{1/2}^2, 2p_{3/2}^4$	0	+	1.7750E + 03	1.7354E + 03	1.7091E + 03	1.6220E + 03	1.5676E + 03
Zn											
S. No	Upper level	2 J	P	Lower level	2 J	P	1E + 22	5E + 22	1E + 23	5E + 23	1E + 24
1	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2, 2p_{1/2}^2, 2p_{3/2}^2$	4	+	1.8992E + 03	1.8574E + 03	1.8313E + 03	1.7414E + 03	1.6852E + 03
2	$2p_{1/2}^1 2p_{3/2}^4$	1	—	$2s_{1/2}^2, 2p_{1/2}^2, 2p_{3/2}^2$	4	+	1.8748E + 03	1.8330E + 03	1.8069E + 03	1.7170E + 03	1.6608E + 03
3	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2, 2p_{1/2}^2, 2p_{3/2}^2$	0	+	1.9125E + 03	1.8707E + 03	1.8446E + 03	1.7546E + 03	1.6985E + 03
4	$2p_{1/2}^1 2p_{3/2}^4$	1	—	$2s_{1/2}^2, 2p_{1/2}^2, 2p_{3/2}^2$	0	+	1.8881E + 03	1.8463E + 03	1.8202E + 03	1.7303E + 03	1.6741E + 03
5	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2, 2p_{1/2}^1, 2p_{3/2}^3$	2	+	1.9213E + 03	1.8795E + 03	1.8534E + 03	1.7635E + 03	1.7073E + 03

Table 3 continued

Zn											
S. No	Upper level	2 J	P	Lower level	2 J	P	1E + 22	5E + 22	1E + 23	5E + 23	1E + 24
6	$2p_{1/2}^1 2p_{3/2}^4$	1	—	$2s_{1/2}^2 2p_{1/2}^1 2p_{3/2}^3$	2	+	1.8969E + 03	1.8551E + 03	1.8290E + 03	1.7391E + 03	1.6829E + 03
7	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2 2p_{1/2}^1 2p_{3/2}^3$	4	+	1.9323E + 03	1.8905E + 03	1.8643E + 03	1.7744E + 03	1.7183E + 03
8	$2p_{1/2}^1 2p_{3/2}^4$	1	—	$2s_{1/2}^2 2p_{1/2}^1 2p_{3/2}^3$	4	+	1.9079E + 03	1.8661E + 03	1.8400E + 03	1.7500E + 03	1.6939E + 03
9	$2p_{1/2}^2 2p_{3/2}^3$	3	—	$2s_{1/2}^2 2p_{3/2}^4$	0	+	1.9621E + 03	1.9203E + 03	1.8941E + 03	1.8042E + 03	1.7480E + 03
10	$2p_{1/2}^1 2p_{3/2}^4$	1	—	$2s_{1/2}^2 2p_{3/2}^3$	0	+	1.9377E + 03	1.8959E + 03	1.8697E + 03	1.7798E + 03	1.7236E + 03

reduces the effect of electron screening and quantum confinement, thereby decreasing plasma environment effect. So, this also implies that for highly charged O-like ions, strongly coupled plasma environment effect will almost negligible or may be insignificant. So it shows that O-like ions of astrophysical important elements such as iron period elements are most affected while elements useful in fusion plasma have no larger impact in the presence of strongly coupled plasma environment. In Table 2, we have compared transition wavelengths at electron density 10^{21} cm^{-3} calculated using ion sphere model in FAC with theoretically calculated and experimentally observed wavelengths by May et al. [77]. They have determined wavelength of O-like Fe from spectra recorded from Hercules laser at ENEA. The difference of FAC and HULLAC [87] results from experimental results is tabulated under the column ΔE_1 and ΔE_2 . We can see that our results are closer to experimental results but ΔE_1 and ΔE_2 are not so large and this implies that difference in transition wavelengths for strongly coupled plasma using ISM in FAC and for weakly coupled plasma using HULLAC at electron density 10^{21} cm^{-3} is very small. This shows that both results from both FAC and HULLAC are reliable and authentic.

3.2. Photoionization

In present work, we have studied photoionization of ground and first excited states of F-like ions from Fe to Zn. In Table 3, we have presented shift in ionization potentials for photoionization to lowest 5 states of O-like ions under the influence of plasma environment at different electron

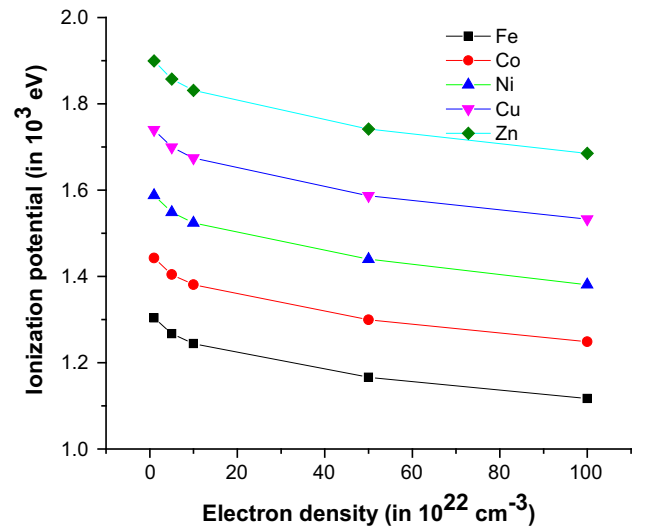
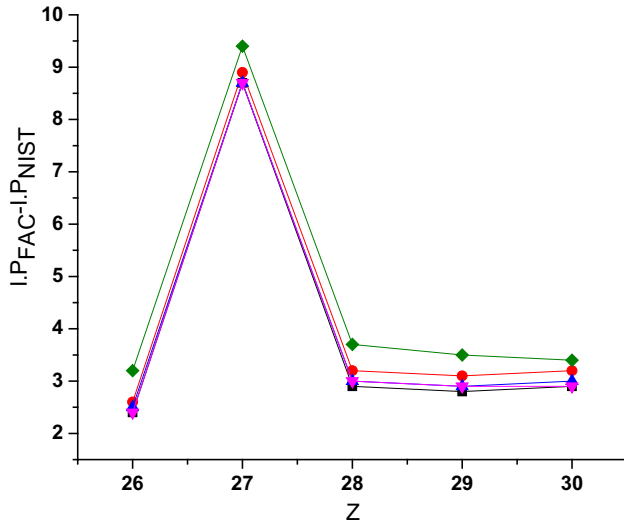
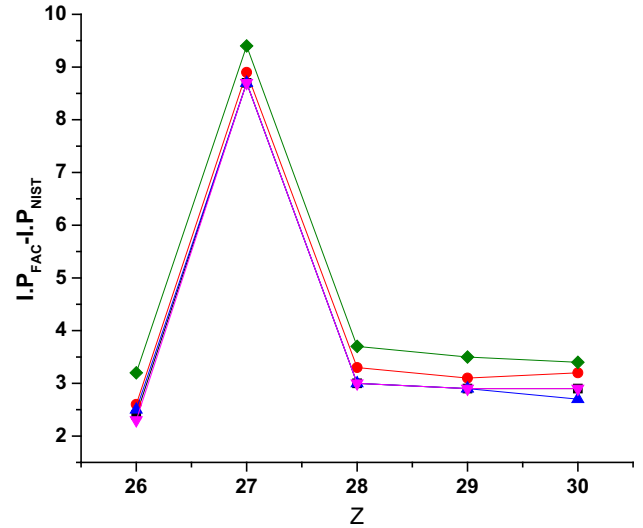


Fig. 2 Variation of shift of ionization potential for F-like ions as a function of electron density under plasma environment

Table 4 Comparison of ionization potential (in eV) without plasma for photoionization of ground and first excited states of F-like ions to lowest 5 levels of O-like ions with electron density (in cm^{-3}). a-NIST, b-[96]

S. No	Fe	Fe _{NIST}	Co	Co _{NIST}	Ni	Ni _{NIST}	Cu	Cu _{NIST}	Zn	Zn _{NIST}
1	1355.4	1357.8 ^a 1355.3 ^b	1495.8	1504.5 ^a 1495.7 ^b	1643.1	1646 ^a 1643 ^b	1797.2	1800 ^a 1797.1 ^b	1958.1	1961 ^a 1958 ^b
2	1342.7	1345.1	1480.7	1489.4	1625.2	1628.2	1776.2	1779.1	1933.7	1936.6
3	1364.5	1367.1	1506.0	1514.9	1654.3	1657.5	1809.4	1812.5	1971.4	1974.6
4	1351.8	1354.4	1490.9	1499.8	1636.4	1639.7	1788.5	1791.6	1947.0	1950.2
5	1366.4	1368.9	1509.1	1517.8	1658.9	1661.9	1815.9	1818.8	1980.2	1983.2
6	1353.7	1356.2	1494.0	1502.7	1641.1	1644.1	1795.0	1797.9	1955.8	1958.5
7	1376.3	1378.7	1519.2	1527.9	1669.3	1672.3	1826.6	1829.5	1991.2	1994.1
8	1363.6	1365.9	1504.1	1512.8	1651.5	1654.5	1805.7	1808.6	1966.8	1969.7
9	1394.9	1398.1	1540.1	1549.5	1692.7	1696.4	1853.0	1856.5	2021.0	2024.4
10	1382.2	1385.4	1525.0	1534.4	1674.9	1678.6	1832.1	1835.6	1996.6	2000.0

**Fig. 3** Ionization potential difference with NIST as a function of atomic number for F-like ions for ground state**Fig. 4** Ionization potential difference with NIST as a function of atomic number for F-like ions for first excited state

densities. In Table 3, ground and first excited states of F-like ions are under the column “upper level” and lowest five states of O-like ions are given in the column “lower level.” 2 J and P represent the double of total angular momentum and parity resp. of the corresponding state. In last five columns, ionization potential is listed for different electron density.

From Fig. 2 and Table 3, we observe that the shift in ionization potential under strongly coupled plasma effect decreases monotonically with increase in electron density for ground state photoionization of F-like ions to ground state of O-like ions because rise in electron density increases instability of the system [88]. This lowering of ionization potential is known as ionization potential depression (IPD) and it can be useful as it affects radiative properties as well as populations of states. The measurable results and predictions of this IPD effect are very important

for analysis, understanding and modeling of various atomic processes occurring under plasma environment for the study of planetary science, shock experiments and warm dense matter [89–95]. From Table 3, it also evident that the shift in ionization potential for ground state is large as compared to that for first excited state for photoionization to same state of O-like ion. In Table 4, we have compared our calculated ionization potential for F-like ions for ground and first excited state with NIST and for ground state with E. BiEmont et al. [96]. “S.No.” in Table 4 is same as in Table 3. We have found some discrepancy with NIST and maximum for F-like Co while in good agreement with E. BiEmont et al. [96]. We have plotted the difference of our calculated ionization potential with NIST for ground and first excited state as a function of atomic number in Figs. 3 and 4, respectively.

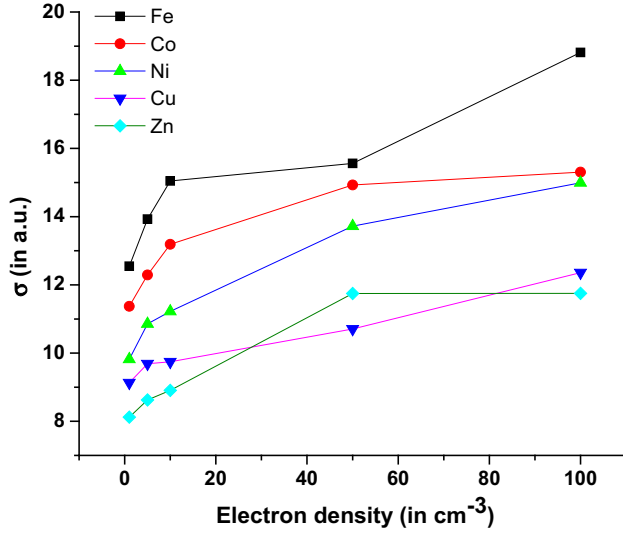


Fig. 5 Variation of photoionization cross sections of transition number 1 (given in Table 3) for F-like ions at 100 eV

We have given photoionization cross-sectional data as supplementary material in tables 5–9. In tables 5–9, we have presented photoionization cross sections for F-like ions Fe to Zn under the effect of strongly coupled plasma for different electron densities. In these tables, “transition number” is the same as “S.No.” in Table 3. From tables 5–9, we analyze that photoionization cross section decreases for all F-like ions for all transition number 1–10 when photo-electron energy increases. It is also observed that photoionization cross sections for the transition numbers 1 and 8 are very large at photo-electron energy 100 eV as compared to other transition numbers. From Fig. 5, we reveal important information that photoionization cross section decreases with increase in nuclear charge for F-like ions because continuum lowering and pressure ionization arises due to nuclear charge screening and this screening effects higher excited states, and hence, photoionization cross section decreases.

The consequence of this is that probability of photon to meet electron will be very less, and hence, large amount of ionization energy will be required to ionize. From Table 2, this can be seen that ionization potential increases with nuclear charge at a particular electron density. At electron density 10^{24} cm^{-3} , the cross section of F-like Fe becomes very large as compared to others and cross sections of F-like Co and F-like Ni, F-like Cu and F-like Zn are almost equal, while the cross section of F-like Co is very close to F-like Fe at 10^{23} cm^{-3} .

From Fig. 6, we investigate that photoionization cross sections for F-like Fe become close and close with increasing photo-electron energy or we can say that at high photo-electron energies, photoionization cross sections close to zero irrespective of the value of electron density.

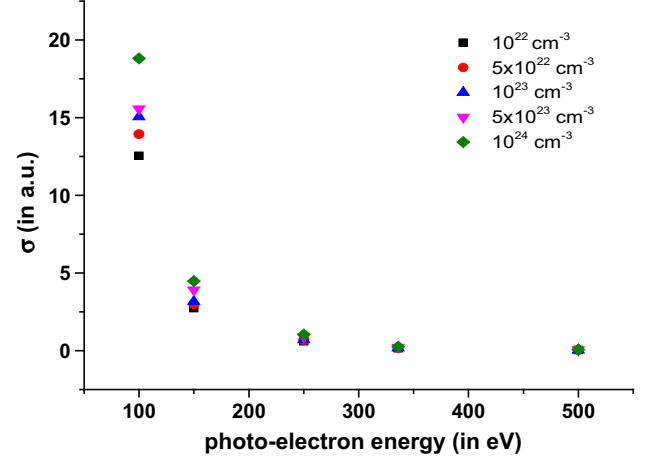


Fig. 6 Variation of photoionization cross section for transition number 1 for F-like Fe

This implies that oscillations die at high energies of photo-electron. There is also decrease in gap in photoionization cross sections with increase in energies of emitted electron. The other F-like ions follow same trend at high photo-electron energies. Further, FAC cross sections provide only background cross sections and can be scaled with available accurate cross sections to check accuracy of cross sections.

4. Conclusions

We have studied plasma embedding effects on excitation energies of O-like ions and photoionization process of F-like ions under the influence of strongly coupled plasma environment. We have implemented an ion sphere model in FAC for this purpose. The main points of the conclusion can be summarized as follows:

1. We have studied plasma effect on excitation energies of lowest 10 levels of Fe XIX to Zn XXIII at electron densities from 10^{22} cm^{-3} to 10^{24} cm^{-3} . Impact of plasma changes the position of levels due to electron screening. This impact is negligible for highly charge O-like ions. Our transition wavelengths from ISM in FAC are closer to wavelengths calculated from HULLAC and experimentally observed.
2. Photoionization of ground state and first excited state of five F-like ions, namely Fe XVIII to Zn XXII, is studied at electron densities from 10^{22} cm^{-3} to 10^{24} cm^{-3} . A significant lowering of ionization potential is predicted that can be useful from astrophysical point of view in investigation and analysis atomic processes in plasma environment.
3. Our calculation shows that photoionization cross section of Fe XVIII is very large at 10^{24} cm^{-3} and at

high values of energies of emitted electron; photoionization cross section becomes very small and does not depend on electron density. While at a particular low photo-electron energy, photoionization cross section increases with increase in electron density and decrease with nuclear charge.

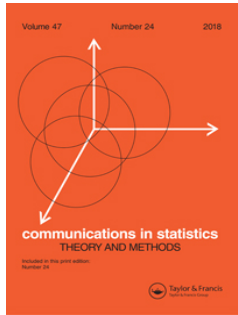
Acknowledgement We are thankful to Shyamlal College, Delhi University and Delhi Technological University for providing facilities and infrastructure for carrying out this work.

References

- [1] Y Y Qi, J G Wang, and R K Janev *Phys. Rev. A* **78** 062511 (2008)
- [2] B Saha, P K Mukherjee, D Bielinska-Waz and J Karwowski *J. Quant. Spectrosc. Radiat. Transf.* **92** 1 (2005)
- [3] C A Rouse *Phys. Rev. A* **4** 90 (1971)
- [4] C Gao, J Zeng and J Yuan *High Energy Density Phys* **7** 54 (2011)
- [5] J -Q Pang, Guo-Xing Han, Ze-Qing Wu and Shi-Chang Li *J. Phys. B* **35** 2117 (2002)
- [6] Yongqiang Li, Jianhua Wu, Yong Hou and Jianmin Yuan *J. Phys. B* **41** 145002 (2008)
- [7] F J Rogers and C A Iglesias *Science* **263** 50 (1994)
- [8] E Storm *J. Fusion Energy* **7** 131 (1988)
- [9] G Mourou and D Umstadter *Phys. Fluids. B* **4** 2315 (1992)
- [10] J Workman, M Nantel, A Maksimchuk, and D Umstadter *Appl. Phys. Lett.* **70** 312 (1997)
- [11] M M Murnane, H C Kapreyn, M D Rosen and R W Falcone *Science* **251** 531 (1991)
- [12] J Seidel, S Arndt and W D Kraeft *Phys. Rev. E* **52** 5387 (1995)
- [13] D Ray *Phys. Rev. E* **62** 4126 (2000)
- [14] T C Killian, T Pattard, T Pohl and J M Rost *Phys. Rep.* **449** 77 (2007)
- [15] B Saha and S Fritzsche *Phys. Rev. E* **73** 036405 (2006)
- [16] D Salzmann *Atomic Physics in Hot Plasma*. (Oxford: Oxford University Press) (1998)
- [17] AF Starace, Photoionization of Atoms (Atomic, Molecular, and Optical Physics Handbook, Edited by G.W.F. Drake (A.I.P., New York, 1996), Chapter 24, pp. 301–309. Copyright 1996 American Institute of Physics)
- [18] JJ Yeh and I Lindau *At. Data Nucl. Data Tables* **32** 1 (1985)
- [19] D Bielinska-Waz, J Karwowski, B Saha, and P K Mukherjee *Phys. Rev. E* **69** 016404 (2004)
- [20] B Saha, P K Mukherjee and G H F Diercksen *Astron. Astrophys.* **396** 337 (2002)
- [21] S Paul and Y K Ho *Phys. Plasmas* **16** 063302 (2009)
- [22] J K Saha, S Bhattacharyya, T K Mukherjee and P K Mukherjee *J. Phys. B* **42** 245701 (2009)
- [23] M.Das, M Das, R K Chaudhuri and S Chattopadhyay *Phys. Rev. A* **85** 042506 (2012)
- [24] U Gupta and A K Rajagopal *J. Phys. B: At. Mol. Phys.* **14** 2309 (1981)
- [25] G J Hatton, N F Lane and J C Weisheit *J. Phys. B: At. Mol. Phys.* **14** 4879 (1981)
- [26] P Martel, L Doreste, E M'inguez and J M Gil *J. Quant. Spectrosc. Radiat. Transfer* **54** 621 (1995)
- [27] J M Ugalde, C Sarasola and X Lopez *Phys. Rev. A* **56** 1642 (1997)
- [28] J M Gil, P Martel, E M'inguez, J G Rubiano, R Rodr'iguez and F H Ruano *J. Quant. Spectrosc. Radiat. Transfer* **75** 539 (2002)
- [29] I Silanes, J M Mercero and J M Ugalde *Phys. Rev. E* **66** 026408 (2002)
- [30] S Kar and Y K Ho *Phys. Rev. E* **70** 066411 (2004)
- [31] J G Rubiano, R Florido, R Rodr'iguez, J M Gil, P Martel and E M'inguez *J. Quant. Spectrosc. Radiat. Transfer* **83** 159 (2004)
- [32] R Rodr'iguez, J M Gil, J G Rubiano, R Florido, P Martel and E M'inguez *J. Quant. Spectrosc. Radiat. Transfer* **91** 393 (2005)
- [33] B Saha and S Fritzsche *J. Phys. B: At. Mol. Opt. Phys.* **40** 259 (2007)
- [34] B F Rozsnyai *Phys. Rev. A* **5** 1137 (1972)
- [35] B F Rozsnyai *J. Quant. Spectrosc. Radiat. Transfer* **27** 211 (1982)
- [36] B F Rozsnyai *Phys. Rev. A* **43** 3035 (1991)
- [37] D A Liberman *Phys. Rev. A* **20** 4981 (1979)
- [38] D Salzmann, R Y Yin and R H Pratt *Phys. Rev. A* **6** 3627 (1985)
- [39] D Salzmann and H Szichman *Phys. Rev. A* **35** 807 (1987)
- [40] H Nguyen, M Koenig, D Benredjem, M Caby and G Coulaud *Phys. Rev. A* **33** 1279 (1986)
- [41] B J B Crowley *Phys. Rev. A* **41** 2179 (1990)
- [42] M S Murillo and J C Weisheit *Phys. Rep.* **302** 1 (1998)
- [43] S Skupsky *Phys. Rev. A* **21** 1316 (1980)
- [44] J C Stewart and K D Pyatt *Astrophys. J.* **144** 1203 (1966)
- [45] S Ichimaru *Rev. Mod. Phys.* **54** 1017 (1982)
- [46] M W C Dharma-Wardana and F Perrot *Phys. Rev. A* **26** 2096 (1982)
- [47] S Sen, P Mandal, P K Mukherjee and B Fricke *Phys. Plasmas* **20** 013505 (2013)
- [48] J Basu and D Ray *Phys. Rev. E* **83** 016407 (2011)
- [49] S Bhattacharyya, A N Sil, S Fritzsche and P K Mukherjee *Eur. Phys. J. D* **46** 1 (2008)
- [50] M Das *Phys. Plasmas* **21** 012709 (2014)
- [51] X Li, Z Xu and F B Rosmej *J. Phys. B: At. Mol. Opt. Phys.* **39** 3373 (2006)
- [52] Y Li, J Wu, Y Hou and J Yuan *J. Phys. B: At. Mol. Opt. Phys.* **41** 145002 (2008)
- [53] M Das, B K Sahoo and S Pal *Phys. Rev. A* **93** 052513 (2016)
- [54] Y Y Qi, J G Wang and R K Janev *Eur. Phys. J. D* **63** 327 (2011)
- [55] Y Y Qi, J G Wang and R K Janev *Phys. Rev. A* **80** 063404 (2009)
- [56] C Y Lin and Y K Ho *Phys. Plasmas* **17** 093302 (2010)
- [57] C Y Lin and Y K Ho *Phys. Rev.* **81** 033405 (2010)
- [58] Y Y Qi, Y Wu, and J G Wang *Phys. Plasmas* **16** 033507 (2009)
- [59] S Sahoo and Y K Ho *Phys. Plasmas* **13** 063301 (2006)
- [60] Y D Jung *Phys. Plasmas* **5** 4456 (1998)
- [61] K G Dyall, I P Grant, C T Johnson, F A Parpia and E P Plummer *Comput. Phys. Commun.* **55** 425 (1989)
- [62] B C Fawcett, K J H Phillips, C Jordan and J R Lemen *MNRAS* **225** 1013 (1987)
- [63] U Feldman and G A Doschek *ApJS* **75** 925 (1991)
- [64] U Feldman, W Curdt, G A Doschek, U Schühle, K Wilhelm and P Lemaire *ApJ* **503** 467 (1998)
- [65] U Feldman, W Curdt, E Landi and K Wilhelm *ApJ* **544** 508 (2000)
- [66] E Landi, M Landini and G Del Zanna *A&A* **324** 1027 (1997)
- [67] E Behar, J Cottam and S M Kahn *ApJ* **548** 966 (2001)
- [68] K P Dere, E Landi, P R Young and G Del Zanna *ApJS* **134** 331 (2001)
- [69] R Mewe, A J J Raassen, J J Drake, J S Kaastra, R L J van der Meer and D Porquet *A&A* **368** 888 (2001)
- [70] J S Kaastra et al. *A&A* **386** 427 (2002)
- [71] W Curdt, E Landi, and U Feldman *A&A* **427** 1045 (2004)
- [72] E Landi and K J H Phillips *ApJS* **160** 286 (2005)
- [73] E Landi and K J H Phillips *ApJS* **166** 421 (2006)
- [74] A J J Raassen and A M T Pollock *A&A* **550** A55 (2013)
- [75] G V Brown, P Beiersdorfer, D A Liedahl, K Widmann, S M Kahn and E J Clothiaux *ApJS* **140** 589 (2002)

-
- [76] K B Fournier et al. *J. Phys. B* **36** 3787 (2003)
- [77] M J May et al. *ApJS* **158** 230 (2005)
- [78] H Chen, M F Gu, E Behar, G V Brown, S M Kahn, and P Beiersdorfer *ApJS* **168** 319 (2007)
- [79] E Träbert, P Beiersdorfer, N S Brickhouse and L Golub *ApJS* **211** 14 (2014)
- [80] Z H Yang, S B Du, M G Su, H W Chang and Y P Guo *AJ* **152** 135 (2016)
- [81] D J Fields, S Mathur, Y Krongold, R Williams and F Nicastro *ApJ* **666** 828 (2007)
- [82] J García, C Mendoza, MA Bautista, TW Gorczyca, TR Kallman and Palmeri *ApJS* **158** 68 (2005)
- [83] J Deprince, M A Bautista, S Fritzsche, J A García, T R Kallman, C Mendoza et al *A&A* **624** A74 (2019)
- [84] F Y Khattak, O A M B Percie du Sert, F B Rosmej and D Riley *J. Phys. Conf. Ser.* **397** 012020 (2012)
- [85] M Belkhiri, C J Fontes and M Poirier *Phys. Rev. A* **92** 032501 (2015)
- [86] M F Gu *Can. J. Phys.* **86** 675 (2008)
- [87] A Bar-Shalom and M Klapisch *J. Oreg JQSRT* **71** 169 (2001)
- [88] Y Ben-Aryeh *Scientific Reports* **9** 20384 (2019)
- [89] *High Energy Density Phys.* **8** 1 (2012)
- [90] S H Glenzer and R Redmer *Rev. Mod. Phys.* **81** 1625 (2009)
- [91] R P Drake *Phys. Plasmas* **16** 055501 (2009)
- [92] H J Lee et al. *Phys. Rev. Lett.* **102** 115001 (2009)
- [93] E García Saiz et al. *Phys. Rev. Lett.* **101** 075003 (2008)
- [94] M D Knudson et al. *Phys. Rev. Lett.* **108** 091102 (2012)
- [95] N Nettelmann, R Redmer and D Blaschke *Phys. Part. Nucl.* **39** 1122 (2008)
- [96] E BiÉmont, Y FrÉmat and P Quinet *At. Data Nucl. Data Tables* **71** 117 (1999)

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



Testing normality in the time series of EMP indices: an application and power-comparison of alternative tests

Sanjay Kumar & Nand Kumar

To cite this article: Sanjay Kumar & Nand Kumar (2021): Testing normality in the time series of EMP indices: an application and power-comparison of alternative tests, Communications in Statistics - Theory and Methods, DOI: [10.1080/03610926.2021.1914097](https://doi.org/10.1080/03610926.2021.1914097)

To link to this article: <https://doi.org/10.1080/03610926.2021.1914097>



Published online: 06 May 2021.



Submit your article to this journal [↗](#)



Article views: 46



View related articles [↗](#)



View Crossmark data [↗](#)



Testing normality in the time series of EMP indices: an application and power-comparison of alternative tests

Sanjay Kumar^{a,b}  and Nand Kumar^c

^aDelhi Technological University, New Delhi, India; ^bDepartment of Economics, Dyal Singh College, Delhi University, New Delhi, India; ^cDepartment of Humanities, Delhi Technological University, New Delhi, India

ABSTRACT

The Exchange Market Pressure Index (EMPI) is an indicator of pressure on a currency. Because of the presence of serial correlation, financial time series may not be normally distributed even for large sample sizes. They may have undefined parameters and hence parametric tests of normality may give misleading results. In this paper, we look at the time series of EMPI of eleven countries of the world, put the data to normality check using tests suggested by various scholars. We also apply a test used exclusively for serially correlated data. No one has used this test earlier. In this context, we also compare the power of these statistical tests, which is another novel contribution of this paper. On the basis of these tests the EMPI time series is found to be non-normal. Two tests are found to be the most powerful. The test which is designed exclusively for time series data is found to be powerful only for China and South Korea, the countries which had the lowest EMPI- standard- deviation in the group of all the eleven countries studied in this paper.

ARTICLE HISTORY

Received 3 August 2020
Accepted 4 April 2021

KEYWORDS

Skewness; kurtosis;
normality; EMPI

JEL

C01; C46; C58; C22

1. Introduction

An EMPI series like any other time series, tends to be serially correlated i.e., dependent thereby belying the basic assumption, of the most of statistical analysis, that observations are independent and identically distributed. One consequence of all this is that statistical inferences tend to be misleading because unless a series has a tendency to cluster round some central values it is not possible to characterize it in terms of a few numbers called the moments of the series. Take an EMPI time series $(EMPI_t)_{t=1}^T$. Let $\mu = \text{Population mean} = \text{Expected value of EMPI} = E(EMPI_t) = \sum_1^T EMPI_t f(EMPI_t)$. Let $\mu_r = E[(EMPI_t - \mu)^r]$ be its r th central moment around the mean and Population Variance $= \text{Var}(EMPI) = \sigma^2 = E(EMPI_t - \mu)^2$. Then coefficients of standardized skewness and excess kurtosis will be given by-

$$\tau = \frac{\mu^3}{\sigma^3} = \frac{E[(EMPI_t - \mu)]^3}{E[(EMPI_t - \mu)^2]^{\frac{3}{2}}} \quad (1)$$

$$\kappa - 3 = \frac{\mu^4}{\sigma^4} = \frac{E[(EMPI_t - \mu)]^4}{E[(EMPI_t - \mu)^2]^2} \quad (2)$$

Take t , orthogonal sample observations on $EMPI_t$, such that $f_{EMPI_t}(EMPI_t) = f_{EMPI_{t-1}}(EMPI_{t-1}) \forall EMPI_t \in \mathbb{R}$ and $f_{EMPI_t}EMPI_{t-1}(EMPI_t EMPI_{t-1}) = f_{EMPI_t}(EMPI_t) f_{EMPI_{t-1}}(EMPI_{t-1}) \forall EMPI_t, EMPI_{t-1} \in \mathbb{R}$ then and $\sqrt{t}\hat{\tau} \rightarrow N(0, 6)$ and $\sqrt{t}(\hat{\kappa} - 3) \rightarrow N(0, 24)$

A parametric test of normality, for t independent observations on a random variable, using, Lagrange multiplier(LM), is suggested by Jarque and Bera (1987) as below:

$$LM/JB = t \left[\frac{\hat{\tau}^2}{6} + \frac{(\hat{\kappa} - 3)^2}{24} \right] \quad (3)$$

Here, $LM/JB_{asy} \sim \chi^2(2)$. For a normally distributed sample the value of LM/JB statistic is zero. Other tests of normality are given by Shapiro and Wilk (1965), Lilliefors (1967), Doornik and Hansen (2008), D'Agostino and Stephens (1986) etc. The uni-variate omnibus test of normality propounded by Doornik and Hansen (2008) can be presented like this let $\{EMPI_1, EMPI_2, EMPI_3, \dots, EMPI_n\}$ be a sample of n independent observations on EMPI. Then,

$$E_p = Z_1^2 + Z_2^2 \sim \chi^2(2) \quad (4)$$

Here Z_1 and Z_2 are skewness and kurtosis of the EMPI sample respectively. Shapiro - Wilk(ibid) have given the W test for normality. The test can be stated for n ordered random samples of EMPI such as $\{EMPI_1 \leq EMPI_2 \leq EMPI_3 \leq \dots \leq EMPI_n\}$. Then,

$$S^2 = \sum_{i=1}^n (EMPI_i - \overline{EMPI})$$

When n is even i.e., $n = 2k$.

$$b = \sum_{i=1}^k a_{n-i+1} (EMPI_{n-i+1} - EMPI_i)$$

When n is odd i.e., $n = 2k + 1$.

$$b = a_n(EMPI_n - EMPI) + \dots + a_{k+2}(EMPI_{k+2} - EMPI_k)$$

The value of the coefficient a in the above two equations is got from the table compiled by Shapiro and Wilk (1965). The test statistics W is given as-

$$W = \frac{b^2}{S^2} \quad (5)$$

Lilliefors (1967) test statistics T is given as

$$T = \sup |F(x) - Z(x)| \quad (6)$$

Here,

$F^*(x)$ = Actual cumulative distribution of the normalized data at each x .

$Z(x)$ = Theoretical cumulative distribution of the data from Z table at each x .

We reject the null hypothesis that the data has a normal distribution if the T value is higher than the same in the table provided by Lilliefors (1967) at an α of 1 percent in two tail test. The power of all these tests will be drastically reduced if we use these tests to check the normality of experimental data where trials are finite and are not independent. This may happen particularly in time series data because of the problem of auto-correlation. Economic time series has a heavy tail look i.e., the moments, in some cases, tend not to converge toward a central value as they tend to do in case of trials involving cross section data or even if they converge they converge at very large data. Thus, we have a problem of size distortion in the data. If probabilities associated with individual trials are conditional (showing reduction in degrees of freedom) i.e., multiplicative in nature then we are almost certain to get a heavy tail distribution. This ultimately, may lead to some or all the moments of the distribution being undefined. Even if the moments are defined using the rules of moments to measure them may underestimate their true value and this underestimation may be higher as we move toward higher moments. Some scholars have suggested that Asymptotic theory of sample extremes i.e., Extreme Value Theory(EVT)¹ and not the central limit theory is suitable to examine data with heavy tails this is because probability of extreme values is higher in the case of the financial time series as opposed to a series with a concentration toward the middle values. Garita and Zhou (2009) noted that assuming normal distribution and estimating frequencies on the basis of sample mean and standard deviation may underestimate the frequency of extreme values of the observations. Mandelbrot (1997) has pointed out that for fat tailed distribution the second moment is undefined and thus estimating it for finite samples may severely affect the predictive power of the model. Vector Autoregressive (VAR) data generating processes for EMPI is used by Tanner (2002), Gochoco-Bautista and Bautista (2005), Kamaly and Erbil (2000) to normalize the series Bai and Ng (2005) have shown that, in the case of time series, measuring the tails using kurtosis statistic is not a sound approach as the true value of κ is likely to be substantially underestimated in practice. In the case of normal distribution, the limiting value of standardized third and fourth moment are 0 and 3 in and their variances are 6 and 24 respectively. In the case of serially correlated data these limiting variances are no longer 6 and 24 but some function of long run variance co-variance matrix; hence, assuming that the data conforms to a normal distribution asymptotically and then testing the normality of a small sample in this framework may not be a proper approach. To properly estimate higher moments in the case of such data we require auto-correlation consistent estimators and literature suggests the use of bootstrap method or kernel density functions for the same with alternative bandwidth parameters. Bai and Ng (2005) suggest a normality test for such data with Bartlett, Parzen and Quadratic Spectral kernel density functions with Newley West and Andrews White and any other pre-chosen bandwidths. These kernel functions are not dependent on value of the moments and hence the problem of moment dependence of a probability distribution function is sorted out. Bai and Ng (2005) test is still a moment based test of normality but they derive equations for auto-correlation consistent estimators for skewness and kurtosis. Hence, this test is an improvement over other tests.

Let EMP_t be weakly stationary time series, then

$$\hat{\pi}_{34} = \hat{\pi}_3^2 + \hat{\pi}_4^2 \rightarrow \chi^2 \quad (7)$$

Here,

$$\hat{\pi}_3^2 \rightarrow \chi^2; \hat{\pi}_4^2 \rightarrow \chi^2$$

We have,

$$\hat{\pi}_3 = \frac{\sqrt{T}\hat{\mu}_3}{s(\hat{\mu}_3)} = \frac{\sqrt{T}\hat{\tau}}{s(\hat{\tau})} \rightarrow N(0, 1) \quad (8)$$

$$\hat{\pi}_4 = \frac{\sqrt{T} \wedge (k - k)}{s(\hat{k})} \rightarrow N(0, 1) \quad (9)$$

In this paper we estimate the test statistics $\hat{\pi}_{34}, \hat{\pi}_3, \hat{\pi}_4$ as suggested by Bai and Ng (2005) and test the normality of the GR EMP index of eleven countries including Australia, Brazil, Canada, China, India, Japan, South Africa, South Korea, Switzerland, United Kingdom and Venezuela. We also test the normality of the EMPI time series of these countries using the tests suggested by Jarque and Bera (1987), Shapiro and Wilk (1965), Lilliefors (1967) and Doornik and Hansen (2008). Bai and Ng (2005) have estimated these statistics for various macro-economic time series data for some countries but no one has estimated and used Bai and Ng statistics in the context of EMPI time series of any country. We first do a stationarity test for the data and thereafter apply the normality test to check whether the EMPI data for these countries have a normal distribution for finite sample. We also present a comparison of the power of these tests for finite EMP time series.

2. The GR exchange market pressure index

The EMP Index is estimated by using the formula:

$$EMP_t = \Delta f_t + \Delta s_t \quad (10)$$

Here, as per Girton and Roper (1977)

EMP_t = Exchange Market Pressure at time t.

Δs_t = First difference of \ln of exchange rate (i.e., percentage change in nominal exchange rate with respect to previous period exchange rate).

Δf_t = First difference of \ln of foreign exchange reserve (i.e., Percentage Change in forex reserve with respect to previous period monetary base).²

3. Data sources

For construction of EMPI we require quarterly data for exchange rate, foreign exchange reserve and monetary base for all countries starting 1992Q1 end 2018Q4. The first and the last quarter data are lost in calculating the first differences. Thus we had original data starting from 1992Q1 to 2018Q4 i.e., 108 observations but finally we have only 106 observations starting 1992Q2 to 2018Q3. We construct the three sets of EMP indices for all the eleven countries. The first EMPI is based on market exchange rate and we call it

EMPIMER, the second EMPI is based on nominal effective exchange rate (NEER) and we call it EMPINEER, the third EMPI is based on real effective exchange rate (REER) and we call it EMPIREER. The data for nominal market exchange rate of INR with USD is taken from Organization for Economic Co-operation and Development (OECD) and the same for other countries is taken from Board of Governors of the Federal Reserve System (FRED) database. Theoretically speaking it is better to gauge the strength or weakness of a currency in terms of effective exchange rate (EER) and not in terms of market exchange rate. There are two indicators of effective exchange rate - NEER and REER. NEER (nominal effective exchange rate) is an export or trade weighted average of bilateral exchange rate and REER (real effective exchange rate) is NEER adjusted by some measure of relative price or cost. In the case of India NEER and REER data using a basket of 36 countries bilateral rates are made available by RBI 2005-06 onwards but the data is available on annual basis only. This data can be accessed from table 143 and table 144 of the Handbook of Statistics on Indian Economy. However, we have used the data of NEER and REER indices provided by the Bank of International Settlement (BIS) for compatibility reasons. The BIS, EER indices currently cover 61 economies (including individual Euro area countries and, separately, the Euro area as an entity). Nominal EERs are calculated as geometric weighted averages of bilateral exchange rates. The weights are derived from manufacturing trade flows and capture both direct bilateral trade and third-market competition by double weighting. Real EERs are the weighted averages of bilateral exchange rates adjusted by relative consumer prices in the two countries. All real effective exchange rate series, except the one for Venezuela, have 2015 (=100) as base year and are estimated using manufacturing consumer price indices. The data is made available at OECD and is sourced from FRED online database. Venezuela's real effective exchange rate series has 2010 (=100) as base year, is made available at BIS and taken from FRED online database. Both broad and narrow nominal and real effective exchange rates are available. We have used the broad indices which are based on 60 countries trade basket. The data for foreign exchange reserve are taken from the 'Survey Based on Standardized Form by Country'. This data is accessed from International Financial Statistics database of International Monetary Fund. The data is available in the head 'International Liquidity Selected Indicators'. For our purpose we have taken the subhead which is the second row of the table and is titled: 'International Reserves, Official Reserve Assets, SDRs, US Dollars'. This data is in Millions of USD. Data about monetary base of India is taken from the 'Database of India Economy' (DBIE).

DBIE is a time series publication of the Reserve Bank of India. Monetary base data are available in the head 'Reserve Money Components and Sources'. The data is available from 6th July 2001 to 25th January 2019 on weekly basis. We have, for our purpose, taken the data of the dates matching first, second, third and fourth quarter of every year. We have backcast this quarterly series to generate data for the quarters prior to 2001 as our data starts from the first quarter of 1992. The back series is generated by using trend growth. Monetary Base of India = Reserve Money = Currency in circulation + Bank's deposits with RBI + Other deposits with RBI. The data about the monetary base of Japan and Brazil is accessed from the Bank of Japan and Brazil Central Bank respectively, all other countries monetary base data are accessed from International Financial Statistic, International Monetary Fund (IMF).

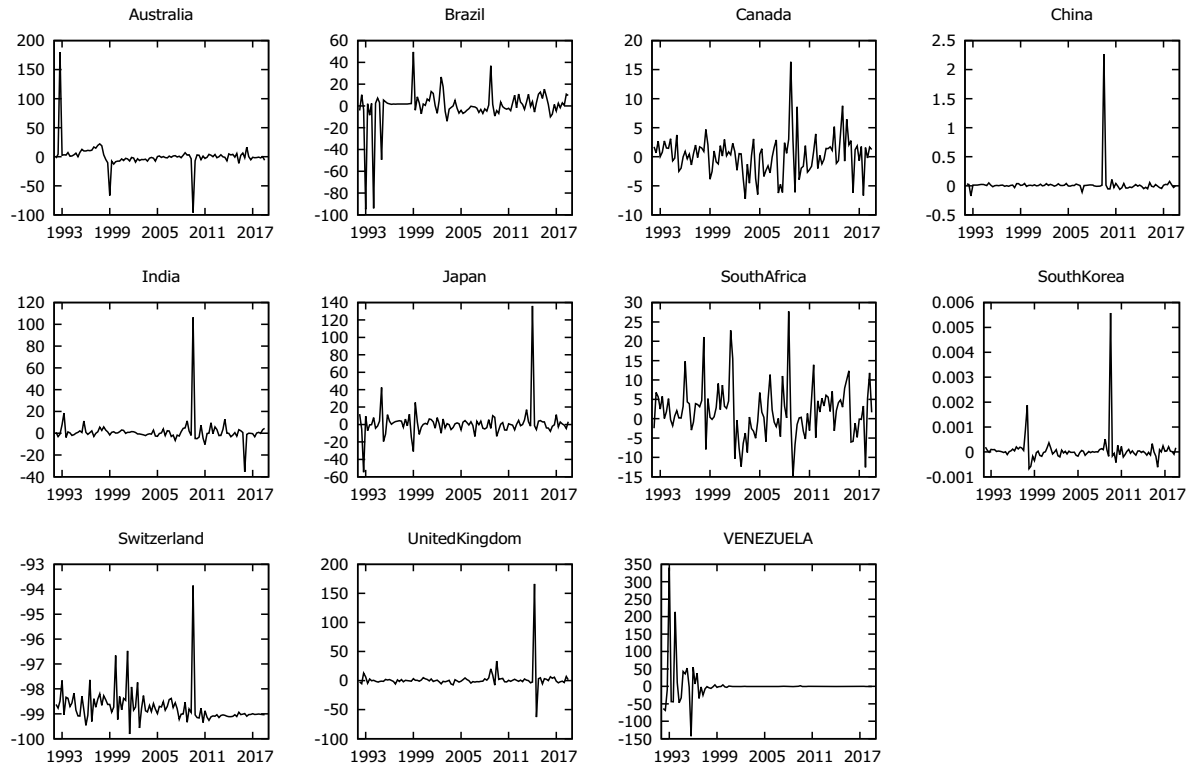


Figure 1. EMPIMER selected countries.

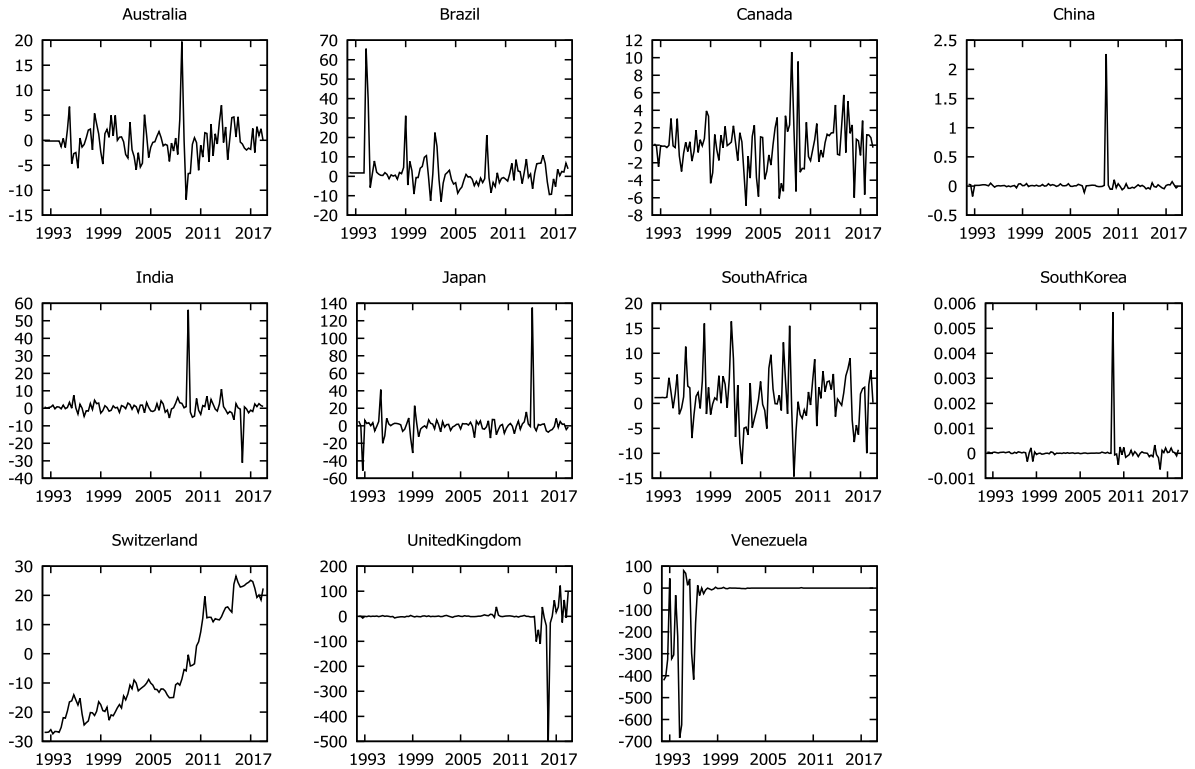


Figure 2. EMPINEER selected countries.

4. Stationarity test: GR EMP index

Before we launch the formal test of normality it is important to put the data to stationarity check. Let $EMPI_t$ be a random time series process which is weakly stationary then $\mu_{EMPI_t} = \mu_{EMPI_{t+v}} \forall v \in \mathbb{R}$, and $\sigma_{EMPI_t}^2 = \sigma_{EMPI_{t+v}}^2 \forall v \in \mathbb{R}$

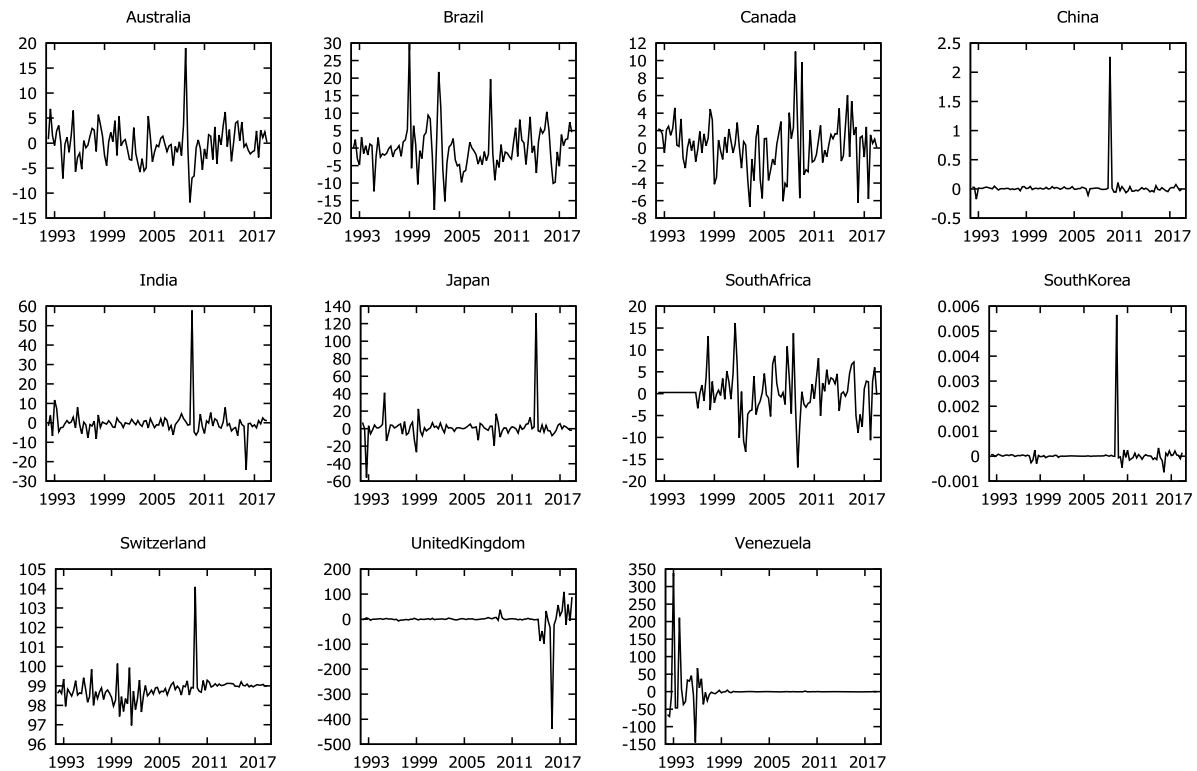


Figure 3. EMPIREER selected countries.

In Figures 1–3 we have presented the time series of charts of the EMPIMER, EMPINEER and EMPIREER based on market exchange rate, nominal exchange rate and real exchange rate resp for all the eleven countries of our interest. We call them EMPIMER (Exchange Market Pressure Index based on Market Exchange Rate), EMPINEER (Exchange Market Pressure Index based on Nominal Exchange Rate) and EMPIREER (Exchange Market Pressure Index based on Real Exchange Rate). All these charts appear to be stationary with respect to time except the EMPINEER chart of Switzerland (see chart no.9, Figure 2). We have also put the data to Dicky-Fuller unit root test to examine the stationarity issue formally. Since graphically the data is trendless and there is no intercept in all the data we have run the unit root test with model without constant and trend. The hypotheses on test are – H_0 Unit root $\rho = 1$ $H_1\rho < 1$. The results are presented in Tables 1–3. The p value of the coefficient of $y(-1)$, in case of all the EMPI for all the countries, is lower than α of 0.01 for one tail hypothesis test, hence we reject the null hypothesis that the data has unit root. Hence all the data about the EMPIMER, EMPINEER, EMPIREER of all the countries are stationary. There is only one country which is exception to this general conclusion. Here we find that the EMPIMER, EMPINEER as well as EMPIREER series for Switzerland are non stationary because in all the three cases the p values are higher than the critical α of 0.01 for one tail test. So only in this exceptional case of Switzerland there is unit root problem.

5. Empirical application of normality tests to EMPI data: moment based tests for IID data

In this section, we put the EMPI data to normality test through different test statistics. We use Doornik and Hansen (2008), Shapiro and Wilk (1965), Lilliefors (1967), Jarque

Table 1. Dicky Fuller test EMPIMER.

Test without constant, model: $EMP_t = \rho EMP_{t-1} + e$, H_0 Unit root $\rho = 1$, $H_1 \rho < 1$, $T = 105, 1992:2-1918:3$						
Country	ρ (Coefficient $y(-1)$)	Std. error	t-ratio	p-value	H_0	Conclusion
Australia	-0.873727	0.0973019	-8.980	3.91e - 031	Reject	Stationary
Brazil	-0.525114	0.136453	-3.848	0.0001	Reject	Stationary
Canada	-0.901953	0.09775271	-9.248	3.65e - 035	Reject	Stationary
China	-0.990664	0.0980450	-10.10	3.43e - 052	Reject	Stationary
India	-1.02379	0.0980973	-10.44	6.08e - 061	Reject	Stationary
Japan	-1.06118	0.0976423	-10.87	1.44e - 074	Reject	Stationary
South Africa	-0.76987	0.0953492	-8.054	9.66e - 021	Reject	Stationary
South Korea	-1.01790	0.0980307	-10.38	1.88e - 059	Reject	Stationary
Switzerland	0.000179354	0.000740940	0.2421	0.7565	Do not Reject	Non stationary
United Kingdom	-1.31839	0.0929508	-14.18	0.0000	Reject	Stationary
Venezuela	-1.49655	.136582	-10.96	4.80e - 022	Reject	Stationary

Table 2. Dicky Fuller test EMPINEER.

Test without constant, model: $EMP_t = \rho EMP_{t-1} + e$, H_0 Unit root $\rho = 1$, $H_1 \rho < 1$, $T = 105, 1992:2-1918:3$						
Country	ρ (Coefficient $y(-1)$)	Std. error	t-ratio	p-value	H_0	Conclusion
Australia	-0.781768	0.0956943	-8.169	8.69e - 022	Reject	Stationary
Brazil	-0.765625	0.110649	-6.919	2.81e - 011	Reject	Stationary
Canada	-0.969185	0.0980146	-9.888	3.33e - 047	Reject	Stationary
China	-0.990130	0.0980447	-10.10	4.65e - 052	Reject	Stationary
India	-0.991499	0.0980586	-10.11	2.30e - 052	Reject	Stationary
Japan	-1.0545	0.0978655	-10.77	2.24e - 071	Reject	Stationary
South Africa	-0.762229	0.0952249	-8.005	2.63e - 020	Reject	Stationary
South Korea	-1.01740	0.0980696	-10.37	3.37e - 059	Reject	Stationary
Switzerland	-0.0132022	0.0136503	-0.9672	0.2962	Do not Reject	Non stationary
United Kingdom	-0.713755	0.179071	-3.986	6.86e-05	Not Reject	Stationary
Venezuela	-0.408150	0.0730013	-5.591	4.43e - 08	Reject	Stationary

Table 3. Dicky Fuller test EMPIREER.

Test without constant, model: $EMP_t = \rho EMP_{t-1} + e$, H_0 Unit root $\rho = 1$, $H_1 \rho < 1$, $T = 105, 1992:2-1918:3$						
Country	ρ (Coefficient $y(-1)$)	Std.error	t-ratio	p-value	H_0	Conclusion
Australia	-0.804598	0.09614760	-8.368	9.59e-024	Reject	Stationary
Brazil	-0.906534	0.175932	-7.430	1.44e - 012	Reject	Stationary
Canada	-0.953507	0.122012	-9.760	1.79e-044	Reject	Stationary
China	-0.990196	0.0980449	-10.10	4.48e-052	Reject	Stationary
India	-1.06098	0.0978635	-10.84	1.20e-073	Reject	Stationary
Japan	-1.04654	0.0978639	-10.69	9.78e-069	Reject	Stationary
South Africa	-0.7966639	0.0960089	-8.298	5.02e-023	Reject	Stationary
South Korea	-1.01739	0.0980680	-10.37	3.36e-059	Reject	Stationary
Switzerland	0.000125343	0.000726363	-0.1726	0.7364	Do not Reject	Non stationary
United Kingdom	-0.713981	0.179359	-3.981	7.01e-05	Reject	Stationary
Venezuela	-0.865825	0.175932	-4.921	1.09e-06	Reject	Stationary

and Bera (1987) tests. These tests are based on the assumption of IID data. We use the sample as well as the Monte Carlo p values for hypothesis testing. The hypotheses of the test are – H_0 = Data is normally distributed, H_1 =Data is not normally distributed. We test these hypotheses at α of 0.05 in one tail test. The test statistics have χ^2 distribution. In case of all the tests the reported p values are less than the critical α of 0.05. Hence, we reject the null hypothesis that the EMPIMER, EMPINEER, and EMPIREER data are normally distributed for these ten countries. The reported Monte-Carlo p values in all the cases however are higher than the critical α of 0.05 hence on the basis of Monte-Carlo p values we are not in a position to reject the null that the EMPIMER,

Table 4. Normality test EMPIMER.

Country	Doornik-Hansen test	Shapiro-Wilk test	Lilliefors test	Jarque-Bera test
Australia	85.1969 $p = 3.16031\text{e-}019$	0.425942 $p = 1.72768\text{e-}018$	0.28341 $p = 0$	8954.97 $p = 0$
Brazil	124.063 $p = 1.14836\text{e-}027$	0.587384 $p = 8.56363\text{e-}016$	0.264823 $p = 0$	1923.39 $p = 0$
Canada	33.6483 $p = 4.93584\text{e-}008$	0.91981 $p = 7.97954\text{e-}006$	0.11271 $p = 0$	100.036 $p = 1.89411\text{e-}022$
China	8178.37 $p = 0$	0.159372 $p = 6.8549\text{e-}022$	0.421724 $p = 0$	42720.2 $p = 0$
India	1067.32 $p = 1.71372\text{e-}232$	0.373338 $p = 3.04053\text{e-}019$	0.28541 $p = 0$	18618.2 $p = 0$
Japan	227.712 $p = 3.57245\text{e-}050$	0.525068 $p = 6.50019\text{e-}017$	0.252391 $p = 0$	8327.31 $p = 0$
South Africa	13.6484 $p = 0.00108716$	0.949482 $p = 0.000506799$	0.11278 $p = 0$	32.84 $p = 7.39392\text{e-}008$
South Korea	2401.46 $p = 0$	0.335713 $p = 9.37846\text{e-}020$	20124.6 $p = 0$	32.84 $p = 7.39392\text{e-}008$
United Kingdom	1005.5 $p = 4.55699\text{e-}219$	0.298535 $p = 3.07979\text{e-}020$	0.342128 $p = 0$	20159.2 $p = 0$
Venezuela	319.504 $p = 4.17391\text{e-}070$	0.384974 $p = 4.42202\text{e-}019$	0.401325 $p = 0$	5656.19 $p = 0$

Table 5. Normality test EMPINEER.

Country	Doornik-Hansen test	Shapiro-Wilk test	Lilliefors test	Jarque-Bera test
Australia	52.0055 $p = 5.09509\text{e-}012$	0.425942 $p = 8.14523\text{e-}007$	0.102159 $p = 0.01$	288.026 $p = 2.85786\text{e-}063$
Brazil	223.33 $p = 3.19454\text{e-}049$	0.679515 $p = 6.90068\text{e-}014$	0.225239 $p = 0$	1758.23 $p = 0$
Canada	15.5818 $p = 0.000413477$	0.956573 $p = 0.00158816$	0.0786645 $p = 0.1$	21.0389 $p = 2.70064\text{e-}005$
China	8161.59 $p = 0$	0.15998 $p = 6.96215\text{e-}022$	0.419846 $p = 0$	42691.1 $p = 0$
India	82.7003 $p = 1.10118\text{e-}018$	0.479601 $p = 1.15506\text{e-}017$	0.260665 $p = 0$	8992.82 $p = 0$
Japan	316.552 $p = 1.82603\text{e-}069$	0.491505 $p = 1.79554\text{e-}017$	0.280174 $p = 0$	9642.29 $p = 0$
South Africa	8.34661 $p = 0.0154013$	0.976023 $p = 0.0517356$	0.0664106 $p = 0.29$	6.77282 $p = 0.0338299$
South Korea	6206.01 $p = 0$	0.191169 $p = 1.5629\text{e-}021$	0.398836 $p = 0$	38934.7 $p = 7.39392\text{e-}008$
United Kingdom	1345.86 $p = 5.62732\text{e-}293$	0.312954 $p = 4.71756\text{e-}020$	0.408784 $p = 0$	18030.3 $p = 0$
Venezuela	567.635 $p = 5.49049\text{e-}124$	0.416158 $p = 1.23967\text{e-}018$	0.462131 $p = 0$	709.252 $p = 0$

EMPINEER and EMPIREER data are normally distributed. Here, in all the tables, we have not reported the test results of the EMPI data for Switzerland as the same has failed to pass the stationarity test (Tables 4–6).

6. Application of bai and Ng normality test to EMPI data

Bai and Ng (2005) applied their test to a sample of 21 macroeconomic time series data related to US economy. These samples relate to different periods. In all the samples except for interest rate and unemployment rate, they have taken the first difference of the \ln (natural log) of the data. On the basis of the application of Bai and Ng (2005) test they reject normality in the US-Japan exchange rate, US CPI inflation rate, US 30-

Table 6. Normality test EMPIREER.

Country	Doornik-Hansen test	Shapiro-Wilk test	Lilliefors test	Jarque-Bera test
Australia	41.9963 $p = 7.59671\text{e-}010$	0.931495 $p = 3.67712\text{e-}005$	0.07643 $p = 0.13$	150.244 $p = 2.37101\text{e-}033$
Brazil	25.4056 $p = 3.04254\text{e-}006$	0.90733 $p = 1.76823\text{e-}006$	0.116766 $p = 0$	124.435 $p = 0$
Canada	13.9047 $p = 0.000956373$	0.956658 $p = 0.00161096$	0.0989492 $p = 0.01$	14.7912 $p = 0.00061395$
China	8164.34 $p = 0$	0.160082 $p = 6.98044\text{e-}022$	0.419709 $p = 0$	42693.9 $p = 0$
India	164.645 $p = 1.76942\text{e-}036$	0.27786 $p = 1.69037\text{e-}020$	0.430296 $p = 0$	10017.5 $p = 0$
Japan	229.541 $p = 1.4318\text{e-}050$	0.476678 $p = 1.03768\text{e-}017$	0.293514 $p = 0$	9364.65 $p = 0$
South Africa	18.0745 $p = 0.000118896$	0.949188 $p = 0.000484048$	0.127164 $p = 0$	17.7757 $p = 0.000138058$
South Korea	6286.24 $p = 0$	0.190803 $p = 1.54793\text{e-}021$	0.389815 $p = 0$	39096.9 $p = 0$
United Kingdom	6286.24 $p = 1.21458\text{e-}286$	0.320018 $p = 5.82809\text{e-}020$	0.398754 $p = 0$	17868.1 $p = 0$
Venezuela	296.82 $p = 3.51794\text{e-}065$	0.386902 $p = 4.70759\text{e-}019$	0.400013 $p = 0$	709.252 $p = 0$

Table 7. Normality test EMPIMER: Bai and Ng long run normality test.

Country	π_3	$\pi_3/5$	π_4	π_{34}	Power π_{34}
Australia	0.824295 $p = 0.204886$	2.526578 $p = 0.005759$	1.195313 $p = 0.115982$	2.704240 $p = 0.258691$	0.1365
Brazil	-1.326321 $p = 0.907633$	2.758026 $p = 0.002908$	1.734317 $p = 0.041431$	3.686219 $p = 0.158324$	0.1080
Canada	1.119130 $p = 0.131542$	1.265320 $p = 0.102878$	1.208428 $p = 0.113441$	1.417404 $p = 0.492283$	0.2610
China	1.084595 $p = 0.139050$	1.203627 $p = 0.114367$	1.159096 $p = 0.123209$	1.324111 $p = 0.515790$	0.2824
India	164.645 $p = 1.76942\text{e-}036$	0.27786 $p = 1.69037\text{e-}020$	0.430296 $p = 0$	10017.5 $p = 0$	0.0500
Japan	229.541 $p = 1.4318\text{e-}050$	0.476678 $p = 1.03768\text{e-}017$	0.293514 $p = 0$	9364.65 $p = 0$	0.0500
South Africa	18.0745 $p = 0.000118896$	0.949188 $p = 0.000484048$	0.127164 $p = 0$	17.7757 $p = 0.000138058$	0.0596
South Korea	6286.24 $p = 0$	0.190803 $p = 1.54793\text{e-}021$	0.389815 $p = 0$	39096.9 $p = 0$	0.0500
United Kingdom	1.092171 $p = 0.137379$	1.629183 $p = 0.051637$	1.124789 $p = 0.130339$	1.402060 $p = 0.496074$	0.2643
Venezuela	296.82 $p = 3.51794\text{e-}065$	0.386902 $p = 4.70759\text{e-}019$	0.400013 $p = 0$	709.252 $p = 0$	0.0502

day interest rate, and US stock returns. The result is given by authors on the basis of 5% critical value which is 1.96 for π_3 , π_4 , and 5.99 for π_{34} . Thus, out of 21 time series, that they examined, six time series are found to be non symmetric, and all these time series are financial time series. Among the financial time series only Canada-US exchange rate and German-US exchange rate are found to be symmetric. In Tables 7–9, we present the result of the application of Bai and Ng (2005) test to EMPI time series of ten countries (Switzerland's series is not put to normality check because it is not stationary). The hypotheses on test are – H_0 = Data is normally distributed, H_1 =Data is not normally distributed. We test these hypotheses at α of 0.05 in one tail test. The test

Table 8. Normality test EMPINEER: Bai and Ng long run normality test.

Country	π_3	$\pi_3/5$	π_4	π_{34}	Power π_{34}
Australia	0.962123 $p = 0.164639$	2.311630 $p = 0.010399$	1.073180 $p = 0.141595$	1.166545 $p = 0.558069$	0.1554
Brazil	1.229560 $p = 0.109431$	2.628032 $p = 0.004294$	1.250587 $p = 0.105543$	2.759713 $p = 0.251615$	0.1342
Canada	0.755401 $p = 0.225004$	3.243699 $p = 0.000590$	1.601438 $p = 0.054640$	4.350909 $p = 0.113557$	0.0972
China	1.032393 $p = 0.150944$	14.65633 $p = 0.000000$	1.105588 $p = 0.134452$	0.331889 $p = 0.847093$	0.9599
India	0.843014 $p = 0.199610$	1.998645 $p = 0.022823$	1.213510 $p = 0.112467$	2.263689 0.322438	0.1598
Japan	0.995534 $p = 0.159738$	1.823241 $p = 0.034133$	1.142342 $p = 0.126656$	2.188209 $p = 0.334839$	0.1650
South Africa	0.692166 $p = 0.244416$	0.621639 $p = 0.267090$	2.354924 $p = 0.009263$	4.243355 $p = 0.119830$	0.1700
South Korea	1.030480 $p = 0.151392$	10.69388 $p = 0.000000$	1.104982 $p = 0.134584$	0.125269 $p = 0.939287$	1.0000
United Kingdom	−1.355688 $p = 0.912401$	3.675629 $p = 0.000119$	0.903110 $p = 0.183234$	0.219689 $p = 0.895973$	0.9981
Venezuela	−1.865685 $p = 0.968957$	27.08616 $p = 0.000000$	1.510452 $p = 0.065464$	3.797907 $p = 0.149725$	0.1059

Table 9. Normality test EMPIREER: Bai and Ng long run normality test.

Country	π_3	$\pi_3/5$	π_4	π_{34}	Power π_{34}
Australia	0.0937125 $p = 0.174347$	1.197001 $p = 0.115653$	1.109317 $p = 0.133647$	1.153937 $p = 0.561598$	0.1570
Brazil	1.426561 $p = 0.076853$	2.132204 $p = 0.016495$	1.764111 $p = 0.038857$	2.363969 $p = 0.306670$	0.1535
Canada	0.553065 $p = 0.290109$	4.154797 $p = 1.63E-05$	1.564884 $p = 0.058805$	4.567623 $p = 0.101895$	0.0945
China	1.032395 $p = 0.150944$	14.65633 $p = 0.000000$	1.105581 $p = 0.134454$	0.331537 $p = 0.847242$	0.9602
India	0.971719 $p = 0.165595$	1.921587 $p = 0.027329$	1.132469 $p = 0.128719$	2.451300 $p = 0.293567$	0.1486
Japan	0.973644 $p = 0.165117$	1.760465 $p = 0.039164$	1.155613 $p = 0.123920$	2.253969 $p = 0.324009$	0.1604
South Africa	−0.270642 $p = 0.606667$	0.123960 $p = 0.450673$	2.623673 $p = 0.004349$	5.181959 $p = 0.074947$	0.0882
South Korea	1.030651 $p = 0.151352$	10.60079 $p = 0.000000$	1.104952 $p = 0.134590$	0.107724 $p = 0.947563$	1.0000
United Kingdom	−1.341607 $p = 0.910138$	3.098803 $p = 0.000972$	0.900951 $p = 0.183807$	4.709935 $p = 0.094897$	0.0928
Venezuela	14.04405 $p = 0.000000$	6279.211 $p = 0.000000$	5.260345 $p = 7.19E-08$	1073.438 $p = 0.000000$	0.0501

Table 10. Power test EMPIMER.

Country	Doornik-Hansen test	Shapiro-Wilk test	Lilliefors test	Jarque-Bera test
Australia	0.0519	0.8780	0.9846	0.0500
Brazil	0.0513	0.7007	0.9905	0.0501
Canada	0.0549	0.4350	1.0000	0.0516
China	0.0500	1.0000	0.8823	0.0500
India	0.0502	0.9284	0.9839	0.0500
Japan	0.0507	0.7688	0.9869	0.0500
South Africa	0.0524	0.3094	0.9947	0.0507
South Korea	0.0500	0.9267	0.0500	0.0507
United Kingdom	0.0502	0.9784	0.9530	0.0500
Venezuela	0.0501	0.8684	0.8467	0.0500

Table 11. Power test EMPINEER.

Country	Doornik-Hansen test	Shapiro-Wilk W	Lilliefors test	Jarque-Bera test
Australia	0.0527	0.8642	0.9846	0.0505
Brazil	0.0507	0.6094	0.9975	0.0501
Canada	0.0611	0.4153	1.0000	0.0581
China	0.0500	1.0000	0.8843	0.0500
India	0.0520	0.8197	0.9916	0.0500
Japan	0.0505	0.8064	0.9858	0.0500
South Africa	0.0720	0.4055	1.0000	0.0779
South Korea	0.0500	0.9124	0.9052	0.0500
United Kingdom	0.0501	0.9712	0.8954	0.0500
Venezuela	0.0503	0.8880	0.8391	0.0502

Table 12. Normality test EMPIREER.

Country	Doornik-Hansen test	Shapiro-Wilk W	Lilliefors test	Jarque-Bera test
Australia	0.0505	0.3181	1.0000	0.0501
Brazil	25.4056	0.4421	1.0000	0.0513
Canada	0.0625	0.4153	1.0000	0.0617
China	0.0500	1.0000	0.8844	0.0500
India	0.0510	0.9866	0.8734	0.0500
Japan	0.0507	0.8230	0.9806	0.0500
South Africa	0.0595	0.4192	1.0000	0.0596
South Korea	0.0500	0.9996	0.9137	0.0500
United Kingdom	0.0500	0.9673	0.9052	0.0500
Venezuela	0.0505	0.9164	0.9040	0.0502

Table 13. Cross country standard deviation of EMPI.

Country	σ EMPIMER	Rank	σ EMPINEER	Rank	σ EMPIREER	Rank
South Korea	0.0006014	1	0.0005614	1	0.0005614	1
China	0.2223	2	0.2224	2	0.2224	2
Switzerland	0.68081	3	16.70	9	0.6942	3
Canada	3.55	4	2.871	3	2.962	4
South Africa	6.794	5	5.300	5	5.054	6
India	11.55	6	6.809	6	6.893	8
Brazil	16.42	7	9.683	7	6.537	7
Japan	16.47	8	16.07	8	15.75	9
United Kingdom	18.01	9	54.65	10	48.22	10
Australia	21.98	10	3.651	4	3.772	5
Venezuela	44.72	11	127.30	11	44.51	11

statistics have χ^2 distribution. In the case of all the tests the reported p values are less than the critical α of 0.05. Hence, we reject the null hypothesis that the EMPIMER, EMPINEER, and EMPIREER data is normally distributed for these ten countries.

7. Power comparison

The reliability of a test can be measured on the basis of its power. A test like Bai and Ng (2005) which is developed exclusively for time series data should show more power in comparison to all other tests which are not developed exclusively for time series data (Tables 10–12). In this section we compare the power of the alternative tests of normality for our sample size of 106. The power of a test depends on the effect size, sample size, and the level of significance. We present the result of the power all the tests for all the countries at a medium effect size of 0.2 and a level of significance of 0.05 in one

tail test. The results show that the Lilliefors test and Shapiro-Wilk test have the highest power in almost all the cases and Jarque-Bera test and Doornik-Hansen test have the lowest powers. The power of Bai-Ng test is also low in almost all the cases, though higher than the power of Doornik-Hansen test and Jarque-Bera test. Only in the case of China and South Korea the power of Bai-Ng test is high (and in the one special case of EMPINEER of United Kingdom). One should also keep in the mind that China and South Korea are two countries with the least volatile exchange market pressure index, judged on the basis of the standard deviation of their EMPIMER, EMPINEER, and EMPIREER. One needs further research to understand that why the power of Bai-Ng test is high in case of EMPINEER of United Kingdom although it has a low rank on basis of standard deviation of EMPINEER. See Table 13 to know about the ranking of countries on the basis of their standard deviation. Thus, Bai and Ng (2005) test, though developed exclusively for time series data, has low power in comparison to other tests in all cases except some selected cases with low standard deviations.

8. Conclusion

In this paper, we tested the normality of the Girton and Roper Exchange Market Pressure Index (EMPI) for eleven countries of the world and also compared the statistical power of the tests applied. EMPI is a financial time series, and financial time series may not be normally distributed even for a large sample size. We created three indices using different exchange rates and called them EMPIMER (Exchange Market Pressure Index based on market exchange rate), EMPINEER (Exchange Market Pressure Index based on nominal exchange rate), and EMPIREER (Exchange Market Pressure Index based on real effective exchange rate). The tests we used to test the normality of EMPI are Jarque and Bera (1987), Shapiro and Wilk (1965), Lilliefors (1967), Doornik and Hansen (2008), and Bai and Ng (2005). The sample size we took is 106. On the basis of all these tests, all the EMPI time series are found to be not normally distributed. The power comparison, of the statistical tests employed, was at the same sample size and the same effect size for all. Shapiro-Wilk test and Lilliefors test were found to be the most powerful without any exception. Bai-Ng test was found to be powerful only in the case of South Korea and China: these are the countries with the lowest standard deviation of all the three EMPIs among all the countries studied here. For one exception, it was found that the Bai-Ng test was also powerful for United Kingdom's EMPINEER, a nation with the highest EMPINEER standard deviation rank of ten, among all the countries examined here. The intuition behind the results is clear. The problem of autocorrelation in EMPI time series data prevents the series from converging to its central value even for large sample sizes. For analyzing patterns in such data, we should bank on frequency domain rather than time domain analysis. The low power of Bai and Ng (2005) test is matter of further examination as this test, supposedly, should have high power as is built exclusively for time series data by improving on skewness-kurtosis based tests of normality.

Notes

1. For a theoretical discussion on Extreme Value Theory see De Haan and Ferreira (2007).

2. Here, Δs_t and Δf_t are calculated on different bases hence before linearly combining them we must multiply them by some weighting factors but G-R Model does not use any such weights. For a discussion on weights of EMP index see Klaassen (2011).

ORCID

Sanjay Kumar  <http://orcid.org/0000-0003-4613-076X>

References

- Bai, J., and S. Ng. 2005. Tests for skewness, kurtosis, and normality for time series data. *Journal of Business & Economic Statistics* 23 (1):49–60. doi:10.1198/073500104000000271.
- D'Agostino, R. B., and M. A. Stephens. 1986. *Goodness-of-fit techniques*. New York: Marcel Dekker Inc.
- De Haan, L., and A. Ferreira. 2007. *Extreme value theory: An introduction*. USA: Springer Science & Business Media.
- Doornik, J. A., and H. Hansen. 2008. An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics* 70:927–39.
- Garita, G., and C. Zhou. 2009. Can financial openness help avoid currency crises? https://mpa.ub.uni-muenchen.de/23166/1/MPRA_paper_23166.pdf
- Girton, L., and D. Roper. 1977. A monetary model of exchange market pressure applied to the postwar canadian experience. *The American Economic Review* 67 (4):537–48.
- Gochoco-Bautista, M. S., and C. C. Bautista. 2005. Monetary policy and exchange market pressure: The case of the philippines. *Journal of Macroeconomics* 27 (1):153–68. doi:10.1016/j.jmacro.2003.09.006.
- Jarque, C. M., and A. K. Bera. 1987. A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique* 55 (2):163–72. doi:10.2307/1403192.
- Kamaly, A., and N. Erbil. 2000. A var analysis of exchange market pressure: A case study for the mena region. Working Paper 2025. Economic Research Forum.
- Klaassen, F. 2011. Identifying the weights in exchange market pressure.
- Lilliefors, H. W. 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* 62 (318):399–402. doi:10.1080/01621459.1967.10482916.
- Mandelbrot, B. B. 1997. The variation of certain speculative prices. In *Fractals and scaling in finance*, 371–418. Springer.
- Shapiro, S. S., and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52 (3/4):591–611.
- Tanner, E. (2002.). *Exchange market pressure, currency crises, and monetary policy: Additional evidence from emerging markets*. Number 2002-2014. International Monetary Fund. doi:10.5089/9781451843132.001.

Therapeutic Targeting of Repurposed Anticancer Drugs in Alzheimer's Disease: Using the Multiomics Approach

Dia Advani and Pravir Kumar*

Cite This: *ACS Omega* 2021, 6, 13870–13887

Read Online

ACCESS |



Metrics & More

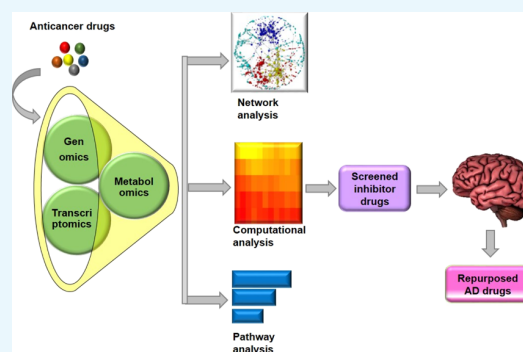


Article Recommendations



Supporting Information

ABSTRACT: Aim/Hypothesis: The complexity and heterogeneity of multiple pathological features make Alzheimer's disease (AD) a major culprit to global health. Drug repurposing is an inexpensive and reliable approach to redirect the existing drugs for new indications. The current study aims to study the possibility of repurposing approved anticancer drugs for AD treatment. We proposed an *in silico* pipeline based on "omics" data mining that combines genomics, transcriptomics, and metabolomics studies. We aimed to validate the neuroprotective properties of repurposed drugs and to identify the possible mechanism of action of the proposed drugs in AD. Results: We generated a list of AD-related genes and then searched DrugBank database and Therapeutic Target Database to find anticancer drugs related to potential AD targets. Specifically, we researched the available approved anticancer drugs and excluded the information of investigational and experimental drugs. We developed a computational pipeline to prioritize the anticancer drugs having a close association with AD targets. From data mining, we generated a list of 2914 AD-related genes and obtained 49 potential druggable targets by functional enrichment analysis. The protein–protein interaction (PPI) studies for these genes revealed 641 interactions. We found that 15 AD risk/direct PPI genes were associated with 30 approved oncology drugs. The computational validation of candidate drug–target interactions, structural and functional analysis, investigation of related molecular mechanisms, and literature-based analysis resulted in four repurposing candidates, of which three drugs were epidermal growth factor receptor (EGFR) inhibitors. Conclusion: Our computational drug repurposing approach proposed EGFR inhibitors as potential repurposing drugs for AD. Consequently, our proposed framework could be used for drug repurposing for different indications in an economical and efficient way.



1. INTRODUCTION

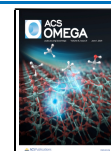
The alarming progression rate, limited therapeutics, and the slow pace of new drug development for Alzheimer's disease (AD) draw the attention of research groups and pharmaceutical companies toward exploring new alternatives. Conventionally, AD is denoted as a central nervous system (CNS) disorder characterized by abnormal amyloid- β ($A\beta$) aggregation, tangle formation of hyperphosphorylated tau, oxidative stress, and hyperactivity glial and microglial cells.¹ The latest reports by the Alzheimer's association suggested that five FDA-approved drugs are currently marketed for AD.² The failure rate of AD therapeutics is more than 99%, and for the disease-modifying therapies, it is 100%. It has been a matter of more than 20 years; no new drug is licensed for AD. The research community is continuously involved in developing new drug discovery strategies; one of the examples is drug repurposing. To encourage the use of repurposed drugs, the National Institute of Aging grants \$2.8 million to Case Western Reserve University School of Medicine to identify potential FDA-approved medicines as repurposed agents for AD. The major classes of drugs investigated for AD as repurposed agents are antihypertensive, antidiabetic, antiasthmatic, retinoid recep-

tors, anticancer agents, antiepileptic, antidepressive, and antimicrobial agents.³ In addition to omics analysis, the concept of pharmacogenomics has gained significant attention in drug repurposing. Studies have suggested that drugs can regulate the expression of small noncoding RNAs such as micro RNAs (miRNAs) and their precursors. For instance, miravirsin is the first miRNA-targeted small molecule that has come in clinical trials and can inhibit miR-122 expression required to replicate hepatitis C virus.⁴ In a study by Yu *et al.*, potential repurposing drugs were identified for breast cancer based on miRNA–disease–drug tripartite relationships.⁵ Likewise, in a recent study, Aydin *et al.* reported miRNA-mediated repurposed drugs for Prolactinoma treatment *via in vitro* experimentation.⁶

Received: March 22, 2021

Accepted: May 10, 2021

Published: May 19, 2021



ACS Publications

© 2021 The Authors. Published by
American Chemical Society

13870

<https://doi.org/10.1021/acsomega.1c01526>
ACS Omega 2021, 6, 13870–13887

Drug repurposing is an opportunistic strategy of identifying new indications of the drugs already approved in the market. A review of different repurposing examples suggested that about 46 drugs have already been repurposed for various indications, and encouraging studies are consistently publishing.⁷ A recent study has revealed that pharmaceutical companies have placed the market for repositioned drugs at \$31.3 billion in 2020, generating about 25% of this industry's annual revenue. Recent estimates suggested that about 30% of the FDA-approved drugs are actually the repurposed drugs.⁸

To date, most of the repurposing studies have been published for parasitic diseases, multiple cancers, tuberculosis, and malaria.⁹ This drug discovery strategy is gaining continuous appreciation as it bypasses the efforts and cost input required for the early stages of drug development. The repurposing of drugs involves two different approaches, computational and experimental.¹⁰ Computational approaches are the combination of systematic steps taken for the initial identification of promising repurposable compounds. The primary methods used for the computational approach are network-based, text mining-based, and semantics-based.¹¹

In the last few years, omics sciences accelerated the drug discovery process by overcoming the challenges associated with it. Recent technological advancements enabled scientists to develop genomics-, transcriptomics-, proteomics-, and metabolomics-based databases. Genomics studies helped us to understand the genetic basis of complex diseases.¹² In the past decade, the genome-wide association studies (GWAS) catalog has revolutionized the area of genomics to identify complex genotype–phenotype associations.¹³ The transcriptomics studies help us to understand the effect of drugs on different cellular states. The expression profiling and genomics studies give the right directionality to gene–phenotype associations.^{14,15} The proteomics studies are extensively used to understand the mechanistic basis of disease.¹⁶ Similarly, the analysis of metabolome provides knowledge of associations of biochemical events with phenotypes.¹⁷

An exciting interplay between cancer and AD gives a direction to use anticancer drugs as repurposed therapeutics. Accumulating evidence has suggested that cancer and AD share some familiar biological hallmarks, and a significant link exists between cancer history and AD neuropathology.^{18,19} In a recent study, Lee *et al.* established an interrelationship between cancer and AD at the transcription level. They compared differentially expressed genes between AD and nine different cancers and found that glioblastoma multiforme shared a strong correlation with AD.²⁰ The repurposing of oncology drugs for AD is underway, and many drugs, for instance, bosutinib, dasatinib, nilotinib, bexarotene, tamibarotene, and thalidomide (ClinicalTrials.gov identifier: NCT02921477, NCT04063124, NCT02947893, NCT01782742, NCT01120002, and NCT01094340, respectively), are in clinical trials for AD.²¹ A study by Lonskaya *et al.* confirmed the therapeutic relevance of tyrosine kinase inhibitors nilotinib and bosutinib in AD, where the drugs facilitated amyloid clearance and reduced neuroinflammation.²² A drug repurposing study by the neuroinformatics approach has proposed that the anticancer drug bexarotene could reduce A β aggregation by interacting with receptors for advanced glycation end products (RAGE) and beta-secretase (BACE-1).²³ A drug repurposing study by Madepalli Lakshmana and the group found that anticancer drug carmustine (BiCNU) could regulate amyloid precursor protein (APP) processing and trafficking to reduce

A β aggregation in the brain.²⁴ Likewise, a study targeting vascular activation in AD has proposed that the anticancer drug sunitinib could reduce vascular activation of various proteins such as amyloid-beta, tumor necrosis factor-alpha (TNF α), interleukin-6 (IL-6), interleukin-1 beta, thrombin, and matrix metalloproteinase 9 and ameliorated cognitive dysfunction in AD transgenic mice. Additionally, a study on the antimetabolic drug, paclitaxel, has revealed the drug's potential in reducing tau-associated pathologies by preventing tau-induced axonal swelling, reversal of microtubule polar orientation, prevention of neurite degeneration, and inhibition of impaired organelle transport and accumulation.²⁵ In parallel, a study on the tyrosine kinase inhibitor, pazopanib, in the AD mouse model has identified the potential of the drug in reducing tau pathology and astrocytic activity. The study has proposed that the drug could not alter microglial activity; however, it could modulate the activity of inflammatory markers and thus provide neuroprotection.²⁶

The motivation of this study is to uncover the hidden neuroprotective potential of anticancer drugs. We adopted an integrated omics data-based repurposing strategy, including genomics, transcriptomics, and metabolomics, and validated our results by different computational methods. Our study was concentrated on FDA-approved anticancer drugs and their repurposing for AD. We developed a bioinformatic pipeline to assign a ranking of the repurposed drugs based on the computational drug repurposing score (CoDReS) validated by network and structural similarity analysis with approved AD drugs. The study also aims to combine the physicochemical analysis, drug-likeness, pathway analysis, and microRNA (miRNA) analysis of repurposing anticancer drugs to understand better the mechanisms involved. The study helped to identify the significant pathways and cancer-related genes associated with the pathogenesis of AD. The study also set a new direction to understand the complex relationship between AD and cancer that would be considered for other neurodegenerative diseases.

2. METHODOLOGY

2.1. Data Extraction. To obtain information on AD-associated genetic variations, we analyzed GWAS studies for AD from NHGRI-EBI GWAS catalog (<http://www.ebi.ac.uk/gwas>).²⁷ The database provides a consistent knowledge of single-nucleotide polymorphism (SNP)-trait associations for various diseases. We extracted GWAS data for (1) PUBMED ID, (2) study accession, (3) genes, (5) SNPs, (6) *P*-value, and (7) OR (odds ratio). Genes are considered significant, which fall under the genomic regions associated with SNPs ($r^2 > 0.6$). For transcriptomics data, NCBI Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo/>) database that contains microarray and next-generation sequence functional genomic data sets was used.²⁸ The collected expression profile of the AD series GSE1297 was analyzed by GEO2R. The GSE1297 series contains microarray analysis data of the hippocampal region of 9 control and 22 AD subjects. The metabolomics data were collected from the Human Metabolome Database (HMDB, <http://www.hmdb.ca>), which contains 114,187 metabolite entries.²⁹ The database was searched for (1) AD linked metabolites, (2) protein name, (3) Uniprot ID, (4) type of metabolite, and (5) gene name.

2.2. Prioritization of Candidate Genes. We utilized two different computational tools to identify the most significant genes associated with AD. The genes obtained from various

omics approaches were then subjected to enrichment analysis by online DAVID functional annotation tool and gene set to diseases (GS2D) tool. DAVID (<https://david.ncifcrf.gov>) provides an integrated platform to extract meaningful biological information from the list of genes enriched in genome-scale studies.³⁰ GS2D (<http://cbdm.uni-mainz.de/geneset2diseases>) is a web tool that performs enrichment analysis based on significant biomedical citations from PubMed.³¹ The gene–disease associations were filtered by a minimum number of citations found (default = 5), the minimum number of gene–disease associations (default = 2), and the maximum false discovery rate (FDR = 0.05). The FDR is used as a matrix in drug repurposing to measure significance of drug-indication scores.³²

The enriched genes were then analyzed for protein–protein interaction (PPI) using the Molecular Interaction Search Tool (MIST) database. MIST (<http://fgrtools.hms.harvard.edu/MIST/>) database can be used to devise significant protein–protein and genetic interactions for different species.³³

2.3. Drug Target Mapping. We have combined the information from genomics, transcriptomics, and metabolomics approaches and had a list of genes associated with AD. To develop a link between AD-related genes with currently available drug projects, we tracked two different databases. DrugBank (www.drugbank.com) (version 5.1.5) contains around 13,554 drug entries incorporating various approved and experimental small molecules and biologics.³⁴ Similarly, the Therapeutic Target Database (TTD) (<http://db.idrblab.net/ttd/>) accommodates 3419 targets and 37316 drug projects.³⁵ We included only those targets for which anticancer drugs are available and excluded the others. All the drugs with clinical, experimental, or withdrawn status were excluded, and only FDA-approved drugs were considered for this study. The information about drugs such as drug name, DrugBank ID, current indication, and drug mode of action was collected.

2.4. Validation of Candidate Drugs. The PPIs from the previous steps were then analyzed by the STRING database (string-db.org) that covers known and predicted interactions for different organisms.³⁶ The experimentally significant interactions (with high interaction scores) were selected, and the others were excluded from the study. The drug–target interactions were evaluated using the STITCH (search tool for interactions of chemicals) (<http://stitch.embl.de/>) database that integrates interactions of 300,000 chemicals and 2.6 million proteins.³⁷ In a complex system, two interacting genes are represented as nodes connected by an edge. The interaction networks were further analyzed, and networks were generated using Cytoscape software v3.3.0 (www.cytoscape.org).

For validation of promising drug candidates on the validation network, we measured network topology parameters such as degree centrality, betweenness, and topological coefficients using the CentiScaPe app on Cytoscape software. A degree is a topological parameter that corresponds to the number of interactions or connections for a given node. Betweenness corresponds to the centrality index of a given node. It represents the shortest path between two adjacent nodes. In biological networks, only a few nodes (hub nodes) have a high degree centrality and the nodes having the shortest path distance are designated as bottlenecks. Both hub nodes and bottlenecks are considered topologically important and biologically significant.³⁸ The topological coefficient is a relative measure that denotes the extent to which a node

shares neighbors with other nodes in the network. The nodes that share no neighbor are assigned a topological coefficient value of 0. The candidate drugs were given ranks based on different topological parameters. The drugs having a higher degree centrality value were considered as topologically important and biologically significant. In short, the drugs (nodes) with higher degree centrality values are regarded as hub nodes with considerable importance in the network.

2.5. Drug Repurposing. The candidate drugs obtained from the previous studies were analyzed for their repurposing potential for AD using the CoDReS tool. CoDReS (<http://bioinformatics.cing.ac.cy/codres>) is a web-based tool that integrates information from the biologically available data sets, calculates affinity scores of protein and ligand pairs, and evaluates drug-likeness and structural similarities.³⁹ The candidate drugs with good repositioning scores were then presented by the hierarchical clustering algorithm of the ChemMine server.⁴⁰ Hierarchical clustering is a powerful approach to find structural and physicochemical similarities of compounds based on atom pair similarity measures. The similarity scores were calculated based on the Z-score values. Also, we calculated the structural similarity with the approved Alzheimer's drugs, namely, donepezil, rivastigmine, galantamine, and memantine. The similarity workbench tool of the ChemMine server was used, and similarity scores were represented as the Tanimoto coefficient, the most widely used metric to compare the molecular structure similarities in medicinal chemistry.⁴¹ The tool utilizes the maximum common substructure (MCS) fingerprint method to find the largest substructures two compounds have in common and present it as the Tanimoto coefficient.

2.6. Literature Validation of the Drug–Disease Relationship. To obtain the information related to neuroprotective functions of anticancer drugs, we have searched the PubMed database using the keywords “anticancer drugs and neuroprotection,” “anticancer drugs and AD,” and anticancer drugs and neurodegenerative disorders. We collected information on whether the proposed repurposing drugs have any neuroprotective mechanism associated with them.

2.7. Swiss ADMET Analysis of Candidate Drugs. The development of drugs for the CNS disorders poses a challenge due to the blood–brain barrier (BBB). While designing a drug for brain diseases, physicochemical properties and brain permeation properties should be optimized. In consideration of this challenge, we analyzed our candidate repurposed drugs for physicochemical properties using the SwissADME analysis tool. SwissADME (<http://www.swissadme.ch/>) is a user-friendly web tool to predict physicochemical properties, pharmacokinetics, and drug-likeness of small molecules.⁴² We collected information about physicochemical properties such as molecular weight, number of rotatable bonds, number of H-bond donor and acceptors present, partition coefficient ($M \log P$), and topological polar surface area (TPSA) and blood–brain permeation, where $M \log P$ was the measure of lipophilicity and TPSA was the measure of the sum of the surfaces of polar atoms present.

2.8. Functional Similarity with MicroRNAs. To further validate our results, we identified miRNAs related to AD from Human microRNA Disease Database (HMDD) (<https://www.cuilab.cn/hmdd>).⁴³ HMDD contains information regarding experimentally validated microRNA–disease associations. We also retrieved information of miRNAs associated with the identified repurposed drugs and then constructed a network

Advani and Kumar, 2021

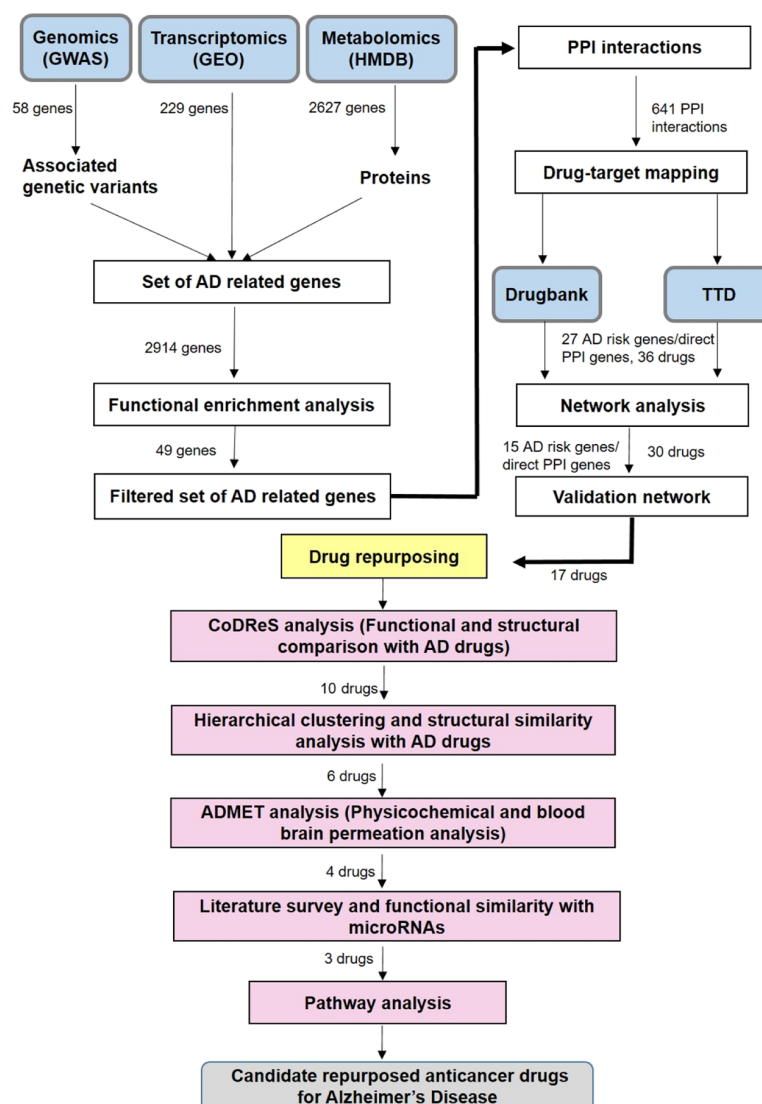


Figure 1. Flow chart of drug repurposing by omics data mining: We retrieved information on AD risk genes from GWAS, transcriptomics, and metabolomics approaches. We found 2914 AD risk genes from which 58 genes were extracted from GWAS, 229 genes were extracted from GEO transcriptomics data, and 2627 genes were related to 128 metabolites from the HMDB database. After functional enrichment analysis, we filtered out 49 AD-associated targets. The PPI network analysis resulted in 641 PPI interactions. We performed drug target mapping to find candidate drugs from DrugBank and TTD databases. Out of 641, 25 PPI interactions were found to be associated with 36 approved anticancer drugs. We excluded the information related to investigational and experimental drugs. We analyzed gene–gene and gene–drug interactions and selected the top 10 PPI interactions that correspond to 30 anticancer compounds. These 30 drugs were then analyzed by the CoDReS web tool that proposes 10 candidate drugs for AD. These drugs were then compared with the available Alzheimer’s therapeutics for structural and functional similarities, where six drugs have shown to be hierarchically clustered. ADMET analysis, pathway analysis, and functional similarity with miRNAs resulted in potential repurposing anticancer drugs against AD.

that combines miRNAs that share common targets between the repurposed drugs and AD. We considered only the miRNAs that were neuroprotective in nature. The disease–miRNA–drug and miRNA–drug relationships were presented in the form of a network using Cytoscape software. The information of AD-related miRNAs, repurposed drugs, and their targets was given as the input.

2.9. Pathway Analysis. To establish a connection of AD-related genes with cancer, we compare the expression pattern of genes with AD and the most common 13 types of cancers prescribed by the National Cancer Institute (NIH).⁴⁴ To

discover the molecular mechanisms regulated by the identified genes, we performed pathway analysis (KEGG,⁴⁵ Bioplane,⁴⁶ and WikiPathways⁴⁷) using the Enrichr tool. Enrichr (<http://amp.pharm.mssm.edu/Enrichr/>) is a web-based enrichment analysis tool that accumulates biological knowledge (genes, diseases, pathways, and drugs) of more than 102 gene set libraries.⁴⁸ The tool has provided information about biologically relevant pathways or enriched pathways for the set of the given genes. These enriched pathways were associated with the given gene list more than would be expected by chance. We also extracted the information of disease signatures

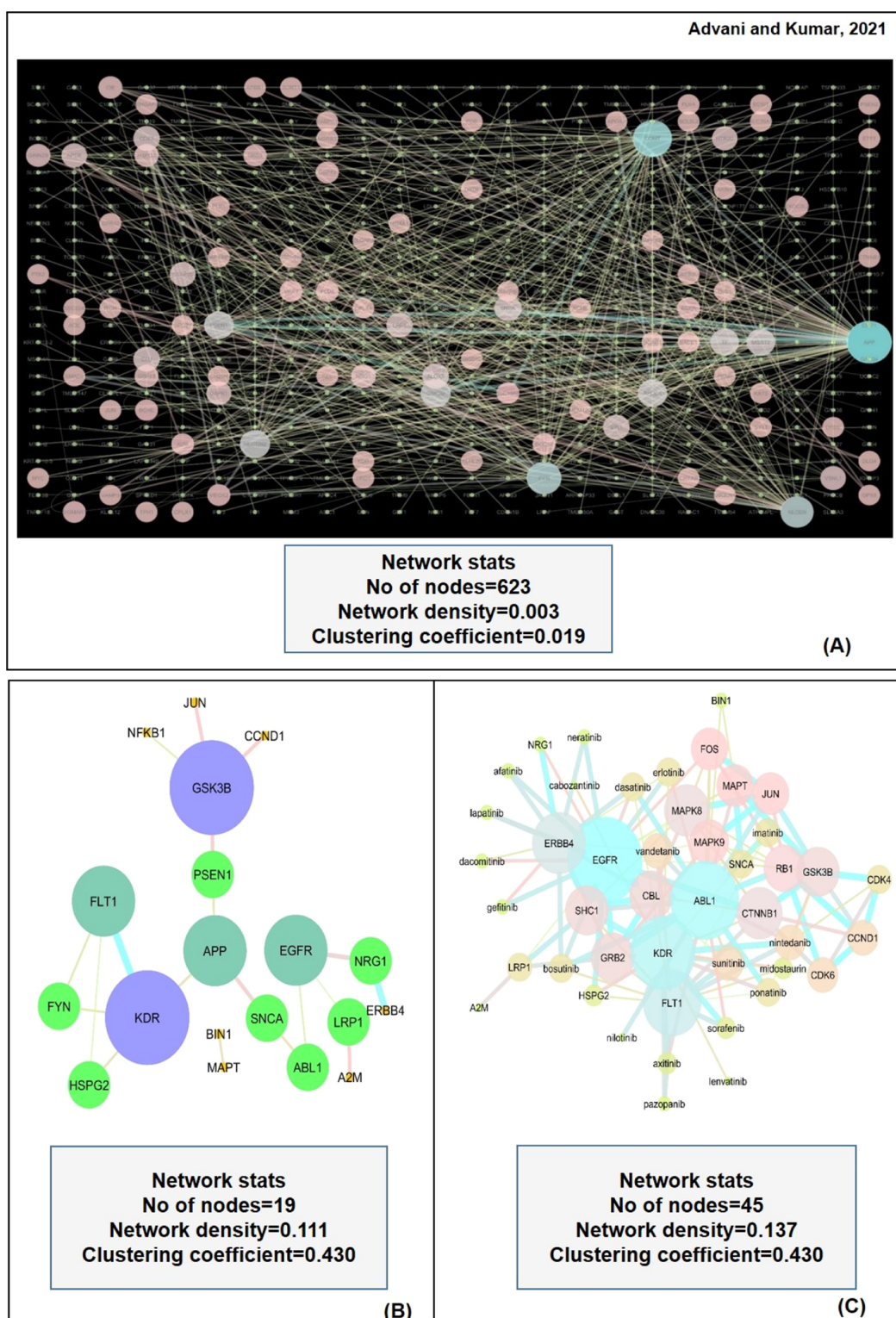


Figure 2. (A) Network is showing PPI interactions for AD-related genes. (B) STRING network of experimentally significant interactions. Glycogen synthase kinase 3 beta (GSK3B), vascular endothelial growth factor receptor 2 (KDR), APP, vascular endothelial growth factor receptor 1 (FLT1), and epidermal growth factor receptor (EGFR) were identified as the hub nodes. (C) STITCH network of drug-gene interactions. Nintedanib, sunitinib, vandetanib, dasatinib, erlotinib, imatinib, ponatinib, and bosutinib were reported as hub nodes as drugs. The size of individual nodes and the thickness of edges correspond to the significance and strength of interactions, respectively.

(DisGeNET and OMIM-based information) related to the given genes using the Enrichr tool. The output of Enrichr is

ranked list terms, and ranking is provided based on *p*-value scores. Enrichr calculates the *p*-value based on Fisher's exact

test that assumes binomial distribution and independence for the probability of the given input gene.

An overview of the complete pipeline is shown in Figure 1.

3. RESULTS

3.1. Omics Data Mining and Enrichment Analysis Revealed AD-Related Genes. The omics data approach enabled us to identify AD-related genes. We collected information about 58 unique genes from 37 GWAS studies. The *P*-value of the identified genes varies from 8×10^{-189} (minimum) to 8×10^{-6} (maximum). We identified 229 genes in the form of differentially coexpressed genes from transcriptomics studies. The data obtained from the HMDB database reported 128 AD-related metabolites that correspond to 2627 genes from metabolomics data. Most of the proteins associated with the retrieved metabolites had unknown functions, while some were enzymes or transporters. We combined the information from different omics approaches, and finally, 2914 genes were found to be associated with AD.

DAVID functional enrichment analysis of 2914 genes revealed that 13 genes from GWAS studies, 18 genes from the transcriptomics approach, and 239 genes from the metabolomics approach have significant associations with AD. Similarly, GS2D functional enrichment analysis revealed that 12 genes from GWAS studies, 4 genes from the transcriptomics approach, and 62 genes from the metabolomics approach were significantly linked with AD.

When we compared the two enrichment analysis methods, 49 AD-related genes were shared in the two enrichment methods (Table S1).

3.2. PPI Network Analysis Revealed Potential Interactors of AD-Risk Genes. We evaluated the PPI network of the 49 AD-risk genes to explore the possibility of any of the genes from the PPI network that serve as a target for approved anticancer drugs. We selected PPI interactions with a high confidence score and excluded the interactions with medium to low confidence. We found 641 PPI interactions from the MIST database results, as shown in Figure 2A. All the PPI genes of 641 interactions, along with 49 AD-risk genes, were searched in the DrugBank database and TTD to find the association with known anticancer drugs. Among the PPI interactors, 17 genes were reported to have approved anticancer medications available in the considered drug repositories. We found that the epidermal growth receptor (EGFR) is the most frequently appeared PPI interactor interacting with four different AD-associated targets APP, alpha-synuclein (SNCA), neuregulin 1 (NRG1), and LDL receptor related protein 1 (LRP1). These PPI interactions were then evaluated by the STRING database and presented on the validation network, as shown in Figure 2B. The topological parameters of genes in STRING, such as degree centrality, betweenness, and topological coefficients, were analyzed by Cytoscape and are presented in Table 1.

The topological parameters were used to identify the hub nodes in the validation network. We identified glycogen synthase kinase beta (GSK3B), kinase insert domain receptor (KDR), APP, EGFR, and Fms-related receptor tyrosine kinase 1 (FLT1) as the top five nodes. GSK3B and KDR had the highest degree centrality values of 4.0 and betweenness values of 0.35 and 0.32, respectively, while APP, EGFR, and FLT1 had degree centrality values of 4 and betweenness values of 0.69, 0.43, and 0.004, respectively. Among the identified genes, GSK3B is a multifunctional protein kinase regulating various cellular processes and is implicated in several diseases. In AD,

Table 1. Topological Parameters of Genes (Nodes) on the STRING Validation Network Using CentiScaPe App on Cytoscape Software^a

Gene	Degree	Betweenness	Topological Coefficient
GSK3B	4	0.35	0.25
KDR	4	0.329166667	0.45
APP	3	0.691666667	0.333333333
EGFR	3	0.433333333	0.333333333
FLT1	3	0.004166667	0.666666667
SNCA	2	0.5	0.5
ABL1	2	0.458333333	0.5
PSEN1	2	0.4	0.5
NRG1	2	0.125	0.5
LRP1	2	0.125	0.5
HSPG2	2	0	0.875
FYN	2	0	0.875
ERBB4	1	0	0
JUN	1	0	0
CCND1	1	0	0
A2M	1	0	0
MAPT	1	0	0
BIN1	1	0	0
NFKB1	1	0	0

^aGenes with significant values are highlighted.

GSK3 is considered a regulator of the two pathological hallmarks, senile plaques and neurofibrillary tangles.^{49,50} The other identified target APP is a single transmembrane protein that acts as a multifunctional cell surface receptor. APP plays a major role in AD pathogenesis as it is associated with A β production, synaptic function, and neuronal homeostasis.^{51,52} The EGFR is a transmembrane molecule that belongs to the HER/ERBB superfamily of receptors. The binding of ligands to this receptor triggers several signaling pathways that promote cell proliferation and cell survival. The other two genes, vascular endothelial growth factor receptor (VEGFR1) or FLT1 and VEGFR2 or KDR, are the two receptors playing a significant role in the signal transduction pathways mediated by the VEGF.⁵³ Some studies have suggested that both FLT1 and KDR are associated with AD neuropathology by inhibiting pro-angiogenic signaling mediated by the VEGF.^{54,55}

3.3. Drug Mapping Identified Potential Repurposing Candidates for AD. Drug target mapping from DrugBank and TTD has shown that 28 direct PPI/AD risk genes were associated with 36 FDA-approved anticancer drugs (Table S2). We omitted the targets related to any investigational, experimental, or withdrawn anticancer drugs. From 36 drugs, 11 drugs were associated with only one direct PPI gene/AD risk gene, while 25 drugs were those that interacted with more than one gene. The retrieved drugs were related to diverse modes of actions, such as inhibitors, antagonists, substrates, and some had unknown functions. The experimentally significant interactions obtained from STRING analysis corresponded to 30 drugs from which 4 drugs (brigatinib, zanubrutinib, osimertinib, and erdafitinib) were not identified by the STITCH database and were excluded from the study.

Of the 26 candidate repurposing drugs, six drugs (cisplatin, encorafenib, vinblastine, paclitaxel, docetaxel, and regorafenib) had not shown any interaction.

Additionally, three drugs (bosutinib, nilotinib, and dasatinib) were in clinical trials for AD or related dementias and were not included in this study. Therefore, the remaining 17 drugs were considered novel candidate repurposing drugs for AD. The candidate drugs with their AD-related targets and PPI targets are summarized in Figure 3.

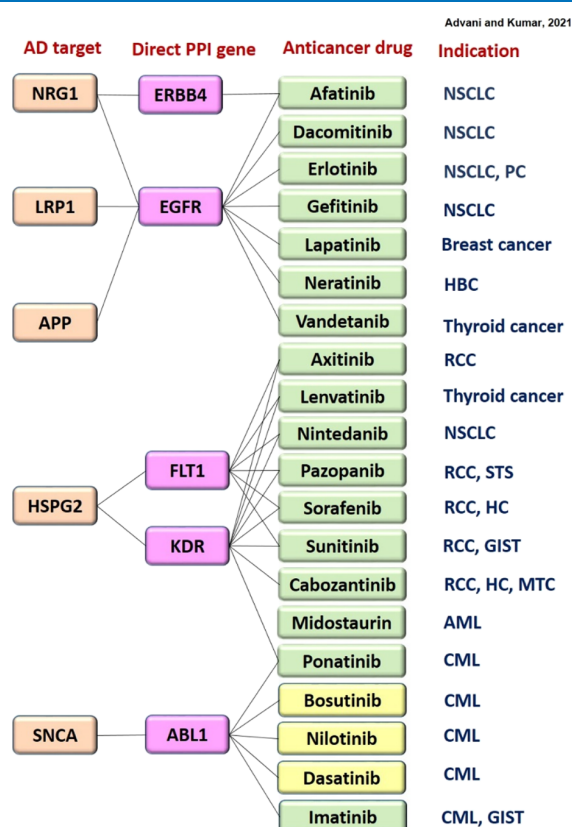


Figure 3. Summary of AD risk genes, genes in direct PPI, and targeted anticancer drugs. Drugs shown in yellow boxes were known in clinical studies as AD therapeutics, and drugs in green boxes were considered as potential repurposing candidates. Some drugs such as afatinib, axitinib, lenvatinib, nintedanib, pazopanib, sorafenib, and ponatinib interact with more than one target. NRG1: neuregulin 1; ERBB4: ErbB2 receptor tyrosine kinase 4; LRP1: LDL receptor-related protein 1; EGFR: epidermal growth factor receptor; HSPG2: heparan sulfate proteoglycan 2; FLT1: Fms-related receptor tyrosine kinase 1; KDR: kinase insert domain receptor; SNCA: synuclein alpha; ABL1: ABL proto-oncogene 1, nonreceptor tyrosine kinase, NSCLC: nonsmall cell lung cancer, PC: pancreatic cancer, HBC: HER-positive breast cancer, RCC: renal cell carcinoma, STS: soft-tissue sarcoma, HC: hepatocellular carcinoma, GIST: gastrointestinal tumors, MTC: medullary thyroid cancer, AML: acute myelogenous leukemia, and CML: chronic myelogenous leukemia.

3.4. Computational Validation of Candidate Repurposed Drugs. The drug-gene validation network was constructed using the STITCH database (Figure 2C) and analyzed using Cytoscape software, and drugs were ranked based on the degree centrality and betweenness values. The results shown in Table 2 have indicated that the known anticancer drugs, dasatinib and bosutinib, were the hub nodes

among known neuroprotective anticancer drugs with the highest value of degree centrality of 4.0 and betweenness values of 0.007 and 0.004, respectively. Similarly, nintedanib, sunitinib, and vandetanib were identified as the important hub nodes among promising drug candidates with a degree centrality of 5.0 and betweenness values of 0.026, 0.021, and 0.011, respectively. We also identified the interactive targets of the topologically important drugs. The most considerable node nintedanib had a strong relationship with the genes KDR, FLT1, GSK3B, cyclin-dependent kinase 4 (CDK4), and ABL proto-oncogene 1 (ABL1). Similarly, sunitinib interacted on the validation network with FLT1, KDR, EGFR, CDK6, and ABL1, while vandetanib had close interactions with ABL1, EGFR, KDR, and FLT1.

3.5. Functional and Structural Analysis Validated the Repurposing Potential of Candidate Drugs. The potential repurposing candidates from the previous steps were evaluated for their functional and structural properties by the CoDRes tool. The tool is based on a disease-specific approach to compare drug–disease relationships concerning a training set of drugs approved or investigated for a disease. We have incorporated this tool to rerank the candidate drugs based on their repurposing scores. The comparative values for different drugs have been provided in (Table S3). Figure 4A–C has illustrated the comparative functional, structural, and CoDRes scores of the candidate drugs, respectively. The values have suggested that most of the drugs have good structural scores, but functional scores have shown significant variations. We found that erlotinib had the highest functional score (1.0), while dacomitinib had the lowest value (0.001). Similarly, sunitinib, sorafenib, imatinib, gefitinib, vandetanib, lenvatinib, pazopanib, axitinib, afatinib, and dacomitinib had the highest values (1.0) in terms of structural score, and lapatinib had the lowest score (0.33). Moreover, erlotinib had the highest CoDRes value (1.0), and lapatinib had the lowest value (0.20). We have selected the top 10 drugs with the highest CoDRes scores for further study. The CoDRes results have indicated that erlotinib would be a good repurposing drug having the highest functional and structural scores.

Additionally, we exploited the ChemMine server to investigate anti-Alzheimer's properties of candidate drugs and compared their clinical potential with donepezil, rivastigmine, galantamine, and memantine. The hierarchical clustering was performed using a clustering threshold of 1. We noticed no drug clusters with typical anti-Alzheimer drugs. We have selected the closest neighbors to Donepezil such as vandetanib, gefitinib, erlotinib, imatinib, afatinib, and sunitinib. Similarly, for another anti-Alzheimer drug rivastigmine, we found sunitinib as the closest match. Likewise, for galantamine, we found vandetanib, erlotinib, and gefitinib as the closest neighbors. We have found no nearest neighbor to memantine. The results are presented in Table 3. The best candidates obtained from clustering analysis have also demonstrated good structural similarity values, as highlighted in red in the table. Finally, we have selected 6 out of 10 drugs for supplementary analysis. The clustered groups were represented in the form of a heat map, as shown in Figure 4D.

3.6. Literature Studies and ADMET Analysis Evaluated the Neuroprotective Potential of Repurposed Drugs. To further validate our results, we have searched for the available information regarding the neuroprotective properties of the drugs proposed from the previous steps. A few bibliographic studies were available regarding neuro-

Table 2. Topological Parameters of Drugs on the Validation Network^a

Rank	Drug name	Degree	Betweenness	Topological Coefficient
1	Nintedanib	5	0.026	0.44
2	Sunitinib	5	0.021	0.402
3	Vandetanib	5	0.011	0.482
4	Dasatinib	4	0.007	0.502
5	Erlotinib	4	0.007	0.502
5	Imatinib	4	0.006	0.548
6	Ponatinib	4	0.006	0.543
6	Bosutinib	4	0.004	0.513
7	Axitinib	3	0.002	0.679
7	Sorafenib	3	0.002	0.679
7	Midostaurin	3	0.002	0.597
8	Pazopanib	2	0	0.875
8	Afatinib	2	0	0.791
8	Dacomitinib	2	0	0.791
8	Gefitinib	2	0	0.791
8	Lapatinib	2	0	0.791
8	Neratinib	2	0	0.791
9	Cabozantinib	1	0	0
9	Lenvatinib	1	0	0
9	Nilotinib	1	0	0

^aPromising drugs with the highest ranks are highlighted in pink, and known neuroprotective anticancer drugs are highlighted in green.

protective functions of anticancer drugs, as summarized in Table 4. Based on these results, we confirmed that all six drugs have repurposing potential for AD. ADMET analysis of the six drugs has confirmed that four drugs (erlotinib, gefitinib, vandetanib, and sunitinib) have good physicochemical properties (molecular weight, no of rotatable bonds, no of H-bond donors, no of H-bond acceptors, TPSA, and M log P) and were able to cross the BBB, as shown in (Table S4). Two drugs, afatinib and imatinib, would not be able to cross the BBB and thus were excluded from the study.

3.7. Functional Similarity Analysis with MicroRNAs.

To further validate our results, we extracted the list of AD-related miRNAs and also searched for the miRNAs related to the repurposed drugs (Table S5). After comparison, we found that erlotinib and gefitinib shared three miRNAs with AD where only one miRNA has neuroprotective functions, while vandetanib shared 33 different miRNAs with AD, as shown in the network in Figure 5. Of the 33 miRNAs, 11 miRNAs have neuroprotective functions. We found that miRNA-200a is the only AD-related miRNA with a neuroprotective function associated with all three drugs. miRNA-200a targets the EGFR gene, and a literature survey has confirmed its neuroprotective role in attenuating amyloid-beta overproduction by down-regulating BACE1 expression and tau hyperphosphorylation by reducing the expression of protein kinase A (PKA).⁶⁵

3.8. Pathway Analysis Confirmed the Repurposing Potential of EGFR Inhibitors. The significant AD-related genes were searched in the DisGeNET database to develop an expression pattern among AD and various types of cancers. The results are presented in the form of a heat map shown in Table 5 where the blue color represents high expression values,

while the red color represents low expression values. We found that CCND1, EGFR, and KDR are among the top genes which are commonly expressed in AD and in a different type of cancer. Furthermore, the experimentally significant gene interactions obtained from the STRING database were considered for pathway analysis by the Enrichr tool. We used KEGG, BioPlanet, and WikiPathway databases for pathway analysis (Table 6).

The most frequently appeared genes in the enriched pathways (biologically relevant) were the EGFR, JUN, and GSK3B. The ERBB signaling pathway, focal adhesion, mitogen-activated protein kinase (MAPK) signaling, Cu homeostasis, and phosphatidylinositol-3-kinase (PI3-Akt) pathways were the top signaling pathways associated with AD pathogenesis. There were many pieces of evidence available for the pathways identified by our study with AD. The pathological role of ErbB4 activity in AD is confirmed by Woo *et al.*, where ErbB4 was accompanied by AD progression.⁶⁶ The role of focal adhesion signaling in AD pathology is established because A β upregulates many proteins related to focal adhesion signaling that induce re-entry of neurons into the cell cycle.⁶⁷ Aberrant activation of focal adhesion kinases is associated with synaptic loss and neuronal dystrophy in AD.⁶⁸ Many studies have proposed that MAPK signaling plays an essential role in AD pathogenesis by regulating tau phosphorylation, APP processing, and neuronal apoptosis.⁶⁹ Several MAPKs interact with AD-related proteins such as tau, APP, presenilin (PS), and apolipoprotein E (ApoE).⁷⁰ The role of Cu in AD pathogenesis is controversial. Some studies have demonstrated that Cu overload is responsible for neurotoxicity in AD brains, while other studies

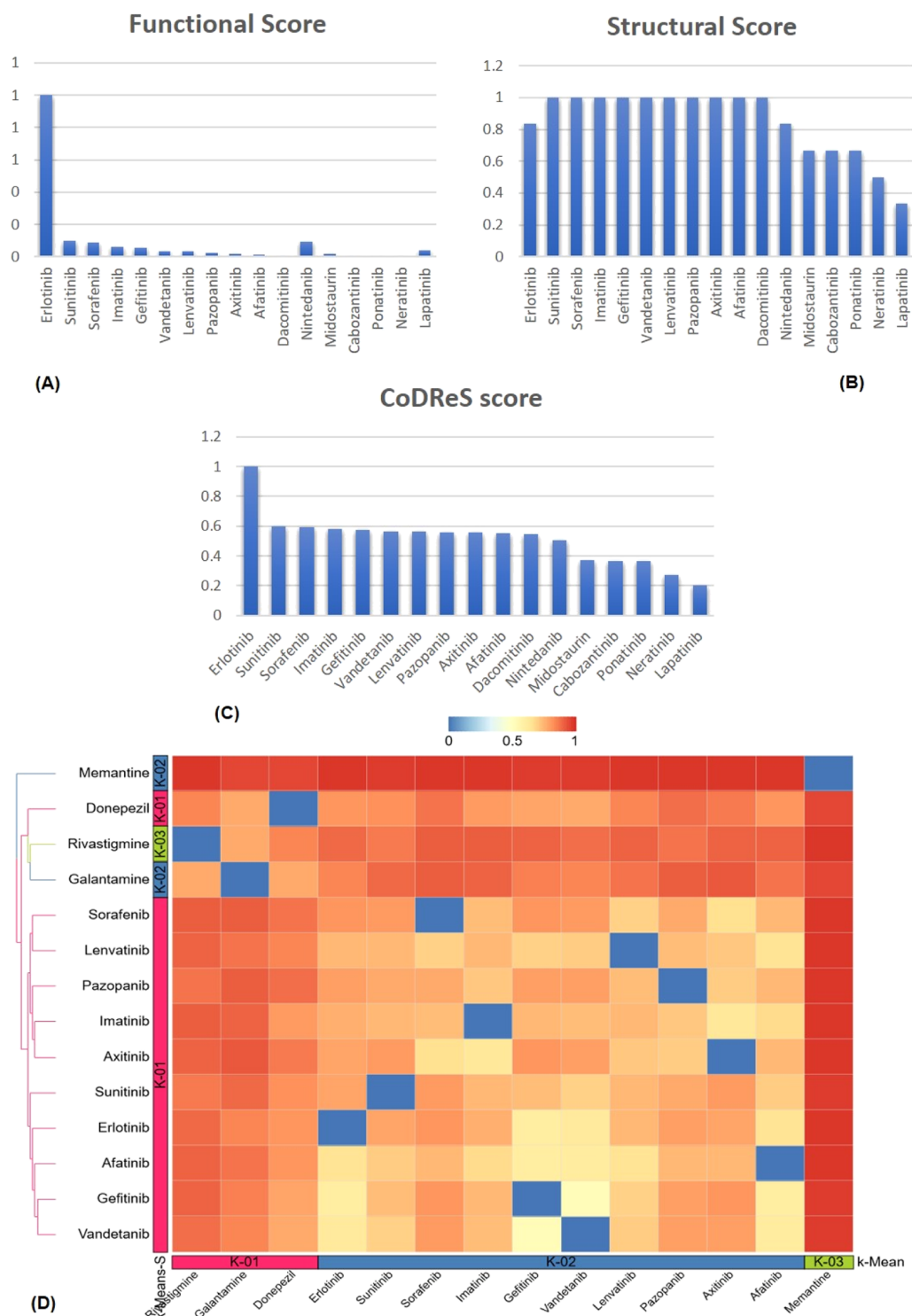


Figure 4. (A) Functional scores of different candidate repurposing drugs as calculated using the CoDReS tool. (B) Structural scores of different candidate repurposing drugs as calculated using the CoDReS tool. (C) CoDReS scores of candidate repurposing drugs. Erlotinib is shown as the most promising repurposing drug with good structural and functional scores. The structural scores of the drugs are more or less similar, while the functional scores have shown great variations. (D) Clustered heat map of candidate repurposing drugs with known Alzheimer's drugs donepezil, rivastigmine, galantamine, and memantine. The heat map is generated using a distance matrix as the input generated by subtracting the similarity coefficient from 1. The colors from blue to red represent the correlation intensities of drugs where blue represents complete correlation and red represents no correlation.

Table 3. Similarity Scores (Tanimoto Coefficient) of Repurposed Drugs with Known Alzheimer's Drugs^a

Drug	Donepezil	Rivastigmine	Galantamine	Memantine
Afatinib	0.176	0.086	0.112	0.009
Axitinib	0.128	0.081	0.067	0
Erlotinib	0.175	0.096	0.147	0.004
Gefitinib	0.207	0.084	0.138	0.014
Imatinib	0.182	0.076	0.881	0.008
Lenvatinib	0.149	0.086	0.112	0.003
Pazopanib	0.104	0.119	0.076	0.001
Sorafenib	0.119	0.070	0.072	0
Sunitinib	0.166	0.129	0.100	0.012
Vandetanib	0.218	0.105	0.149	0.017

^aHighlighted drugs have more or less similar scores to known AD drugs.

Table 4. Literature Studies for Neuroprotective Functions of Potential Repurposing Candidates

drug	neuroprotective function	references
afatinib	inhibition of oxygen/glucose-induced neuroinflammation and EGFR activation	56
erlotinib	reduction in A β -induced memory loss in AD	57
gefitinib	improvement in cognition and memory functions	57
	may improve AD pathogenesis by inhibiting the β -secretase activity	58
imatinib	inhibition of A β accumulation by the selective inhibition of BACE activity	59
	promotes degradation of A β by inducing the activity of A β -degrading enzyme neprilysin	60
	inhibition of brain c-Abl, reduction in circulating levels of A β , shifts APP processing to non-amyloidogenic pathway	61
sunitinib	provides neuroprotection by inhibiting NO production	62
	inhibition of acetylcholinesterase activity and attenuation of cognitive impairments in scopolamine-induced AD mice	63
vandetanib	may inhibit acetylcholinesterase activity in AD	64

have proposed Cu deficiency as a contributing factor to AD pathogenesis.⁷¹ Likewise, the role of the PI3K pathway is confirmed by studies where abnormal activities of the pathway were responsible for A β production and sequestration.⁷² The PI3K pathway activation has therapeutic potential to treat AD as some of the drugs such as donepezil, coenzyme Q10, and human telomerase reverse transcriptase (hTERT) are known to treat AD by GSK3B inhibition and PI3K activation.⁷³

DisGeNET and OMIM databases were used to find the most closely associated diseases with the identified genes (Table 7). The DisGeNET results reported that out of 15 genes, 13 genes were associated with AD (P -value 7.44×10^{-12}), while OMIM disease analysis identified 3 genes (P -value 5.77×10^{-5}) related to AD. Functional classification of identified genes from STRING interactions and their associated drugs retrieved from the STITCH network has revealed that kinases and their inhibitors are the major class of targets and targeted drugs associated with AD, respectively (Figure 6).

4. DISCUSSION

Drug repurposing is a productive approach to identify novel therapeutic uses of available drugs. The common biological pathways of different diseases and the advancements in system

biology tools open up new horizons to analyze the off-target effects of approved drugs for various indications. Over the last decade, several studies have been published, emphasizing the shared molecular mechanism of cancer and AD. Indeed, drug repurposing of anticancer drugs as neuroprotective agents has been applied to overcome AD-related clinical consequences. However, the complexity of different neuropathological states and limited understanding of different cellular signaling mechanisms in AD posed a big challenge to develop repurpose therapeutics. In the present study, we used an integrated approach to reveal potential AD-related targets. We opted for a comprehensive data analysis approach to identify neuroprotective anticancer drugs and analyzed the data with network-based and pathway-based tools. We identified 49 AD-related genes by combining GWAS, transcriptomics, and metabolomics studies. We reported 17 cancer-related genes that have direct interactions with the identified AD-related targets. We identified 36 approved anticancer drugs that have associations with these targeting genes. For further study, we selected the experimentally significant genes with the highest interaction scores, as shown in the STRING network. We found 30 anticancer drugs as respective targets of the experimentally significant genes.

Computational validation by CoDReS ranked the repurposing drugs based on their functional and structural properties. Among the proposed drugs, dasatinib (phase I/II), nilotinib (phase II), and bosutinib (phase I) are in clinical trials as repurposed therapeutics for AD, thus validating the authenticity of our drug repurposing approach. The top 10 drugs obtained from CoDReS scoring were analyzed for their similarities with the known AD drugs and clustered based on their similarity scores. We selected the closest neighbors, vandetanib, erlotinib, gefitinib, afatinib, imatinib, and sunitinib. The literature studies have confirmed the repurposing potential of these anticancer drugs. The ADMET analysis of these six drugs revealed that afatinib and imatinib did not possess good physicochemical properties and were not BBB-penetrant. Thus, we proposed vandetanib, erlotinib, gefitinib, and sunitinib as potential repurposing drugs.

The pathway analysis identified the EGFR and GSK3B as the most frequently appeared genes in AD-associated pathways. The CCND1, EGFR, and KDR are found as the most commonly expressed genes in AD and in 13 most common types of cancers. Network analysis of PPI interactions revealed that GSK3B, KDR, APP, EGFR, and FLT1 were the hub genes

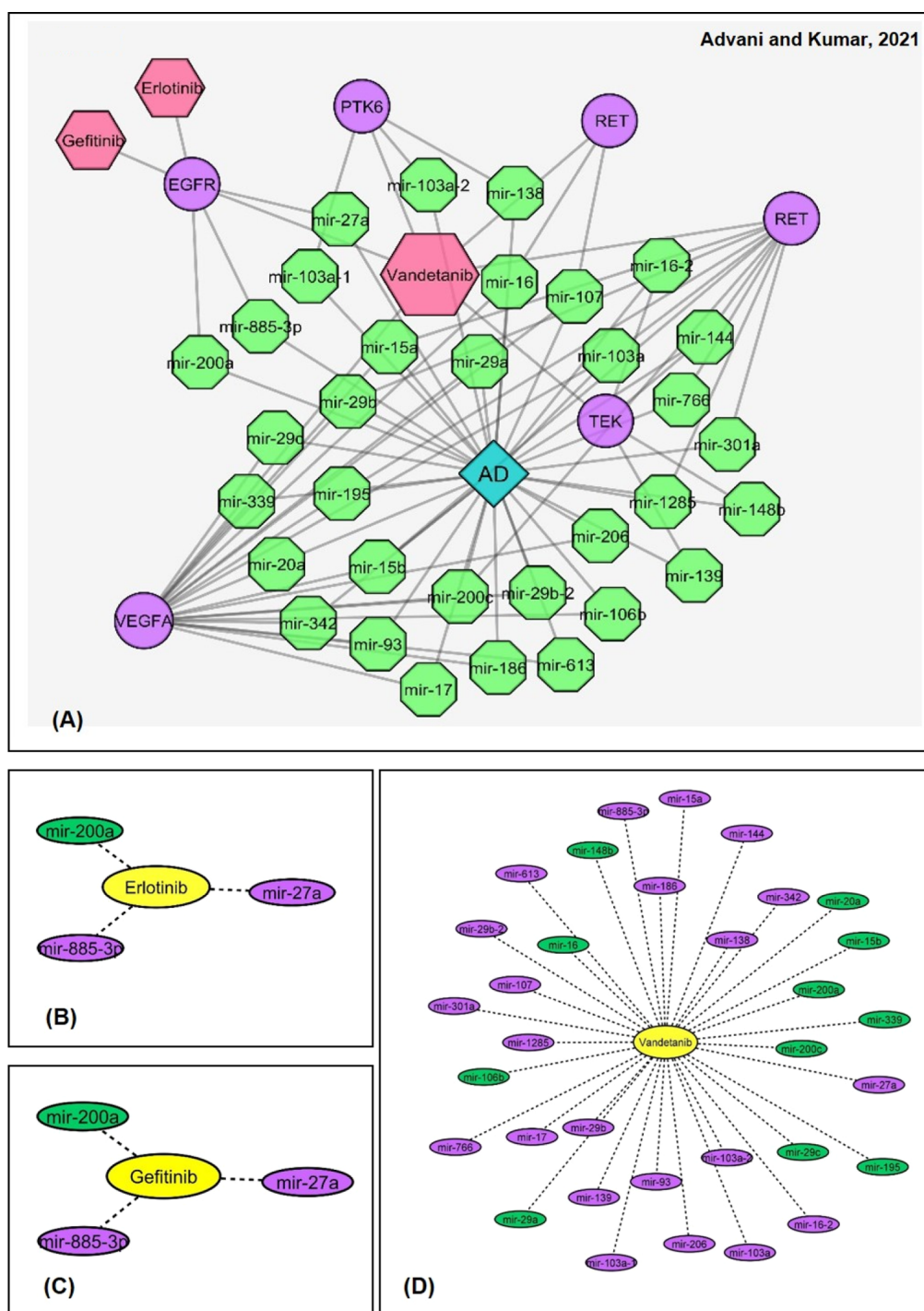


Figure 5. (A) Network is showing the interrelationship of miRNAs associated with AD and those associated with repurposed anticancer drugs erlotinib, gefitinib, and vandetanib. The network shows that vandetanib shares many common targets such as EGFR, PTK6, RET, TEK, and VEGFA with AD-related miRNAs, while both erlotinib and gefitinib share functional similarity through the EGFR gene. (B–D) Association of erlotinib, gefitinib, and vandetanib with miRNAs, respectively, where miRNAs shown in green are neuroprotective, while miRNAs shown in purple are neurodegenerative as identified through literature analysis. miRNA-200a is the only one that shows association with all three repurposed drugs.

in the PPI network. Literature studies have supported the neuroprotective potential of these targets and their associated drugs. In short, our integrated omics analysis with computational validation tools had prioritized the role of GSK3B and EGFR in AD pathogenesis. ErBb signaling, focal adhesion, MAPK pathway, Cu homeostasis, and PI3-Akt were the over-

representative pathways targeted by these genes that we prioritized by pathway analysis using different databases. However, the therapeutic relevance of targeting the EGFR in AD is not well established. Still, some studies have supported the fact that the EGFR prevents A β and ApoE-induced cognitive deficits and considered a preferred target for treating

Table 5. Heat Map Showing the Expression Pattern of Shared Genes between AD and 13 Most Common Cancer Types^a

	ABL1	A2M	BIN1	CCND1	ERBB4	EGFR	FLT1	GSK3B	HSPG2	JUN	KDR	LRP1	MAPT	NRG1	SNCA
AD															
Bladder cancer															
Breast cancer															
Colorectal cancer															
Endometrial cancer															
Kidney cancer															
Leukemia															
Liver cancer															
Lung cancer															
Melanoma															
NHL															
Pancreatic cancer															
Prostatic cancer															
Thyroid cancer															

^aAD: Alzheimer's disease; NHL: non-Hodgkin lymphoma.Table 6. Pathway Analysis of STRING Interactions Based on *p*-Values^a

S.No.	Pathway name	Genes involved	P-value ^{bb}
KEGG pathway analysis			
1	ErBb signaling pathway	GSK3B, JUN, ERBB4, ABL1, NRG1, EGFR	2.39E-11
2	Focal adhesion	GSK3B, JUN, FLT1, CCND1, KDR, EGFR	4.18E-11
3	MAPK signaling pathway	MAPT, JUN, FLT1, ERBB4, KDR, EGFR	4.38E-11
BioPlanet pathway analysis			
1	ErBb signaling pathway	GSK3B, JUN, CCND1, ERBB4, ABL1, NRG1, EGFR	2.52E-13
2	Focal adhesion	GSK3B, JUN, CCND1, FLT1, KDR, EGFR	1.08E-08
3	PI3-Akt pathway	GSK3B, ERBB4, NRG1, EGFR	2.73E-08
WikiPathway analysis			
1	ErBb signaling pathway	GSK3B, JUN, CCND1, ERBB4, ABL1, NRG1, EGFR	1.99E-13
2	Cu homeostasis	APP, GSK3B, JUN, CCND1, MAPT	2.87E-10
3	Focal adhesion	JUN, GSK3B, FLT1, CCND1, KDR, EGFR	4.06E-09

^aGenes in red are the most frequently appeared genes in the enriched pathways. ^bHere, the *p*-value represents the probability of any gene belonging to a biological pathway.Table 7. Disease-Based Analysis of STRING Interactions Based on *p*-Values

S.No.	Disease name	Genes involved	P-value ^a
DisGeNET analysis			
1	Amyloidosis	APP, BIN1, EGFR, ERBB4, FLT1, GSK3B, HSPG2, LRP1, MAPT, NRG1, SNCA	7.19E-13
2	Melanoma	ABL1, APP, BIN1, CCND1, EGFR, ERBB4, FLT1, GSK3B, HSPG2, JUN, KDR, LRP1, NRG1, SNCA	2.26E-12
3	Alzheimer's Disease	ABL1, APP, BIN1, CCND1, EGFR, ERBB4, GSK3B, HSPG2, JUN, LRP1, MAPT, NRG1, SNCA	7.44E-12
4	Central Neuroblastoma	APP, BIN1, CCND1, EGFR, ERBB4, FLT1, GSK3B, JUN, KDR, LRP1, MAPT, SNCA	3.57E-11
5	Non-small cancer lung carcinoma	ABL1, APP, BIN1, CCND1, EGFR, ERBB4, FLT1, GSK3B, JUN, KDR, LRP1, NRG1, SNCA	3.63E-11
OMIM disease analysis			
1	Dementia	APP, CCND1, EGFR, MAPT, SNCA	4.52E-09
2	Parkinson's Disease	CCND1, EGFR, MAPT, SNCA	7.39E-07
3	Alzheimer's Disease	APP, CCND1, EGFR	5.77E-05
4	Schizophrenia	CCND1, EGFR, NRG1	6.46E-05
5	Myopathy	BIN1, CCND1, EGFR	9.09E-05

^aHere, the *p*-value represents the probability of any gene belonging to a biological disease.

AD.^{57,74} We also established a new connection of the EGFR with AD-related targets such as APP, SNCA, LRP1, and NRG.

Many bibliographic mentions also supported this finding. A recently published study has identified that APP-EGFR

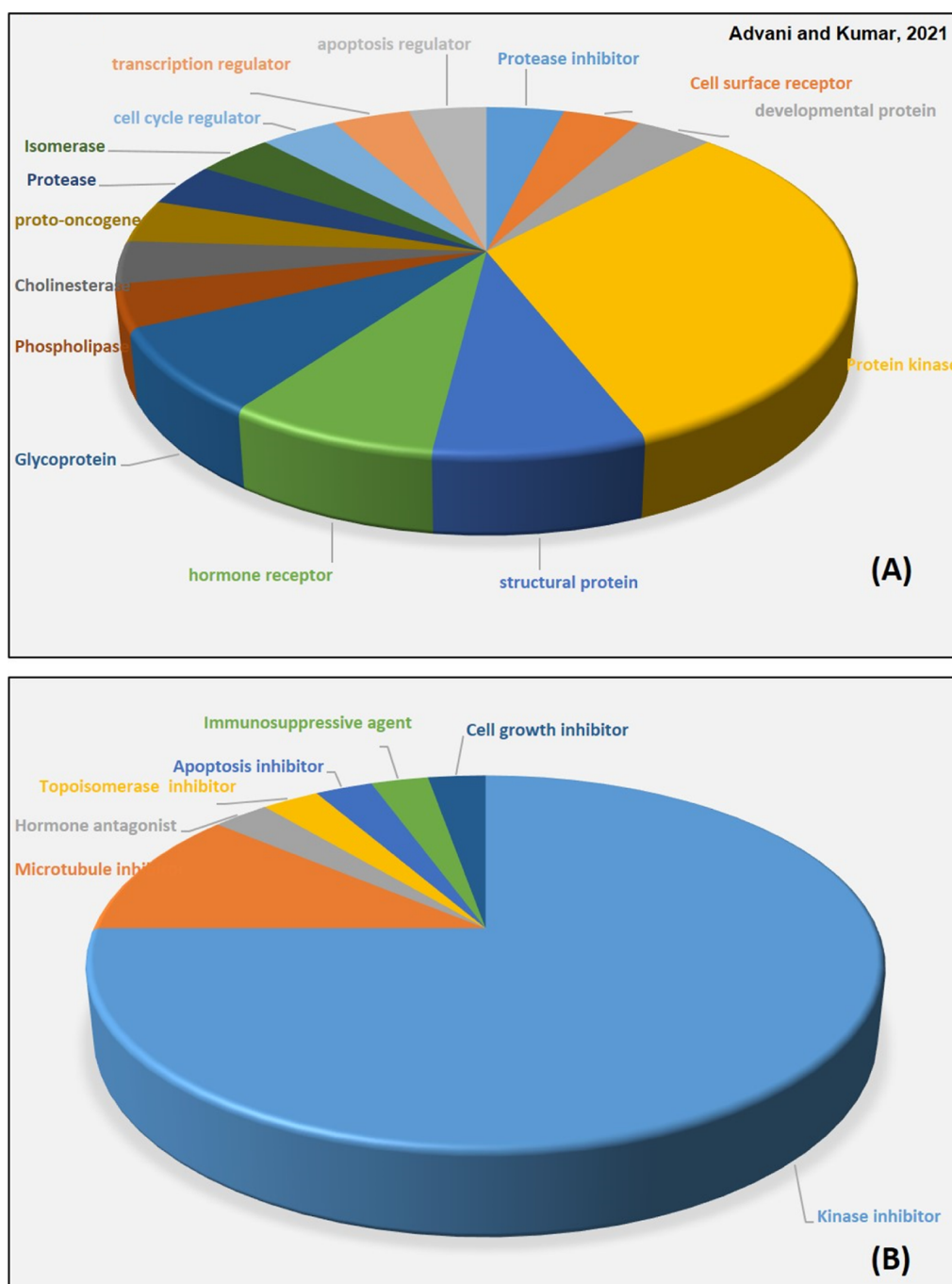


Figure 6. (A) Figure showing the functional categories of AD-related genes/PPI genes. The relative area of each segment corresponds to the relative fraction of a particular target class. As shown, protein kinases represent the major functional target protein class. (B) Functional classification of candidate repurposing anticancer drugs for AD. As expected, kinase inhibitors are the most prevalent drugs having neuroprotective functions.

interaction promoted extracellular signal-regulated kinase (ERK) signaling and contributed to neuritogenesis and neuronal differentiation.⁷⁵ Some studies have reported that the EGFR has structural and expression similarities with ErbB4, the primary receptor of NRG1, in several brain regions. Some studies have found that the EGFR was coexpressed with ErbB4 in several GABAergic neurons.^{76,77} This finding would be helpful to establish new connections of EGFR inhibitors with NRG1. Although the role of the EGFR in SNCA gene

polymorphisms in AD brains is not explored, a study by Yan *et al.* confirmed that SNCA plays a significant role in EGFR signaling in lung adenocarcinoma cells.⁷⁸

Our proposed repurposed drug list had three EGFR inhibitors—vandetanib, erlotinib, and gefitinib. Among the proposed drugs, vandetanib, a tyrosine kinase inhibitor, is currently marketed to treat tumors of the thyroid gland. Likewise, erlotinib, an EGFR inhibitor, is used for treating nonsmall cell lung cancer (NSCLC) and pancreatic cancer.

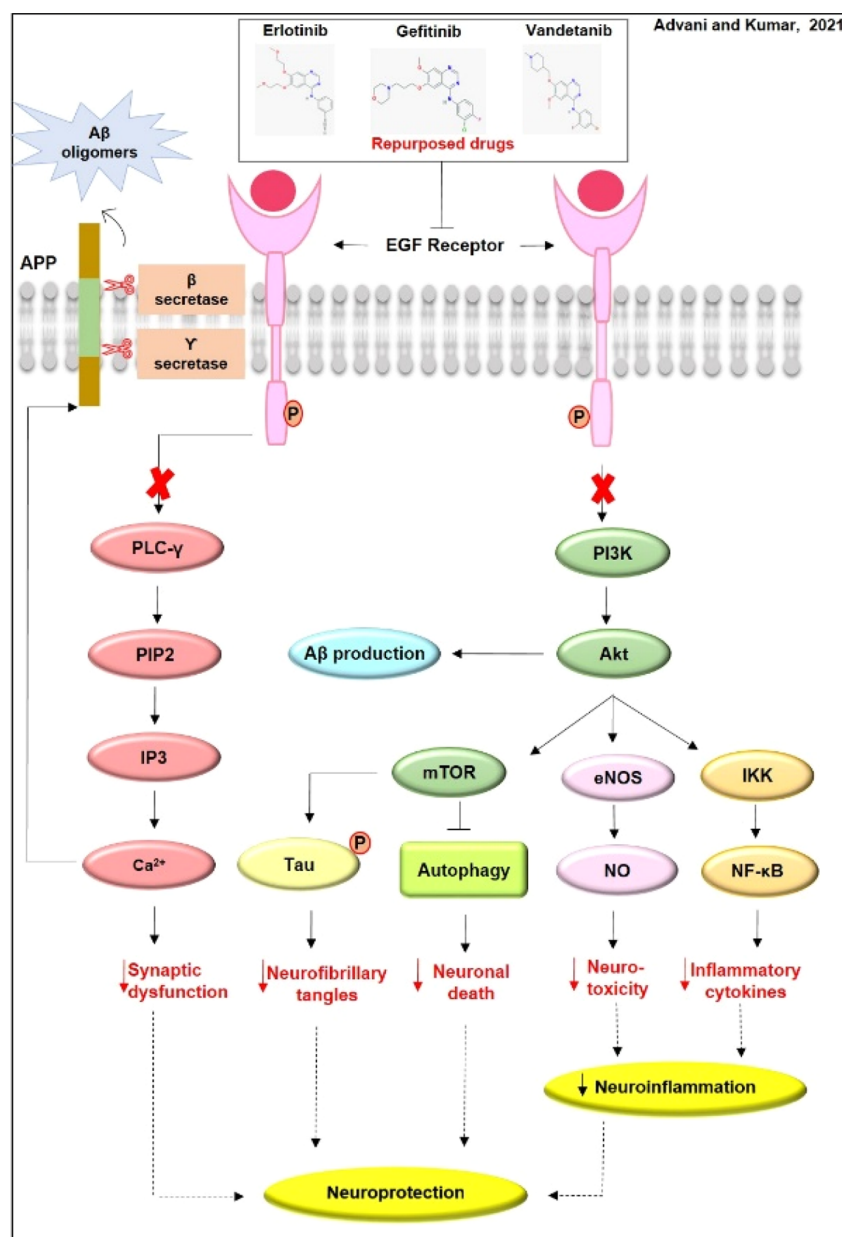


Figure 7. Schematic representation of the proposed mechanism of neuroprotective functions of EGFR inhibitors in AD. The binding of a ligand to the EGFR causes conformational changes in the receptor and activates various signaling cascades. Activation of the PI3K/Akt axis activates mTOR that is a major inhibitor of the autophagic process. The inhibition of autophagy leads to neuronal death. Activated Akt further induces endothelial nitric oxide synthase (eNOS) that generates nitric oxide (NO), a neurotoxin. The activated Akt instigates inflammatory cytokine production by inducing NF-κB production. The activated EGFR induces Ca²⁺ release from the endoplasmic reticulum by inducing phospholipase C gamma (PLC-γ) production. Excessive release of Ca²⁺ causes synaptic dysfunction and Aβ production from APP. All the events trigger neuroinflammation and neurodegeneration. Pharmacological inhibition of the EGFR by inhibitors, erlotinib, gefitinib, and vandetanib, may reverse the downstream signaling cascades of the EGFR and provide neuroprotection, a reduction in synaptic dysfunction, reduced tau phosphorylation, inhibition of neuronal death, and inhibition of neuroinflammatory processes. Dotted arrows represent the proposed neuroprotective functions of the repurposed drugs.

Similarly, gefitinib, an inhibitor of EGFR tyrosine kinase, is approved to treat locally advanced or metastatic NSCLC. Structural similarities of these drugs with approved AD drugs and physicochemical and BBB analyses also supported the therapeutic potential of these drugs. Earlier studies have proposed that erlotinib and gefitinib rescued EGFR-induced Aβ toxicity and memory loss in *Drosophila* and mouse

models,⁵⁷ but the exact molecular mechanism and affected signaling pathways are yet to be elucidated.

Furthermore, some recent computational studies have predicted the potential drug–disease relations based on miRNA data. Based on this fact, we searched for miRNAs that were related to AD and correlated the gene targets of these miRNAs with the gene targets of the proposed

repurposed drugs. From this analysis, we identified some neuroprotective microRNAs and established their relationship with the repurposed drugs. We identified miRNA-200a as a potential neuroprotective candidate that shares targets with all three repurposed EGFR inhibitors. In such a way, miRNA–disease–drug relations helped us to establish a link between repurposed drugs and AD concerning the miRNA axis.

To find out the significance of the results, we curated the available literature and proposed the potential neuroprotective functions of the repurposing drugs in AD pathogenesis, as shown in Figure 7. We suggested that tau phosphorylation, autophagy, and neuroinflammation were the significant AD-related biological mechanisms regulated by the proposed EGFR inhibitor drugs. PI3-Akt signaling, NF-kappa B pathway, and Ca²⁺ signaling were the significant pathways targeted by the proposed drugs.

5. CONCLUSIONS

Repurposed drugs can be a promising way of treating complex diseases such as AD. Our study has proposed an integrated omics-based data mining approach to identify the possible relationship of anticancer drugs with AD-associated genes. We further integrated network-based and pathway-based analysis methods to validate the overlap of anticancer drugs with AD-related pathways. The resulting drugs were validated based on computational repurposing tools, similarity scores, and physicochemical analysis. Additionally, literature validation, the functional similarity with miRNAs, and pathway analysis supported the hypothesis that EGFR inhibitors vandetanib, erlotinib, and gefitinib might play therapeutic roles by targeting AD-related proteins. Furthermore, we elucidated the mechanistic basis of these drugs in ameliorating AD-associated neurotoxicity and neuroinflammation. Additionally, our comprehensive approach also proposed a connection between AD-related targets and the reported repurposing drugs. As far as experimental aspects are concerned, *in vitro* and animal studies are warranted to confirm their neuroprotective potential.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acsomega.1c01526>.

Functional enrichment analysis of AD-associated genes, list of candidate repurposing anticancer drugs, computational drug repositioning scores, physiochemical properties of repurposed drugs, and AD-related miRNAs, drugs, and targets (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Pravir Kumar – Molecular Neuroscience and Functional Genomics Laboratory, Delhi Technological University, Delhi 110042, India; orcid.org/0000-0001-7444-2344; Phone: +91-9818898622; Email: pravirkumar@dtu.ac.in, kpravir@gmail.com

Author

Dia Advani – Molecular Neuroscience and Functional Genomics Laboratory, Delhi Technological University, Delhi 110042, India

Complete contact information is available at:

<https://pubs.acs.org/doi/10.1021/acsomega.1c01526>

Author Contributions

P.K. conceived and designed the manuscript. D.A. collected and analyzed data. D.A. and P.K. wrote the manuscript, discussed the results, and analyzed the entire data.

Funding

D.A. has received Senior Research Fellowship (SRF) by Department of Biotechnology (DBT), Govt. of India (Fellow ID: DBT/2018/DTU/1067).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to thank the senior management of Delhi Technological University (DTU) and the Department of Biotechnology (DBT), Indian Government, for their constant support and financial assistance.

■ ABBREVIATIONS

AD, Alzheimer's disease; ABL1, ABL proto-oncogene 1; A β , amyloid- β ; ApoE, apolipoprotein E; APP, amyloid precursor protein; BBB, blood–brain barrier; BACE-1, beta-secretase; CDK4, cyclin-dependent kinase 4; CNS, central nervous system; CoDReS, computational drug repositioning score; EGFR, epidermal growth factor receptor; ERK, extracellular signal-regulated kinase; FDR, false discovery rate; FLT1, Fms-related receptor tyrosine kinase 1; GEO, Gene Expression Omnibus; GS2D, gene set to diseases; GWAS, genome-wide association studies; GSK3B, glycogen synthase kinase 3 beta; HMDB, Human Metabolome Database; HMDD, Human microRNA Disease Database; hTERT, human telomerase reverse transcriptase; IL-6, interleukin-6; KDR, kinase insert domain receptor; LRP1, LDL receptor-related protein 1; MAPK, mitogen-activated protein kinase; MCS, maximum common substructure; MIST, Molecular Interaction Search Tool; *M log P*, partition coefficient; NIH, National Cancer Institute; NRG1, neuregulin 1; NSCLC, nonsmall cell lung cancer; OR, odds ratio; PI3K-Akt, phosphatidylinositol-3-kinase; PPI, protein–protein interaction; PKA, protein kinase A; PS, presenilin; RAGE, receptors for advanced glycation end products; SNCA, alpha-synuclein; SNP, single-nucleotide polymorphism; STITCH, search tool for interactions of chemicals; TNF- α , tumor necrosis factor- α ; TPSA, topological polar surface area; TTD, Therapeutic Target Database

■ REFERENCES

- (1) Wang, J.; Gu, B. J.; Masters, C. L.; Wang, Y.-J. A systemic view of Alzheimer disease - insights from amyloid- β metabolism beyond the brain. *Nat. Rev. Neurol.* **2017**, *13*, 612.
- (2) Zhang, M.; Schmitt-Ulms, G.; Sato, C.; Xi, Z.; Zhang, Y.; Zhou, Y.; St George-Hyslop, P.; Rogaeva, E. Drug Repositioning for Alzheimer's Disease Based on Systematic "omics" Data Mining. *PLoS One* **2016**, *11*, No. e0168812.
- (3) Durães, F.; Pinto, M.; Sousa, E. Old Drugs as New Treatments for Neurodegenerative Diseases. *Pharmaceuticals* **2018**, *11*, 44.
- (4) Lanford, R. E.; Hildebrandt-Eriksen, E. S.; Petri, A.; Persson, R.; Lindow, M.; Munk, M. E.; Kauppinen, S.; Ørum, H. Therapeutic Silencing of MicroRNA-122 in Primates with Chronic Hepatitis C Virus Infection. *Science* **2010**, *327*, 198.

- (5) Yu, L.; Zhao, J.; Gao, L. Predicting Potential Drugs for Breast Cancer Based on MiRNA and Tissue Specificity. *Int. J. Biol. Sci.* **2018**, *14*, 971.
- (6) Aydin, B.; Arslan, S.; Bayraklı, F.; Karademir, B.; Arga, K. Y. MiRNA-Mediated Drug Repurposing Unveiled Potential Candidate Drugs for Prolactinoma Treatment. *Neuroendocrinology* **2021**, DOI: 10.1159/000515801.
- (7) Ashburn, T. T.; Thor, K. B. Drug Repositioning: Identifying and Developing New Uses for Existing Drugs. *Nat. Rev. Drug Discovery* **2004**, *3*, 673.
- (8) Rudrapal, M.; J. Khairnar, S.; G. Jadhav, A. Drug Repurposing (DR): An Emerging Approach in Drug Discovery. *Drug Repurposing—Hypothesis, Molecular Aspects and Therapeutic Applications*; IntechOpen, 2020.
- (9) Karatzas, E.; Bourdakou, M. M.; Kolios, G.; Spyrou, G. M. Drug Repurposing in Idiopathic Pulmonary Fibrosis Filtered by a Bioinformatics-Derived Composite Score. *Sci. Rep.* **2017**, *7*, 12569.
- (10) Pushpakom, S.; Iorio, F.; Eyers, P. A.; Escott, K. J.; Hopper, S.; Wells, A.; Doig, A.; Williams, T.; Latimer, J.; McNamee, C.; Norris, A.; Sanseau, P.; Cavalla, D.; Pirmohamed, M. Drug Repurposing: Progress, Challenges and Recommendations. *Nat. Rev. Drug Discovery* **2019**, *18*, 41.
- (11) Xue, H.; Li, J.; Xie, H.; Wang, Y. Review of Drug Repositioning Approaches and Resources. *Int. J. Biol. Sci.* **2018**, *14*, 1232.
- (12) Paananen, J.; Fortino, V. An Omics Perspective on Drug Target Discovery Platforms. *Briefings Bioinf.* **2019**, *21*, 1937.
- (13) Tam, V.; Patel, N.; Turcotte, M.; Bossé, Y.; Paré, G.; Meyre, D. Benefits and Limitations of Genome-Wide Association Studies. *Nat. Rev. Genet.* **2019**, *20*, 467.
- (14) Alexander-Dann, B.; Pruteanu, L. L.; Oerton, E.; Sharma, N.; Berindan-Neagoe, I.; Módos, D.; Bender, A. Developments in Toxicogenomics: Understanding and Predicting Compound-Induced Toxicity from Gene Expression Data. *Mol. Omics* **2018**, *14*, 218.
- (15) Pritchard, J.-L. E.; O'Mara, T. A.; Glubb, D. M. Enhancing the Promise of Drug Repositioning through Genetics. *Front. Pharmacol.* **2017**, *8*, 896.
- (16) Schirle, M.; Bantscheff, M.; Kuster, B. Mass Spectrometry-Based Proteomics in Preclinical Drug Discovery. *Chem. Biol.* **2012**, *19*, 72.
- (17) Patti, G. J.; Yanes, O.; Siuzdak, G. Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263.
- (18) Nudelman, K. N. H.; McDonald, B. C.; Lahiri, D. K.; Saykin, A. J. Biological Hallmarks of Cancer in Alzheimer's Disease. *Mol. Neurobiol.* **2019**, *56*, 7173.
- (19) Monacelli, F.; Cea, M.; Borghi, R.; Odetti, P.; Nencioni, A. Do Cancer Drugs Counteract Neurodegeneration? Repurposing for Alzheimer's Disease. *J. Alzheimer's Dis.* **2016**, *55*, 1295.
- (20) Lee, S. Y.; Song, M.-Y.; Kim, D.; Park, C.; Park, D. K.; Kim, D. G.; Yoo, J. S.; Kim, Y. H. A Proteotranscriptomic-Based Computational Drug-Repositioning Method for Alzheimer's Disease. *Front. Pharmacol.* **2020**, *10*, 1653.
- (21) Advani, D.; Gupta, R.; Tripathi, R.; Sharma, S.; Ambasta, R. K.; Kumar, P. Protective Role of Anticancer Drugs in Neurodegenerative Disorders: A Drug Repurposing Approach. *Neurochem. Int.* **2020**, *140*, 104841.
- (22) Lonskaya, I.; Hebron, M. L.; Selby, S. T.; Turner, R. S.; Moussa, C. E.-H. Nilotinib and Bosutinib Modulate Pre-Plaques Alterations of Blood Immune Markers and Neuro-Inflammation in Alzheimer's Disease Models. *Neuroscience* **2015**, *304*, 316.
- (23) Bibi, N.; Rizvi, S. M. D.; Batool, A.; Kamal, M. A. Inhibitory Mechanism of An Anticancer Drug, Bexarotene Against Amyloid β Peptide Aggregation: Repurposing Via Neuroinformatics Approach. *Curr. Pharm. Des.* **2019**, *25*, 2989.
- (24) Hayes, C. D.; Dey, D.; Palavicini, J. P.; Wang, H.; Patkar, K. A.; Minond, D.; Nefzi, A.; Lakshmana, M. K. Striking Reduction of Amyloid Plaque Burden in an Alzheimer's Mouse Model after Chronic Administration of Carmustine. *BMC Med.* **2013**, *11*, 81.
- (25) Shemesh, O. A.; Spira, M. E. Rescue of Neurons from Undergoing Hallmark Tau-Induced Alzheimer's Disease Cell Pathologies by the Antimitotic Drug Paclitaxel. *Neurobiol. Dis.* **2011**, *43*, 163.
- (26) Javidnia, M.; Hebron, M. L.; Xin, Y.; Kinney, N. G.; Moussa, C. E.-H. Pazopanib Reduces Phosphorylated Tau Levels and Alters Astrocytes in a Mouse Model of Tauopathy. *J. Alzheimer's Dis.* **2017**, *60*, 461.
- (27) Buniello, A.; MacArthur, J. A. L.; Cerezo, M.; Harris, L. W.; Hayhurst, J.; Malagone, C.; McMahon, A.; Morales, J.; Mountjoy, E.; Sollis, E.; Suveges, D.; Vrousou, O.; Whetzel, P. L.; Amode, R.; Guillen, J. A.; Riat, H. S.; Trevanion, S. J.; Hall, P.; Junkins, H.; Flicek, P.; Burdett, T.; Hindorf, L. A.; Cunningham, F.; Parkinson, H. The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019. *Nucleic Acids Res.* **2019**, *47*, D1005. <http://www.ebi.ac.uk/gwas>
- (28) Barrett, T.; Wilhite, S. E.; Ledoux, P.; Evangelista, C.; Kim, I. F.; Tomashevsky, M.; Marshall, K. A.; Phillippy, K. H.; Sherman, P. M.; Holko, M.; Yefanov, A.; Lee, H.; Zhang, N.; Robertson, C. L.; Serova, N.; Davis, S.; Soboleva, A. NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Res.* **2013**, *41*, D991. <http://www.ncbi.nlm.nih.gov/geo/>
- (29) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; Sayeeda, Z.; Lo, E.; Assempour, N.; Berjanskii, M.; Singhal, S.; Arndt, D.; Liang, Y.; Badran, H.; Grant, J.; Serra-Cayuela, A.; Liu, Y.; Mandal, R.; Neveu, V.; Pon, A.; Knox, C.; Wilson, M.; Manach, C.; Scalbert, A. HMDB 4.0: The Human Metabolome Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608. <http://www.hmdb.ca>
- (30) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* **2009**, *4*, 44. <https://david.ncifcrf.gov>
- (31) Andrade-Navarro, M. A.; Fontaine, J. F. Gene Set to Diseases (GS2D): Disease Enrichment Analysis on Human Gene Sets with Literature Data. *Genom. Comput. Biol.* **2016**, *2*, No. e33. <http://cbdm.uni-mainz.de/geneset2diseases>
- (32) Cavalla, D. Predictive Methods in Drug Repurposing: Gold Mine or Just a Bigger Haystack? *Drug Discov. Today* **2013**, *18*, 523–532.
- (33) Hu, Y.; Vinayagam, A.; Nand, A.; Comjean, A.; Chung, V.; Hao, T.; Mohr, S. E.; Perrimon, N. Molecular Interaction Search Tool (MIST): An Integrated Resource for Mining Gene and Protein Interaction Data. *Nucleic Acids Res.* **2018**, *46*, D567–D574. <http://fgtrtools.hms.harvard.edu/MIST/>
- (34) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maclejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074. www.drugbank.com
- (35) Wang, Y.; Zhang, S.; Li, F.; Zhou, Y.; Zhang, Y.; Wang, Z.; Zhang, R.; Zhu, J.; Ren, Y.; Tan, Y.; Qin, C.; Li, Y.; Li, X.; Chen, Y.; Zhu, F. Therapeutic Target Database 2020: Enriched Resource for Facilitating Research and Early Development of Targeted Therapeutics. *Nucleic Acids Res.* **2020**, *48*, D1031. <http://db.idrblab.net/ttd/>
- (36) Szklarczyk, D.; Gable, A. L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N. T.; Morris, J. H.; Bork, P.; Jensen, L. J.; Mering, C. v. STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Res.* **2019**, *47*, D607. string-db.org
- (37) Szklarczyk, D.; Santos, A.; von Mering, C.; Jensen, L. J.; Bork, P.; Kuhn, M. STITCH 5: Augmenting Protein-Chemical Interaction Networks with Tissue and Affinity Data. *Nucleic Acids Res.* **2016**, *44*, D380. <http://stitch.embl.de/>
- (38) Jadamba, E.; Shin, M. A Systematic Framework for Drug Repositioning from Integrated Omics and Drug Phenotype Profiles Using Pathway-Drug Network. *BioMed Res. Int.* **2016**, *2016*, 7147039.
- (39) Karatzas, E.; Minadakis, G.; Kolios, G.; Delis, A.; Spyrou, G. M. A Web Tool for Ranking Candidate Drugs Against a Selected Disease

Based on a Combination of Functional and Structural Criteria. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 939. <http://bioinformatics.cing.ac.cy/codres>

(40) Backman, T. W. H.; Cao, Y.; Girke, T. ChemMine Tools: An Online Service for Analyzing and Clustering Small Molecules. *Nucleic Acids Res.* **2011**, *39*, W486.

(41) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 3186.

(42) Daina, A.; Michielin, O.; Zoete, V. SwissADME: A Free Web Tool to Evaluate Pharmacokinetics, Drug-Likeness and Medicinal Chemistry Friendliness of Small Molecules. *Sci. Rep.* **2017**, *7*, 42717. <http://www.swissadme.ch/>

(43) Huang, Z.; Shi, J.; Gao, Y.; Cui, C.; Zhang, S.; Li, J.; Zhou, Y.; Cui, Q. HMDD v3.0: A Database for Experimentally Supported Human MicroRNA-Disease Associations. *Nucleic Acids Res.* **2019**, *47*, D1013–D1017. <https://www.cuilab.cn/hmdd>

(44) American Cancer Society. *Cancer Facts & Figures 2020*; American Cancer Society, 2020.

(45) Kanehisa, M.; Sato, Y.; Furumichi, M.; Morishima, K.; Tanabe, M. New Approach for Understanding Genome Variations in KEGG. *Nucleic Acids Res.* **2019**, *47*, D590.

(46) Huang, R.; Grishagin, I.; Wang, Y.; Zhao, T.; Greene, J.; Obenauer, J. C.; Ngan, D.; Nguyen, D.-T.; Guha, R.; Jadhav, A.; Southall, N.; Simeonov, A.; Austin, C. P. The NCATS BioPlanet - An Integrated Platform for Exploring the Universe of Cellular Signaling Pathways for Toxicology, Systems Biology, and Chemical Genomics. *Front. Pharmacol.* **2019**, *10*, 445.

(47) Slenter, D. N.; Kutmon, M.; Hanspers, K.; Riutta, A.; Windsor, J.; Nunes, N.; Mélius, J.; Cirillo, E.; Coort, S. L.; D'Igles, D.; Ehrhart, F.; Giesbertz, P.; Kalafati, M.; Martens, M.; Miller, R.; Nishida, K.; Rieswijk, L.; Waagmeester, A.; Eijssen, L. M. T.; Evelo, C. T.; Pico, A. R.; Willighagen, E. L. WikiPathways: A Multifaceted Pathway Database Bridging Metabolomics to Other Omics Research. *Nucleic Acids Res.* **2018**, *46*, D661.

(48) Kuleshov, M. V.; Jones, M. R.; Rouillard, A. D.; Fernandez, N. F.; Duan, Q.; Wang, Z.; Koplev, S.; Jenkins, S. L.; Jagodnik, K. M.; Lachmann, A.; McDermott, M. G.; Monteiro, C. D.; Gundersen, G. W.; Ma'ayan, A. Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update. *Nucleic Acids Res.* **2016**, *44*, W90. <http://amp.pharm.mssm.edu/Enrichr/>

(49) Llorens-Martin, M.; Jurado, J.; Hernández, F.; Ávila, J. GSK-3 β , a Pivotal Kinase in Alzheimer Disease. *Front. Mol. Neurosci.* **2014**, *7*, 46.

(50) Takashima, A. GSK-3 Is Essential in the Pathogenesis of Alzheimer's Disease. *J. Alzheimer's Dis.* **2006**, *9*, 309.

(51) Müller, U. C.; Deller, T.; Korte, M. Not Just Amyloid: Physiological Functions of the Amyloid Precursor Protein Family. *Nat. Rev. Neurosci.* **2017**, *18*, 281.

(52) Murphy, M. P.; Levine, H. Alzheimer's Disease and the Amyloid- β Peptide. *J. Alzheimer's Dis.* **2010**, *19*, 311.

(53) Kanno, S.; Oda, N.; Abe, M.; Terai, Y.; Ito, M.; Shitara, K.; Tabayashi, K.; Shibuya, M.; Sato, Y. Roles of Two VEGF Receptors, Flt-1 and KDR, in the Signal Transduction of VEGF Effects in Human Vascular Endothelial Cells. *Oncogene* **2000**, *19*, 2138.

(54) Harris, R.; Miners, J. S.; Allen, S.; Love, S. VEGFR1 and VEGFR2 in Alzheimer's Disease. *J. Alzheimer's Dis.* **2017**, *61*, 741.

(55) Mahoney, E. R.; Dumitrescu, L.; Moore, A. M.; Cambrono, F. E.; De Jager, P. L.; Koran, M. E. I.; Petyuk, V. A.; Robinson, R. A. S.; Goyal, S.; Schneider, J. A.; Bennett, D. A.; Jefferson, A. L.; Hohman, T. J. Brain Expression of the Vascular Endothelial Growth Factor Gene Family in Cognitive Aging and Alzheimer's Disease. *Mol. Psychiatry* **2019**, *26*, 888.

(56) Chen, Y.-J.; Hsu, C.-C.; Shiao, Y.-J.; Wang, H.-T.; Lo, Y.-L.; Lin, A. M. Y. Anti-Inflammatory Effect of Afatinib (an EGFR-TKI) on OGD-Induced Neuroinflammation. *Sci. Rep.* **2019**, *9*, 2516.

(57) Wang, L.; Chiang, H.-C.; Wu, W.; Liang, B.; Xie, Z.; Yao, X.; Ma, W.; Du, S.; Zhong, Y. Epidermal growth factor receptor is a preferred target for treating Amyloid- β -induced memory loss. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 16743.

(58) Niu, M.; Hu, J.; Wu, S.; Zhang, X.; Xu, H.; Zhang, Y.; Zhang, J.; Yang, Y. Structural Bioinformatics-Based Identification of EGFR Inhibitor Gefitinib as a Putative Lead Compound for BACE. *Chem. Biol. Drug Des.* **2014**, *83*, 81.

(59) Netzer, W. J.; Bettayeb, K.; Sinha, S. C.; Flajolet, M.; Greengard, P.; Bustos, V. Gleevec shifts APP processing from a β -cleavage to a nonamyloidogenic cleavage. *Proc. Natl. Acad. Sci.* **2017**, *114*, 1389.

(60) Eisele, Y. S.; Baumann, M.; Klebl, B.; Nordhammer, C.; Jucker, M.; Kilger, E. Gleevec Increases Levels of the Amyloid Precursor Protein Intracellular Domain and of the Amyloid- β -degrading Enzyme Nephrilysin. *Mol. Biol. Cell* **2007**, *18*, 3591.

(61) Estrada, L. D.; Chamorro, D.; Yañez, M. J.; Gonzalez, M.; Leal, N.; Von Bernhardi, R.; Dulcey, A. E.; Marugan, J.; Ferrer, M.; Soto, C.; Zanlungo, S.; Inestrosa, N. C.; Alvarez, A. R. Reduction of Blood Amyloid- β Oligomers in Alzheimer's Disease Transgenic Mice by c-Abl Kinase Inhibition. *J. Alzheimer's Dis.* **2016**, *54*, 1193.

(62) Sanchez, A.; Tripathy, D.; Yin, X.; Luo, J.; Martinez, J. M.; Grammas, P. Sunitinib enhances neuronal survival in vitro via NF- κ B-mediated signaling and expression of cyclooxygenase-2 and inducible nitric oxide synthase. *J. Neuroinflammation* **2013**, *10*, 857.

(63) Huang, L.; Lin, J.; Xiang, S.; Zhao, K.; Yu, J.; Zheng, J.; Xu, D.; Mak, S.; Hu, S.; Nirasha, S.; Wang, C.; Chen, X.; Zhang, J.; Xu, S.; Wei, X.; Zhang, Z.; Zhou, D.; Zhou, W.; Cui, W.; Han, Y.; Hu, Z.; Wang, Q. Sunitinib, a Clinically Used Anticancer Drug, Is a Potent AChE Inhibitor and Attenuates Cognitive Impairments in Mice. *ACS Chem. Neurosci.* **2016**, *7*, 1047.

(64) Hassan, M.; Raza, H.; Abbasi, M. A.; Moustafa, A. A.; Seo, S.-Y. The Exploration of Novel Alzheimer's Therapeutic Agents from the Pool of FDA Approved Medicines Using Drug Repositioning, Enzyme Inhibition and Kinetic Mechanism Approaches. *Biomed. Pharmacother.* **2019**, *109*, 2513–2526.

(65) Wang, L.; Liu, J.; Wang, Q.; Jiang, H.; Zeng, L.; Li, Z.; Liu, R. MicroRNA-200a-3p Mediates Neuroprotection in Alzheimer-Related Deficits and Attenuates Amyloid-Beta Overproduction and Tau Hyperphosphorylation via Coregulating BACE1 and PRKACB. *Front. Pharmacol.* **2019**, *10*, 806.

(66) Woo, R.-S.; Lee, J.-H.; Yu, H.-N.; Song, D.-Y.; Baik, T.-K. Expression of ErbB4 in the Neurons of Alzheimer's Disease Brain and APP/PS1 Mice, a Model of Alzheimer's Disease. *Anat. Cell Biol.* **2011**, *44*, 116.

(67) Caltagaroni, J.; Jing, Z.; Bowser, R. Focal Adhesions Regulate A β Signaling and Cell Death in Alzheimer's Disease. *Biochim. Biophys. Acta, Mol. Basis Dis.* **2007**, *1772*, 438.

(68) Grace, E. A.; Busciglio, J. Aberrant Activation of Focal Adhesion Proteins Mediates Fibrillar Amyloid β -Induced Neuronal Dystrophy. *J. Neurosci.* **2003**, *23*, 493.

(69) Kim, E. K.; Choi, E.-J. Pathological Roles of MAPK Signaling Pathways in Human Diseases. *Biochim. Biophys. Acta, Mol. Basis Dis.* **2010**, *1802*, 396.

(70) Zhu, X.; Lee, H.-g.; Raina, A. K.; Perry, G.; Smith, M. A. The Role of Mitogen-Activated Protein Kinase Pathways in Alzheimer's Disease. *Neurosignals* **2002**, *11*, 270–281.

(71) Bagheri, S.; Squitti, R.; Haertlé, T.; Siotto, M.; Saboury, A. A. Role of Copper in the Onset of Alzheimer's Disease Compared to Other Metals. *Front. Aging Neurosci.* **2018**, *9*, 446.

(72) Choi, H.; Ho Koh, S. Interaction between Amyloid Beta Toxicity and the PI3K Pathway in Alzheimer's Disease. *J. Alzheimer's Dis. Park.* **2016**, *6*, 269.

(73) Yu, H.-J.; Koh, S.-H. The Role of PI3K/AKT Pathway and Its Therapeutic Possibility in Alzheimer's Disease. *Hanyang Med. Rev.* **2017**, *37*, 18.

(74) Thomas, R.; Zuchowska, P.; Morris, A. W. J.; Marottoli, F. M.; Sunny, S.; Deaton, R.; Gann, P. H.; Tai, L. M. Epidermal Growth Factor Prevents APOE4 and Amyloid-Beta-Induced Cognitive and Cerebrovascular Deficits in Female Mice. *Acta Neuropathol. Commun.* **2016**, *4*, 111.

(75) da Rocha, J. F.; Bastos, L.; Domingues, S. C.; Bento, A. R.; Konietzko, U.; da Cruz e Silva, O. A. B.; Vieira, S. I. APP Binds to the

EGFR Ligands HB-EGF and EGF, Acting Synergistically with EGF to Promote ERK Signaling and Neuritogenesis. *Mol. Neurobiol.* **2021**, *58*, 668.

(76) Fox, I. J.; Kornblum, H. I. Developmental Profile of ErbB Receptors in Murine Central Nervous System: Implications for Functional Interactions. *J. Neurosci. Res.* **2005**, *79*, 584.

(77) Iwakura, Y.; Nawa, H. ErbB1-4-Dependent EGF/Neuregulin Signals and Their Cross Talk in the Central Nervous System: Pathological Implications in Schizophrenia and Parkinson's Disease. *Front. Cell. Neurosci.* **2013**, *7*, 4.

(78) Yan, Y.; Xu, Z.; Hu, X.; Qian, L.; Li, Z.; Zhou, Y.; Dai, S.; Zeng, S.; Gong, Z. SNCA Is a Functionally Low-Expressed Gene in Lung Adenocarcinoma. *Genes* **2018**, *9*, 16.

Thermal and Electrical Behaviour of the Persistent Current Switch for a Whole-Body Superconducting MRI Magnet

Ajit Nandawadekar¹, V. Soni, N. Suman, Sankar Ram T, R. Kumar, S.K. Saini, R G Sharma, Mukhtiar Singh, and Soumen Kar²

Abstract—A prototype persistent current switch is developed for an actively shielded whole-body 1.5 T MRI magnet having an operating current of ~ 500 A. The switch is developed using a six-strand CuNi-NbTi conductor. The total length of the conductor used in the PCS is ~ 40 m using a bifilar winding technique having a room temperature resistance of $15\ \Omega$ and an inductance of $6.6\ \mu\text{H}$. Two numbers of thermo-foil heaters having a resistance of $90\ \Omega$ each are placed between the layers of the winding pack of the switch. The wet-winding technique is followed for the switch using a cryogenic grade epoxy. The characteristics of the switch are performed using a 4 K test rig for its applicability in the 1.5 T MRI magnet. The normal resistance of the switch is measured to be $12.5\ \Omega$ at 15 K which is 15% less than the estimated value. The total energy loss onto the switch is estimated to be 0.24% of the ramping energy of the magnet at 6 V charging voltage. The thermal switching profile of the switch is studied and correlated with the total energy loss.

Index Terms—Magnetic Resonance Imaging, Persistence Current Switch, Superconducting Magnet.

I. INTRODUCTION

A SUPERCONDUCTING solenoid magnet is the heart of any whole-body clinical MRI scanner which needs a very high spatial homogeneity along with the field stability better than 0.1 ppm/hr [1]–[2]. Such high temporal field stability is difficult to be achieved by any commercially available power supply due to its inherent ripple. Any variation in the magnetic field would lead to a change in the resonance frequency of the resonating hydrogen ions resulting in artifacts in the image. The superconducting magnet can provide the highest degree of temporal field stability in the persistent mode of operation. The superconducting switch or the persistent current switch (PCS) across the superconducting coils makes it possible to operate in

the persistent mode to achieve the desired temporal field stability of the MRI magnet [3]–[7]. The PCS plays the most crucial role even during the ramping up of the magnet by maintaining its resistive state i.e., the *OFF* state. The PCS is set at the superconducting state i.e., the *ON* state, once the magnet reaches its desired field for its operation in the persistent mode. An actively shielded multi-coil 1.5 T superconducting MRI magnet system is recently designed for a whole-body clinical scanner [8]. A prototype PCS is developed using a multi-strand CuNi-NbTi conductor for the 1.5 T MRI magnet to analyze its thermal and electrical behaviour to generate input parameters for designing of the final PCS.

In this paper, the design parameters of the PCS are discussed in detail along with its analytical calculation of total energy dissipation. Thermo-foil heaters are used at various intermediate layers of the winding pack of the epoxy impregnated switch to study the thermal behavior in correlation with the location of heaters inside the winding layer. Also, a comparative analysis of the temperature profile along with its switching behavior is discussed for the thermo-foil resistors located at the various layers of the PCS. The normal resistance has been measured across the PCS at various temperatures.

II. DESIGN OF THE PCS

The PCS was made of the NbTi-Cu-30%Ni conductor having six strands with an overall diameter of 1.6 mm. Fig. 1 shows the critical current curve of the CuNi-NbTi conductor. The conductor had a critical current of more than 2 kA at 1 T. In the final application, the PCS will be placed at ~ 0.6 T of the background field of the magnet. The critical current of the conductor with four strands was be 1.35 kA at 1 T. The overall normal resistances per strand (R_{strand}) of the conductor were $2.49\ \Omega/\text{m}$ and $2.21\ \Omega/\text{m}$ respectively at 300 K and 15 K. The strands of the conductor behaved as parallel resistors as shown in Fig. 2.

The normal resistance (R_{PCS}) of the PCS is estimated using Eq. (1);

$$R_{\text{PCS}} = \frac{R_{\text{strand}}}{n} \quad (1)$$

where, n is the number of strands.

Manuscript received December 1, 2020; revised March 29, 2021; accepted April 16, 2021. Date of publication April 30, 2021; date of current version June 11, 2021. This work was supported by the Ministry of Electronics and Information Technology, Govt. of India under Project 11(15)/2014-ME&HI. (Corresponding author: Ajit Nandawadekar.)

Ajit Nandawadekar and Mukhtiar Singh are with the Delhi Technological University, Shahbad Daulatpur 110042, India (e-mail: ajitnandawadekar@gmail.com).

V. Soni, N. Suman, Sankar Ram T, R. Kumar, S.K. Saini, R G Sharma, and Soumen Kar are with the Inter-University Accelerator Centre, New Delhi 110067, India (e-mail: kar.soumen@gmail.com).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TASC.2021.3076748>.

Digital Object Identifier 10.1109/TASC.2021.3076748

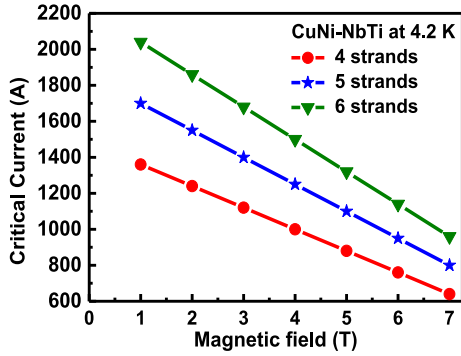


Fig. 1. The critical current curve of the multi-strand CuNi-NbTi conductor [Bruker OST].

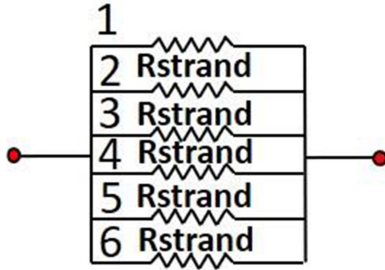


Fig. 2. The equivalent resistance diagram of the six-strands CuNi-NbTi conductor where R_{strand} is the normal resistance of each strand.

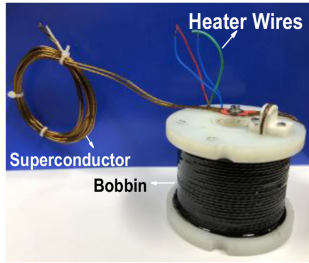


Fig. 3. The prototype persistent current switch developed for the 1.5 T MRI magnet.

Using Eq. (1), the resistance of the PCS conductor with 4, 5, and 6 strands are respectively estimated to be $0.553 \Omega/m$, $0.442 \Omega/m$, and $0.368 \Omega/m$ at 15 K. The critical current of the conductor used in the PCS at its background field needs to be at least 1.5 times higher than the operating current of the magnet to take care of in case of any disturbance. Hence, the critical current and its associated normal resistance can be chosen by selecting appropriate number of strands that to be used for connecting the magnet for its persistent operation. Fig. 3 shows the photograph of the PCS. The bobbin of the PCS was made of insulating material. The total length of the conductor used in the bifilar winding was 40 m. The normal resistance of the PCS with the six-strand conductor was 16.6Ω and 14.6Ω respectively at 300 K and 15 K. For this experimental study, all the six strands of the conductor were used to generate the design parameters for the final PCS. The bifilar winding technique was utilized to achieve the low inductance ($\sim 6.6 \mu H$) necessary to have minimum field perturbation. The parameters of the PCS are summarized in

TABLE I
DESIGN PARAMETERS OF THE PCS

Parameter	Value
No. of layer	8
Total no of turns	264
Turns per layer	33
Total wire length (m)	40
Normal Resistance (Ω) at 300K	16.6
Inductance (μH)	6.6

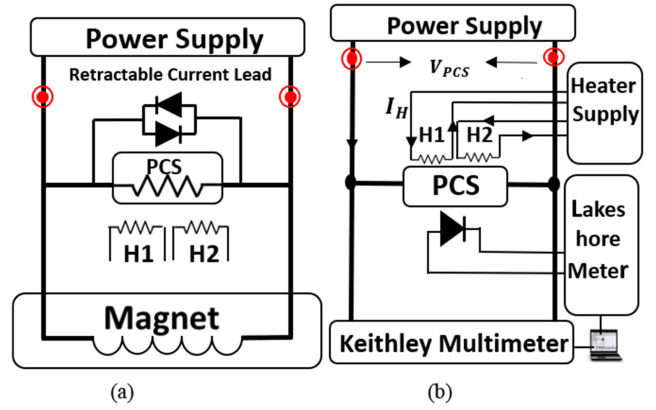


Fig. 4. (a) A simplified electrical schematic of the superconducting magnet with the PCS and, (b) the schematic of the measurement test set up.

Table I. The winding pack of the PCS was impregnated with Sytcast 2850 FT epoxy mixed using catalyst 24.

Two thermo-foil polyimide heaters having electrical resistance of 90Ω each were placed between the layers of the winding of the PCS. Heater-1 (H1) was placed between the 2nd and 3rd layer and the heater-2 (H2) was placed between the 5th and 6th layer. The heaters could be switched between the *ON* and the *OFF* state. Few layers of fibre glass cloth were wrapped on the last layer of the PCS to thermal insulate the winding from direct contact with the liquid helium bath. A calibrated silicon diode (DT-670, Lakeshore Cryotronics Inc.) temperature sensor was fixed onto the intermediate layers of the PCS to monitor its temperature. Fig. 4(a) shows the simplified schematic of the electrical connection of the superconducting magnet with the PCS. During ramp up of the magnet, the PCS must be “open” i.e., *OFF* state which was achieved by making the PCS resistive. It must be “closed” i.e., *ON* state for persistence operation which was achieved by making the switch superconducting. During ramp up of the magnet, the PCS will be at the resistive state until the magnet reaches the desired field. A finite amount of current would flow through the PCS during the ramp up of the magnet. Hence, there will be an energy dissipation into the PCS during the ramp up of the magnet which would eventually result in evaporating the liquid helium. The total energy (Q_T) dissipated into the PCS is calculated by using Eq. (2);

$$Q_T = Q_H + Q_C \quad (2)$$

where, Q_H is the dissipated energy into the PCS heater and, Q_C is the dissipated energy into the PCS at its *OFF* state due to the charging voltage of the magnet, V_C . Eq. (2) can be expressed

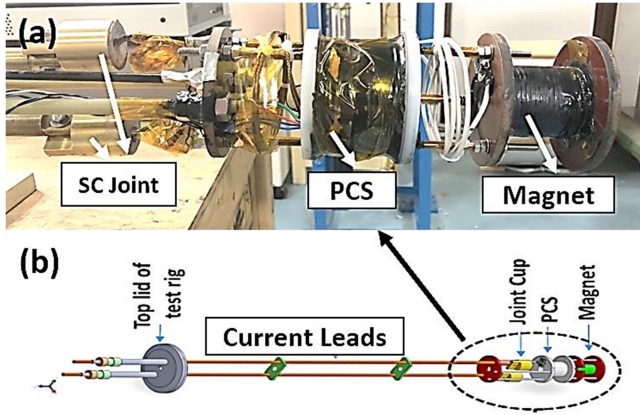


Fig. 5. (a) The 4K rig for testing the PCS and, (b) the schematic representation of the 4K rig used for the thermal and the electrical characterization of the PCS.

as;

$$Q_T = I_H^2 R_H + \frac{V_C^2}{R_{PCS}} \quad (3)$$

where, I_H is the current to the PCS heater, R_H is the value of electrical resistance of the heater, and R_{PCS} is the value of normal resistance of the PCS at its restive state i.e., OFF state. The relation between the normal resistance of PCS and the energy losses are defined in Eqs. (4)–(6) [10]–[11].

$$E_R = \frac{2E_0 L}{R_t} \quad (4)$$

$$P_{loss} = \frac{E_R}{E_0} \times 100\% \quad (5)$$

$$R_{PCS} \geq \frac{200L}{P_{loss} \times t} \quad (6)$$

where, E_R is the total energy losses in the PCS, E_0 is the total stored energy of the magnet, L is the self-inductance of the magnet which is 42 H in our case, and P_{loss} is the percentage of energy loss in the PCS during ramp up of the magnet.

III. EXPERIMENTAL TEST SETUP

Fig. 4(b) shows the schematic of the experimental test setup. The thermo-foil polyimide heaters (H1 and H2) were connected to the heater power supply (Keithley-2450) as shown in Fig. 4(b). During ramp up of the MRI magnet as shown in Fig. 4(a), at the charging voltage (V_C) of 6 V, the current through PCS, (I_{PCS}) was 0.41 A if the normal resistance of the PCS is 14.6 Ω at 15 K. Hereinafter, the current through the PCS during ramp up will be referred to as ‘PCS current’. Hence, during the testing, the PCS was energized with 100–500 mA current (I_{PCS}) at its OFF state i.e., resistive state using a power supply (Kepco, 72 V/6 A). The voltage and the temperature of the PCS were respectively measured using a digital voltmeter (Keithley 2000) and a temperature monitor (model 218, Lakeshore Cryotronics Inc.). The temperature and the corresponding voltage profile were recorded using a laptop for later analysis of the behaviour.

Fig. 5(a)–(b) respectively show the photograph and the schematic of the 4 K test rig developed for characterizing the

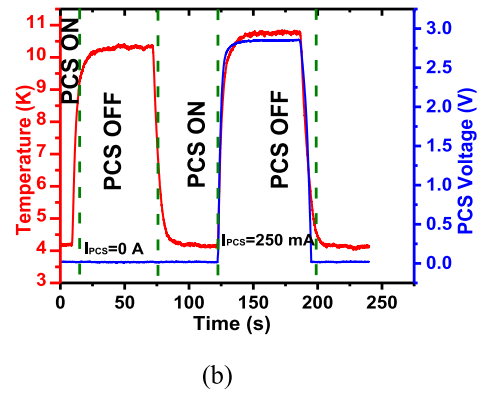
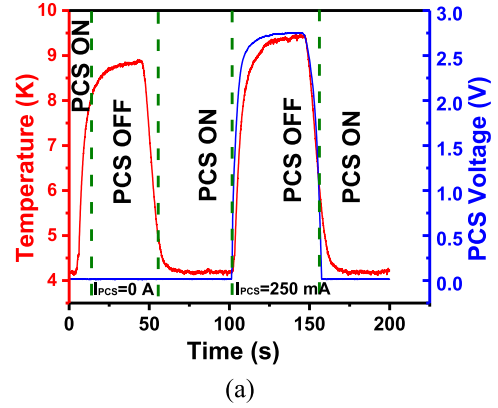


Fig. 6. (a) The temperature and the voltage profile of the PCS (a) with 250 mA of current through the H1 heater and 250 mA of current through the PCS at the resistive state and, (b) with 250 mA of current through the heater H2 and 250 mA of current through the PCS at its resistive state.

thermal and electrical behaviour of the PCS. It consists of a pair of current leads, two joint cups, and the PCS. The terminals of the current leads are connected to the PCS through the joint cups. The 4 K rig as shown in Fig. 5(a) was inserted into a helium dewar for its testing at 4.2 K. During testing of the PCS, the temperature and voltage of the PCS was initially measured at a certain value of heater power without energizing the PCS i.e., without sending any current through the PCS ($I_{PCS} = 0$) at its resistive state. This is referred to as the 1st cycle of each set of measurement. Similarly, in the 2nd cycle, the temperature and the corresponding voltage of the PCS are measured at a certain value of heater power while energizing the PCS by sending a current ($I_{PCS} > 0$) equivalent to the PCS current during ramp up of the magnet. The normal resistance (R_{PCS}) of the PCS could be measured at the 2nd cycle of each measurement by measuring the voltage drop (V_{PCS}) across it using Eq. (7).

$$R_{PCS} = \frac{V_{PCS}}{I_{PCS}} \quad (7)$$

IV. RESULTS AND DISCUSSION

Fig. 6(a) shows the temperature profile of the PCS and corresponding voltage profile with 250 mA of current through the H1 heater. At the 1st cycle ($I_{PCS} = 0$), the temperature of PCS reached to 9 K as shown in Fig. 6. The temperature reached to

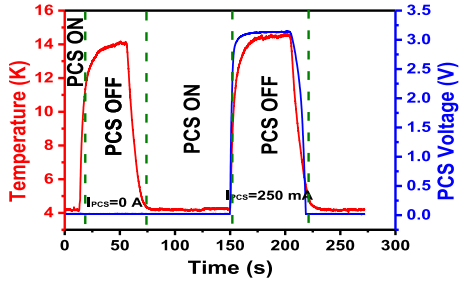


Fig. 7 The temperature and voltage profile of the PCS with 350 mA of current through heater H2 and 250 mA through the PCS.

9.5 K when the PCS was energized with 250 mA of current at its normal state as shown in the 2nd cycle in Fig. 6(a). The current of 250 mA mimics the PCS current during the ramping of the magnet. The temperature of the PCS was increased by 0.5 K due to the current flow through its resistive state. The voltage drop across the PCS was measured to be 2.75 V which signified a normal resistance of 11 Ω of the PCS. As soon as the H1 thermo-foil heater was energized, the transition time from its ON state to OFF state happened within 2 s. Similarly, Fig. 6(b) shows the temperature profile of the PCS and the corresponding voltage profile of the PCS with 250 mA of current through the H2 heater and the PCS.

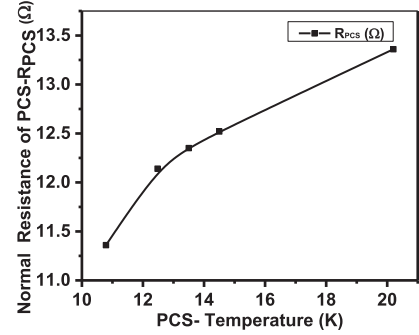
The temperature of the PCS stabilized at 10.5 K. The corresponding voltage across the PCS was 2.9 V, which corresponds to 11.6 Ω of normal resistance of the PCS. With the H2 heater, the equilibrium temperature of the PCS was 1 K higher than its equilibrium temperature with the heater H1 at the same heater power.

Fig. 7 shows the temperature profile of the PCS and corresponding voltage profile with 350 mA of current through the H2 heater. The equilibrium temperature of the PCS reached to 15 K shown in Fig. 7. The corresponding voltage across the PCS is 3.25 V which corresponded to a normal resistance of 13 Ω . The normal resistance per meter measured was 0.325 Ω /m that is 12% less than that of the actual value of the normal resistance at 15 K. This signifies a non-uniform temperature distribution inside the winding of the PCS at its resistive state.

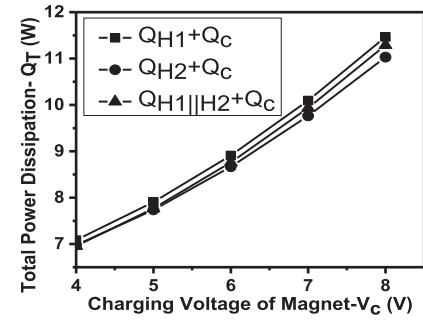
This signifies a non-uniform temperature distribution inside the winding of the PCS at its resistive state. Hence, the PCS needs much better thermal isolation to have uniform temperature distribution. Although the higher degree of thermal isolation would increase the transition time from the normal conducting state to the superconducting state.

Similarly, the temperature of the PCS which was measured with the various currents through the H2 heater.

Fig. 8(a) shows the normal resistance of the PCS measured at various equilibrium temperatures attained at the various heater (H2) power. During ramp up of the magnet, there will be a dissipation of heat energy into the PCS at its resistive state as defined in Eq. (3). The total energy dissipation (Q_T) into the PCS at the various charging voltages in the range of 4–8 V is shown in Fig. 8(b). At 6 V, the total energy dissipation is estimated to be 8.5 W. At 6 V, the total ramp-up time will be 2940 s for the magnet having 42 H of inductance to reach at its desired current of 420 A.



(a)



(b)

Fig. 8. (a) The normal resistance of the PCS with its corresponding temperature and, (b) the estimated energy dissipation into the PCS during ramp-up of the MRI magnet at the various charging voltages.

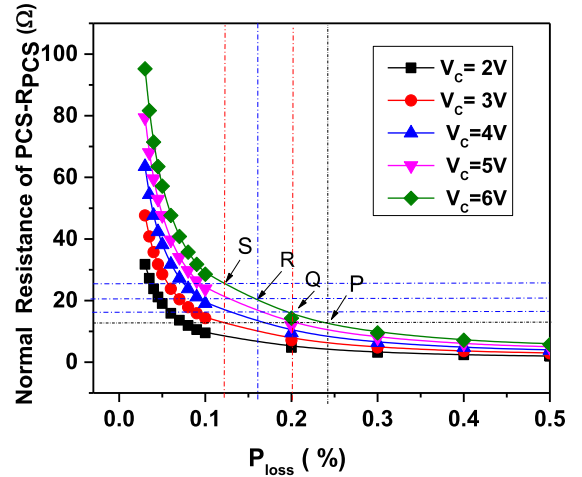


Fig. 9. The estimated energy dissipation on PCS at various temperatures during ramping of the MRI magnet.

Hence, the amount of liquid helium that would be evaporated during the ramping of the MRI magnet would be 9.71. The dissipation of heat energy will be higher for the H1 heater for the same charging voltage because the H1 heater needs higher amplitude of heater power to attain same temperature on the PCS.

Fig. 9 shows the relation of the power loss (P_{loss}) as defined in Eq. (5), with the normal resistance of the PCS during ramp-up of the magnet at the various charging voltages. The point 'P' on Fig. 9 indicates the power loss (P_{loss}) of 0.23% in the PCS

($\sim 13 \Omega$ normal resistance at 15 K) while charging the magnet at 6 V. The percentage of energy losses into the PCS at 17 Ω , 20 Ω , 25 Ω of normal resistance will be 0.18%, 0.16%, and 0.12% as shown in Fig. 9. Hence to reduce the energy loss on the PCS while charging the magnet at 6 V, the normal resistance needs to be increased. The normal resistance can be increased by improving the thermal isolation of the PCS which can be implemented in the final PCS.

The normal resistance of the same PCS can also be increased by using only four strands while connecting with the MRI magnet. The estimated normal resistance of the same PCS would be $\sim 22 \Omega$ with four strands of the conductor. Considering the 12% reduction in actual normal resistance of the PCS, the effective normal resistance will be $\sim 19.4 \Omega$ would eventually result in 0.16% of energy loss during charging the magnet with 6 V of charging voltage. The critical current of the four strand conductor at the background field of 1 T will still be sufficient for the persistent operation of the MRI magnet.

V. CONCLUSION

A prototype superconducting switch for a whole-body MRI magnet has been designed, fabricated, and tested in liquid helium. A comparative analysis of the temperature profile and switching behavior was analyzed for the various thermo-foil heaters located at the various layers of the PCS. Heater H2 gives optimum performance by generating a better temperature profile dissipating a moderate amount of heat during the ramping process. Based on the measured normal resistance and the heat dissipation, the input parameters were generated for the final PCS.

ACKNOWLEDGMENT

The authors would like to thank the Cryogenic Group of Inter-University Accelerator Centre, New Delhi for their support in providing liquid helium for this experimental study. The authors would also like to convey thanks to SAMEER-Mumbai for their support in carrying out this activity for the IMRI project.

REFERENCES

- [1] Y. Lvovsky, E. W. Stautner, and T. Zhang, "Novel technologies and configurations of superconducting magnets for MRI," *Supercond. Sci. Technol.*, vol. 26, 2013, Art. no. 093001.
- [2] H. Maeda, M. Urata, Y. Oda, M. Kageyama, and S. Kabashima, "Instability of persistent current switch," *IEEE Trans. Magn.*, vol. 27, no. 2, pp. 2124–2127, Mar. 1991.
- [3] T. K. Ko, Y. S. Oh, and S. J. Lee, "Optimal design of the superconducting persistent current switch with respect to the heater currents and the operating currents," *IEEE Trans. Appl. Supercond.*, vol. 5, no. 2, pp. 262–265, Jun. 1995.
- [4] B. Dorri and E. T. Laskaris, "Persistent superconducting switch for cryogen-free MR magnets," *IEEE Trans. Appl. Supercond.*, vol. 5, no. 2, pp. 177–180, Jun. 1995.
- [5] C. Cui, Y. Lei, L. Li, Z. Ni, F. Gao, and Q. Wang, "Performance test of superconducting switch for NMR magnet," *IEEE Trans. Appl. Supercond.*, vol. 22, no. 3 Jun. 2012, Art. no. 9502004.
- [6] S. Liu, X. Jiang, G. Chai, and J. Chen, "Superconducting joint and persistent current switch for a 7-T animal MRI magnet," *IEEE Trans. Appl. Supercond.*, vol. 23, no. 3 Jun. 2013, Art. no. 4400504.
- [7] C. Cui, J. Cheng, S. Chen, L. Li, and X. Hu, "Design and test of superconducting persistent current switch for experimental Nb₃Sn superconducting magnet," in *IEEE Trans. Appl. Supercond.*, vol. 26, no. 4 Jun. 2016, Art. no. 0605704.
- [8] S. Kar *et al.*, "Development of high homogeneity and high stability 1.5T superconducting magnet for whole-body MRI scanner," *Indian J. Cryogenics*, vol. 44, no. 1, pp. 193, 2019. [Online]. Available: 10.5958/2349-2120.2019.00034.7
- [9] C. Li *et al.*, "Persistent current switch for HTS superconducting magnets: Design, control strategy, and test results," *IEEE Trans. Appl. Supercond.*, vol. 29, no. 2 Mar. 2019, Art. no. 4900704.
- [10] C. Li *et al.*, "Design for a persistent current switch controlled by alternating current magnetic field," *IEEE Trans. Appl. Supercond.*, vol. 28, no. 4 Jun. 2018, Art. no. 4603205.
- [11] M. N. Wilson, *Superconducting Magnets*, Oxford: Clarendon, 1983, pp. 272–274.

Time Efficient IOS Application For CardioVascular Disease Prediction Using Machine Learning

Vansh Kedia
Department of Computer Engineering
Delhi Technological University
Delhi, India
Vanshkedia1999@gmail.com

Swesh Raj Regmi
Department of Information Technology
Delhi Technological University
Delhi, India
sweshregmi@gmail.com

Khushi Jha
Department of Computer Engineering
Delhi Technological University
Delhi, India
khushi.ktm2@gmail.com

Aman Bhatia
Department of Software Engineering
Delhi Technological University
Delhi, India
aman.b25@hotmail.com

Siddhant Dugar
Department of Computer Engineering
Delhi Technological University
Delhi, India
siddhant.dugar241@gmail.com

Bickey Kumar Shah
Department of Computer Engineering
Delhi Technological University
Delhi, India
Bickeyshah721@gmail.com

Abstract—This paper intends to utilize the vast amount of data generated in the healthcare industry by building machine learning models to predict the incidents of cardiovascular disease in people and hence allow them to take suitable preventive actions. The proposed research work has integrated these functionalities to build a mobile-based ios application using which a person enters details and views system prediction making it an efficient and easy to use interface for the people with time and accuracy. Making the system time efficient in the IOS is of greater importance in the paper. Cardiovascular disease is a class of heart-related disease involving blockage in blood vessels causing health problems like heart attack, chest pain, stroke, and possible heart failure. They are one of the biggest causes of morbidity and mortality in the world and their incident is based on lifestyle hence making identification and prevention difficult. Future scope involves expanding the model to include an integrated prediction including another disease like diabetes and suggest feedback and health tips to the users for healthier lifestyle habits and preventive actions.

Keywords—Machine learning, classification, data model, Cardiovascular Disease (CVD), feature selection, Training model.

I. INTRODUCTION

The disease case that leads to the blood vessels blockage and heart attack with chest pain and results in other heart disease and failure of heart that might lead to death or some other serious issue is the cardiovascular disease. Its instance is caused due to buildup of fatty plaques in arteries leading to constriction of path and stiffening of walls inhibiting blood flow and causing pressure on the heart and arteries. It is one of the reasons and causes of mortality and morbidity among the population of the different countries in the world taking an estimated 17.9 million lives each year. Thus, it has become imperative to deal with it and find ways to prevent and cure it. This can be made possible by utilizing the vast amount of health

data to build a model and make informed decisions and predictions about the possibility of cardiovascular disease in a given person. Therefore, Machine Learning can help utilize healthcare data and give a breakthrough in early identification and prevention. To make an efficient system for predicting the incidence of cardiovascular disease, it intends to input user characteristics and determine the incidence of the following disease in the person.

Cardiovascular disease is a class of lifestyle-based diseases that are mainly dependent on factors reflecting a person's lifestyle like weight, cholesterol, blood pressure, diet, mental health, diabetes occurrence, smoking/alcohol intake, others being age, family history, ethnic background. The proposed system aims to map these factors to the incidence of cardiovascular disease. CVD are those diseases that are related to arterial, blood vessels, valves, and blood circulation in the heart. These are the reasons for the death due to CVD.

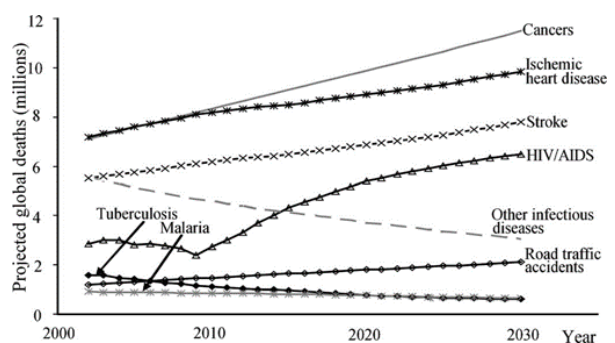


Fig.1 Predicted death from 2002-2030

A person can enter his/her details and our system automatically processes them through the model and generate a result for the person, helping him/her to take appropriate preventive action. A suitable Machine Learning model has to be

designed to determine the incidence of cardiovascular disease in people. The model has to be trained and tested on a dataset and checked for accuracy on testing data, intended to have maximum possible accuracy. An integrated application also needs to be developed containing an interface for user and model's processing allowing people to enter details and confirm the presence of disease. The datasets contain 70,000 records of patient data each containing 11+ features and characteristics which have to be analyzed, utilized, and processed to design the model. The following problem can be solved by building a Machine Learning model using supervised learning techniques to perform Binary Classification on user data, by applying different algorithms (like logistic regression, naive Bayes, support vector machines, decision trees, etc). The assigned data is divided into 2 sets, for training and testing respectively. It is preprocessed, analyzed, and visualized for training the model. The model is then designed using different algorithms or their combination, trained and tested to determine accuracy. Thus, we develop our model which with a good degree of accuracy can predict the incidence of cardiovascular disease in a person. This model can then be integrated into a mobile-based application [2-3]. The application helps create an integrated system that allows the user to enter his details and run the model on his/her characteristics generating prediction result for the user. This system contains the user interface, model, storage together integrated to create a robust system to generate intended results. Cardiovascular disease is a serious burning problem in an underdeveloped country. Different and new technologies and waves of medical Sciences are working on this matter. Many things have been discovered so far. With time more advancement is expected with growing technology in Machine learning and artificial intelligence.

II. LITERATURE REVIEW

A lot of the research-oriented works have been carried out on cardiovascular disease [10]. Research for the model prediction of early detection of cardiovascular disease was seen in 2008 by Tuan D. Pham using Machine learning and Mass spectroscopy data. The research was the combinational application of spectroscopy and Machine algorithm. Association rules discovery for train and test was discovered by Carlos Ordonez in 2006[13]. Through this research, the data set was compared with the state of supervised and unsupervised learning methods. The technology and methods that were used in those days started to be

seen as fruitful. In 2012, the research work was carried out on Cardiovascular risk prediction using a genetic algorithm. An excellent level and with great accuracy of a risk factor is discovered through this method [8]. This research work predicted the risk and cause of CVD from generation to generation. CVDs health informatics are evolving in different fields from data storage to data transmission. In 2020 the clinical implication of Machine learning is predicted for the cardiovascular disease that reads the data into the supervised form and then gives greater accuracy to the data for which the disease is to be predicted [12]. The cases of research have been fruitful since 2002. Every time and every year medical science has achieved well.[5]

III. METHODOLOGY

The paper deal with the IOS application development through machine learning algorithms. Different technology from web app development to machine learning algorithms has been used to create a user-friendly interface. Machine learning work and Interface development work has been classified into different parts of study and implementations to make the learning easier and comfortable. Data loading, data preprocessing to data transfer with swift language implementation is divided into sub learning.

A. Machine Learning model development

The Machine Learning model is the most important component of the system, it accurately determines prediction results based on the training set, applied algorithms.

Its development is composed of the following phases:

1) *Data preprocessing and visualization*

2) *Model Testing and accuracy*

Different factors and features are responsible for the cause of cardiovascular disease. For detecting the cause, we cannot take every factor in the process of the machine learning algorithm [1]. If we mention different factors, then we will have to go through principal component analysis that will make the minimum features and makes the algorithm implementation easy. The important features adopted for the accuracy calculations are age, height, weight, gender, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake, and physical activity. All of these have their variable type while prediction. Some of them are objective features, some are subjective features, some of them are examination and cardiovascular disease detection is the target variable.

Feature	Variable Type	Variable	Value Type
Age	Objective Feature	age	int (days)
Height	Objective Feature	height	int (cm)
Weight	Objective Feature	weight	float (kg)
Gender	Objective Feature	gender	categorical code
Systolic blood pressure	Examination Feature	ap_hi	int
Diastolic blood pressure	Examination Feature	ap_lo	int
Cholesterol	Examination Feature	cholesterol	1: normal, 2: above normal, 3: well above normal
Glucose	Examination Feature	gluc	1: normal, 2: above normal, 3: well above normal
Smoking	Subjective Feature	smoke	binary
Alcohol intake	Subjective Feature	alco	binary
Physical activity	Subjective Feature	active	binary
Presence or absence of cardiovascular disease	Target Variable	cardio	binary

Fig.2 Dataset Description

1) Data Preprocessing and Visualization

This is the initial phase which involves working on the dataset which has been imported. Since the model is trained and tested on the dataset and modifications/tweaks made based on data qualities, distribution, characteristics, it becomes important to have a dataset following all consistencies.

2) Model Development, training, and testing

This step involves building a model (running on various machine learning algorithms) and testing that model on testing data so that its effectiveness to determine result and accuracy to predict can be improved. Also, data needs to be analyzed and features' nature identified so that based on the type of features, availability, and nature, the best possible algorithm can be applied. For this, it is very important to have an understanding of the dataset and its features. A similar understanding for various classification algorithms is needed so that an appropriate match can be made, and a model can be constructed using the algorithm. Then it involves testing our model to see the obtained accuracy. Based on the accuracy, changes or modifications are made like changes in algorithm, parameter tuning, etc.

B. IOS Application Development

The application provides an interface for the user to know prediction for his/her case in real-time based on entered details, thus providing flexibility and an easy to use the system. Its development is composed of the following phases:

1) Frontend Development

This phase involves developing a frontend user interface, which interacts with the user to take in suitable details related to the user (user characteristics) and display output on the screen.

2) Backend Development

This phase involves developing the backend to handle requests coming from the frontend side, handling and storing data, and integrating the system with the Machine Learning Model[4]. This can then be integrated with the frontend to produce a complete working application.

C. Tools and Technologies Used

Jupyter notebook was used for the machine learning model prediction and for the ios development work swift was used for the front end and core machine learning for the backend work.

Activity	Tools and framework
Machine learning	Python
IOS App Development	Frontend->swift UI Backend->COREML

Table.1 Tool and Technologies

The Jupyter Notebook is an open-source web application that allows us to create and share documents that contain live code, equations, visualizations, and narrative text. Uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more [9].

SWIFT UI is a framework in Swift which allows us to write and code apps in a declarative approach. In this framework, we specify Swift UI about the working and the appearance of our app with the help of code.

CoreML is a framework using which Machine Learning models can be integrated into an IOS app. It supports natural language processing (NLP), image analysis, and various other conventional models to provide top-notch on-device performance with minimal memory footprint and power consumption [5-6].

D. Project Implementation

1) Data Validation

This step involves viewing and checking the data entries, values and validating them for further use after suitable corrections. Inbuilt libraries in Python like NumPy, pandas are used to view and validate. After loading the data into the notebook the null values and missing values are detected. After the detection of the missing values, either the column of the missing data is dropped or it is filled with the suitable value. It is done as per the requirement of the dataset. The datasets are then preprocessed and solved. Each column of the data set is transformed into the numerical value that machine language understands. Every column of data is transformed into the binary form as the machine understands only the binary language i.e 0 and 1[7-11]. The duplicate and irrelevant data are removed to make and implement the algorithms in an easy manner and the final data is validated.

2) Data Visualization and Outlier Detection

Using libraries like matplotlib and seaborn, data is visualized using different graphs, charts, scale diagrams, models. Using it, we can determine outlier entries that are redundant for the model as they are having significantly different values from normal entries, hence statistically they make a large difference in the model impacting the prediction negatively [9].

Following are possible causes of Outliers:

- Data entry errors (human errors while entering data)
- Measurement errors (instrument errors leading to wrong entries)
- Experimental errors (data extraction or experiment planning/executing errors)
- Data processing errors (data manipulation or data set unintended mutations)

To eliminate the problem while modeling we eliminate the highest and lowest data of the particular features as they are quite big and do not fit the averages values of the datasets. Many outliers of height and weight are combined to form a new feature.

3) Feature Extraction and selection

This step involves modifying and adjusting column values to designing appropriate features which can be utilized in building model and finding the best possible prediction. Here, we select features from available column entries. This involves tasks like:

- Scaling (to appropriate limits)
- Interdependence removal (features with collinear relation are separated)

- Deriving new features (from existing to reduce feature count in models)

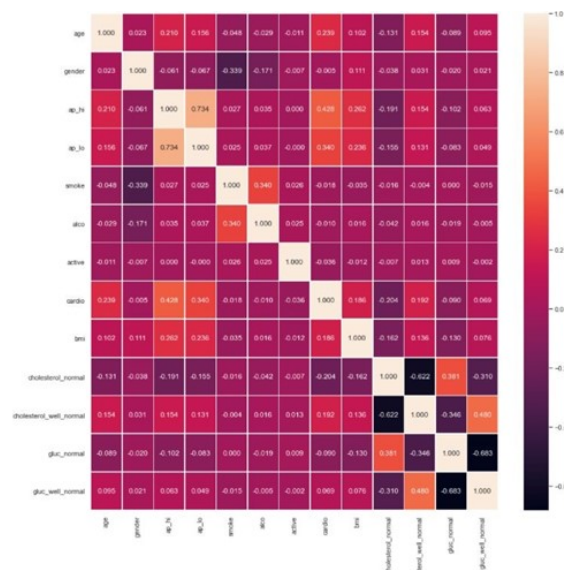


Fig.3 Correlation matrix

In the figure.3 mentioned it is a correlation matrix that gives the index of correlation between different features. The correlation matrix helps to identify that how one feature is related to another and if the given features are related by more than 90% then any one of the data columns is reduced and this process is carried out through principal component analysis [8].

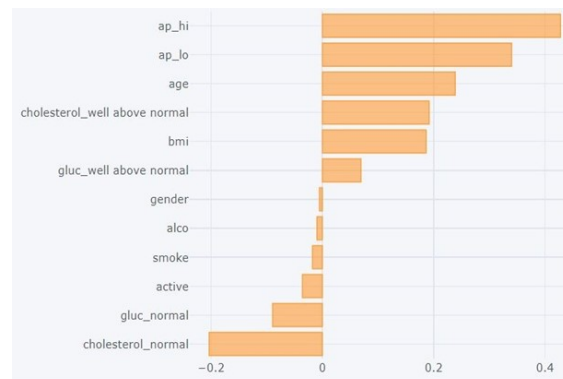


Fig.4 Feature correlation and heatmap

4) Model Training and Testing

This step involved finding a suitable model for the dataset based on feature engineering and model needs and requirements. The training and testing are performed to find the best possible model. Also, improvisations are made in testing to improve model accuracy and prediction.

5) Model Selection and Data Splitting

The data is divided into training and testing data separately. Once models are selected, they are trained on the training data to tune weights and determine biases. classification models that were applied:

- Logistic Regression
- Decision tree
- Random forest
- XGBoost

Before the part of data splitting the task of the standard scalar is performed that converts all the data into one single unit dimension and after that computer easily reads that data form. This is one of the most important steps throughout the work[10].

6) Training and Testing

Once the model is built, it is trained on the available data. After training, the model is weighed accordingly to the data and is made ready for testing. Parameters like accuracy, f1score are determined and on basis of them, a suitable model is selected.

Following steps are followed,

- Testing the different models developed on testing data, to observe accuracy and effective result.
- To remove deficiencies and insufficiency in different models based on results in the testing phase and tune the models better for prediction.
- Select the most appropriate model from different built models based on best accuracy and the overall result after rigorous testing and identification.

7) Integration Processes

The frontend of this app is written in the Swift programming language and the main framework to develop this app is swiftUI. Swift is a programming language that was introduced in 2014 and developed by Apple Inc. It is a very powerful, general-purpose, and open-source programming language. It is also user friendly and is easy to pick up by new programmers. The frontend of the iOS application has been developed to interact with the user. The user enters his/her details and they are taken and feed into the model to predict the output i.e. prediction for the incidence of cardiovascular disease.

Backend development involved building and training a Machine Learning model and then integrating it with the application frontend to

create an integrated system for the user. The subtasks are:

- Design system backend to handle user requests and efficiently answer them.
- Handle data storage and access for the application and develop robust techniques to ensure security and integrity.
- Integrate the Machine Learning model with the system backend and allow effective processing of the model to generate appropriate results communicated to the user through the frontend interface.

The backend of the app uses the CoreML framework, By using CoreML the system,

- Accepts data from the front end side and run prediction.
- Integrates the machine learning models into the app.
- Run the model and sends the result.
- A user inputs the values in the app which works as a test feature on our trained model and outputs the data.

IV. RESULT ANALYSIS

The application developed gives the prediction that for features mentioned it gives correct accuracy of cardiovascular disease or not. The result obtained was efficient up to 72.70% through the XGboost algorithm of Machine learning. Logistic regression was efficient at about 72.39%, decision tree to 62.87%, and random forest to 69.18%.

	Accuracy in %	F1-Score
XGBoost	72.700000	0.720000
Logistic Regression	72.390000	0.710000
Random Forest	69.180000	0.690000
Decision Tree	62.870000	0.630000

Table.2 Accuracy Result.

IOS supporting this application is created taking time into the consideration. High Mathematical and computations methods and data have been used in the algorithms that predict the result into the milliseconds. The application is used in the IOS has a greater effect. If the application is installed and used in other android phones it comparatively provides less efficiency.

V. CONCLUSION AND FUTURE WORK

The application can be made more efficient and advanced using more inclined features that are inclined to the prediction of cardiovascular disease. Like bone marrow problems, Valve problems, and many more. Besides it, we can use Deep learning and CNN methods using the hyperparameter to make and increase the efficiency by more than 90%. Medical sciences and scientists are working on these methods and research for easier prediction of CVD through laser. CardioVascular disease has been a serious burning problem in underdeveloped and developed countries. The medical sciences are trying to reach every part of the world to get rid of this and research works, and scientists have made their work easier and comfortable through this mobile application prediction.

ACKNOWLEDGMENT

The authors are indebted to the central library of Delhi Technological University for the research lab and materials required to complete the research work.

REFERENCES

- [1] P. Saranya and P. Asha, "Survey on Big Data Analytics in Health Care," *Proc. 2nd Int. Conf. Smart Syst. Inven. Technol. ICSSIT 2019*, vol. 01, no. 02, pp. 46–51, 2019, doi: 10.1109/ICSSIT46314.2019.8987882.
- [2] V. T., "Neural Network Analysis for Tumor Investigation and Cancer Prediction," *J. Electron. Informatics*, vol. 2019, no. 02, pp. 89–98, 2019, doi: 10.36548/jes.2019.2.004.
- [3] B. N. Steele, M. T. Draney, J. P. Ku, and C. A. Taylor, "Internet-based system for simulation-based medical planning for cardiovascular disease," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 2, pp. 123–129, 2003, doi: 10.1109/TITB.2003.811880.
- [4] C. Ordonez, "Association rule discovery with the train and test approach for heart disease prediction," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 2, pp. 334–343, 2006, doi: 10.1109/TITB.2006.864475.
- [5] T. H. Shaffer, M. D. Altose, D. H. Lederer, and N. S. Chemiack, "The Interaction of FRC and Ventilation on Occlusion Pressure in Conscious Man," *IEEE Trans. Biomed. Eng.*, vol. BME-24, no. 5, pp. 444–448, 1977, doi: 10.1109/TBME.1977.326180.
- [6] T. D. Pham *et al.*, "Computational prediction models for early detection of risk of cardiovascular events using mass spectrometry data," *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 5, pp. 636–643, 2008, doi: 10.1109/TITB.2007.908756.
- [7] L. N. Pu, Z. Zhao, and Y. T. Zhang, "Investigation on cardiovascular risk prediction using genetic information," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 5, pp. 795–808, 2012, doi: 10.1109/TITB.2012.2205009.
- [8] C. J. Hartley, M. Naghavi, O. Parodi, C. S. Pattichis, C. C. Y. Poon, and Y. T. Zhang, "Guest editorial cardiovascular health informatics: Risk screening and intervention," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 5, pp. 791–794, 2012, doi: 10.1109/TITB.2012.2216057.
- [9] Y. T. Zhang, Y. L. Zheng, W. H. Lin, H. Y. Zhang, and X. L. Zhou, "Challenges and opportunities in cardiovascular health informatics," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 3, pp. 633–642, 2013, doi: 10.1109/TBME.2013.2244892.
- [10] B. J. Mortazavi *et al.*, "Prediction of Adverse Events in Patients Undergoing Major Cardiovascular Procedures," *IEEE J. Biomed. Heal. Informatics*, vol. 21, no. 6, pp. 1719–1729, 2017, doi: 10.1109/JBHI.2017.2675340.
- [11] M. Yasin, T. Tekeste, H. Saleh, B. Mohammad, O. Sinanoglu, and M. Ismail, "Ultra-Low Power, Secure IoT Platform for Predicting Cardiovascular Diseases," *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 64, no. 9, pp. 2624–2637, 2017, doi: 10.1109/TCSI.2017.2694968.
- [12] P. Bizopoulos and D. Koutsouris, "Deep learning in cardiology," *arXiv*, vol. 12, pp. 168–193, 2019.
- [13] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
- [14] S. Adhikari, S. Thapa, and B. K. Shah, "Oversampling based Classifiers for Categorization of Radar Returns from the Ionosphere," *Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESc 2020*, no. Icesc, pp. 975–978, 2020, doi: 10.1109/ICESc48915.2020.9155833.
- [15] G. Joo, Y. Song, H. Im, and J. Park, "Clinical implication of machine learning in predicting the occurrence of cardiovascular disease using big data (Nationwide Cohort Data in Korea)," *IEEE Access*, vol. 8, pp. 157643–157653, 2020, doi: 10.1109/ACCESS.2020.3015757.
- [16] S. Thapa, P. Singh, D. K. Jain, N. Bharill, A. Gupta, and M. Prasad, "Data-Driven Approach based on Feature Selection Technique for Early Diagnosis of Alzheimer's Disease," *Proc. Int. Jt. Conf. Neural Networks*, 2020, doi: 10.1109/IJCNN48605.2020.9207359.
- [17] A. Ghimire, S. Thapa, A. K. Jha, A. Kumar, A. Kumar, and S. Adhikari, "Pandemic," pp. 1083–1092, 2021.
- [18] Z. Huang *et al.*, "Parkinson's Disease Classification and Clinical Score Regression via United Embedding and Sparse Learning From Longitudinal Data," pp. 1–15, 2021.
- [19] A. Kumar and A. Kumar, "Dog Breed Classifier for Facial Recognition using Convolutional Neural Networks," pp. 508–513, 2020.
- [20] B. K. Shah, V. Kedia, R. Raut, S. Ansari, and A. Shroff, "Evaluation and Comparative Study of Edge Detection Techniques," vol. 22, no. 5, pp. 6–15, 2020, doi: 10.9790/0661-2205030615.

Toxic Speech Detection using Traditional Machine Learning Models and BERT and fastText Embedding with Deep Neural Networks

Pranav Malik*

Department of Information Technology
Delhi Technological University
Delhi, India
prnvmlik@gmail.com

Aditi Aggrawal*

Department of Information Technology
Delhi Technological University
Delhi, India
aditiagg99@gmail.com

Dinesh K. Vishwakarma*

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

* All authors have contributed equally

Abstract— The introduction of social media brought about a revolution in the world of digitalization and communication. These platforms were initially developed with a purpose of connecting people across the global boundaries while allowing them to express their views and opinions and learn from others' ideas. With the incoming of the pandemic, the usage of these sites has risen significantly by it by the businesses, educational institutions, students or general public. The increasing ubiquity of social media platforms like Twitter and Facebook has been an issue of major concern since a long time. Along with providing a way for enhanced communication, these platforms also allow internet users to voice their opinions which get circulated among the masses within seconds. Moreover, given the different backgrounds, beliefs, ethnicity and cultures that the users on these platforms come from, many of them tend to use mean, aggressive and hateful content during their discussions with people not hailing from a background similar to theirs. The amount of hate speech and offensive content has been increasing exponentially. Terms like "profane", "hate", and "offensive" are used interchangeably, and hence these have been classified under a broader category of "Toxic" content. A major part of our dataset focuses on conversations prevailing among the youth. After the preprocessing of this dataset using NLP and embeddings (Bert and fastText), a bunch of Machine Learning (LR, SVM, DT, RF, XGBoost) and Deep Learning algorithms (CNN, MLP, LSTM) have been performed, with CNN giving the best results.

Keywords— hate speech, offensive, toxic, classification, deep learning, natural language processing, twitter, facebook

I. INTRODUCTION

In today's era of online connections, with the growing prevalence of social media sites like Facebook, Twitter, Instagram, Youtube and Snapchat, more than half of the population of the world seeks to connect and converse through these platforms. This ability of being able to connect with a mass audience by generating and sharing content to interact over large distances, has changed the way these users are involved in public affairs, politics and also with each other. All this has led to provoking violence and amplified the propagation of hateful content. Most of these social media platforms are motivated to draw attention as a part of their business model. Since this

offensive content attracts the attention of the masses, it becomes more audible on such platforms.

Many a times, it is the inability of people to understand and acknowledge opinions, ideas and views of people hailing from different gender, socio-economic backgrounds and cultures that acts as the driving cause of the spread of this toxicity and hence the hate content targets a specific gender, religion, ethnic group or racial community. Of the population being targeted, adolescents form a major group that are vulnerable to such hatred. Being extremely involved on these platforms, the youth community is a major contributor and receiver of this content. Therefore a major part of our dataset contains tweets circulating among the youth.

Public research surveys suggest that around 25% of respondents tend to feel unsafe in their community because of the spread of online hate and more than 80% are trying to establish ways to monitor and combat cyber bullying. In the year of 2018, platforms like Facebook and YouTube had to take down 3 million posts and 25,000 of videos to maintain the credibility of their content's quality. The propagation of such content is leading to increased violence in matters such as communal riots and lynching globally. Moreover, children under the age of 25 years are twice as likely to be involved in suicidal acts because of this toxicity across the web. Melania Trump, the former first lady of the United States, has made it her own goal to fight against the spread of online hate speech and cyberbullying. A number of international institutions, including the UN Human Rights Council and the Online Hate Prevention Institute are engaged in understanding the nature, proliferation, and prevention of online hate speech.

As observed through the analysis of our dataset, the prevailing use of various slangs and abbreviations, like "wtf", "asap", "tbh", "idc" etc., emojis to make the content more interactive and expressive, extensive repetitions as a form of emotional emphasis in a statement, like "Hate him sooo muchhh" is used as an indication of extent of despise the sender holds towards the addressee, and the extensive use of punctuations, like "Shut Up You !!!!!", all of this makes it

difficult for the researchers and the government to study the content and check it for the presence of any toxicity.

For this research, following two datasets are combined, each of which is publicly available: ALONE Dataset (Adolescents ON twittEr) and the English dataset from Fire 2020's HASOC shared task. The ALONE dataset consists of 688 toxic interactions among the adolescents on the platform of Twitter. The HASOC dataset includes 5852 entries of toxic content from Twitter and Facebook.

Hate speech has no international definition legally, but is a deliberate act aimed at discrimination, violence and hostility. Offensive speech is the derogatory text that uses abusive slurs or curse words. Profanity is the use of obscene or in-appropriate words meant to demean an individual or community. The three terms are often confused with each other, as clearly indicated by the three examples in Table I. Hence it was decided to combine them to form a major group of toxic content.

TABLE I. EXAMPLES OF POSTS

S.No	Post	Label
1	Trump doesn't know what the Hell he is talking about!!!!He can't help but spew B*llshit!!! He's a Disgrace & 2020 cannot get here fast enough #FuckfaceVonClownstick #fucktrump #LockTrumpUp #ImpeachTheMF #ImpeachTheMFTraitor	Profane
2	Now we can add #FlagRapist to the resume of #Traitor, #Pedophile, #Racist, #Rapist and serial sexual predator of women @realDonaldTrump!	Offensive
3	Of course, #traitor, #peodphile, #racist, #rapist and serial sexual predator of women, @realDonaldTrump will do nothing about this as he has complete contempt for the rule of law, unless he can use it as a tool to bash an opponent!	Hate

This research work contributes through a series of steps. Before drawing a final comparison among the results of applied ML and DL classifiers, the following four steps have been performed sequentially:

1. First step was to pre-process the data going through tokenization, performing basic stemming and lemmatization techniques, introducing part-of-speech and then finally incorporating resulting tweets into TF-IDF vectorizer.
2. Second step was to perform various Machine Learning algorithms including Logistic Regression, Support Vector Machine, Decision Trees, Random Forest, and Gradient Boosting. Along with that Ensemble Learning methods of Stacking and Bagging were also performed.
3. Third step was to perform a couple of word embedding techniques (context based and non-context based) comprising of BERT Embedding and fastText, passed the resulting vectors through layers of various deep neural networks.
4. The final step aims at getting better accuracy and recall values with the use of fine - tuned deep neural networks

such as LSTM, CNN and MLP. Lastly, a comparison of the results was drawn to establish the most accurate algorithm.

Manual filtering can be very time consuming with very low accuracy. Automatic systems are thus required to carry out this process in a much more efficient way while saving a lot of time and effort. Recently, there has been a significant development in the fields of ML and DL which provide us a means to analyse the text semantically and make predictions while understanding the content. Despite this development, it remains difficult to compare the performance of these algorithms. The research analysis has been carried out on a combination of two datasets. The ALONE dataset used in this work is a relatively new dataset and we compare the performance of various ML and DL classifiers that will act as state of the art models for evaluations in future studies. The introduction of this data helps us to focus more on the hate pertaining among the youth of the society. The other dataset used is from HASOC'20 and the focus has been to classify the tweets in either of the two major categories of Toxic Speech and Non-Toxic Speech while not indulging into the sub categories for now.

II. RELATED WORK

The area of hate speech detection has been relatively new in terms of research but despite that, it holds many significant contributions within its domain. There have been studies on social media sites of Twitter[1], Facebook[2], YouTube[3][20] and Reddit[4][5]. Fortune and Nunes[6] analysed the motivation behind researching in the area of Hate Speech Detection and also lay out a detailed view of future scope for the same. In [7], Waseem worked on a dataset of 16K tweets categorised as sexism, racism or neither. He performed logistic regression performed the best with various techniques such as character and word n-grams. Gaydhani et al. [8] also performed logistic regression on a combination of three datasets. She observed that Logistic regression with n-gram range of 1 to 3 and TF-IDF vectorizer resulting in an accuracy of 95.6%. In [1], Davidson created a dataset from twitter of size 24k which classifies into hate, offensive or neither after that performed NLP techniques on tweets such as stemming, part of speech tagging, tf-idf vectorizer, vader sentimental analysis etc followed by running various ML classification algorithms from which Logistic Regression with l2 regularisation performed the best. However, they concluded that it was difficult to distinguish Hate and Offensive content based on lexical methods.

As an initial step in Hate Speech Detection using neural networks, [18] highlights the work of Djuric et al. He proposes a two-step approach. He creates a low dimensional text embedding with similar comments and words are learnt using a CBOV (continuous Bag of Words) neural model. This text embedding is then used to train a binary classifier to distinguish hate content from non-hate content. In his work in [18] Djuric, worked on various embeddings combined with deep learning models on a 16K annotated tweets dataset and found out that random embedding combined with LSTM followed by gradient boosted Decision trees performed the best. Zhang et al. [9] introduced a CNN+GRU (convolutional neural network, gated

recurrent unit network) model to provide an improved accuracy over the existing models of CNNs and GRUs independently used for hate speech detection. Their model improved the F1-score by up to 13%.

Tharindu, Marcos and Hansi worked on an annotated dataset from facebook and twitter in HASOC competition with labels Hate & Offensive (HOF) and not Hate & Offensive (NOT). In [10] they performed minimal pre-processing to retain the multi lingual characteristics of dataset and then performed various deep learning architectures such as Pooled GRU, stacked LSTM with attention, LSTM and GRU with attention, 2D Convolution with Pooling, GRU with Capsule, LSTM with Capsule and Attention and BERT in which fine tuned BERT outperformed all with a 0.78 accuracy. As described in [19], Wijesiriwardene created a dataset ALONE (Adolescents ON twittEr) which focuses on conversations from youth on twitter, it classifies the tweets as toxic, non-toxic and unclear. This dataset also provides various metadata and multimodal aspects. We have tried to perform various machine learning and deep learning algorithms on an entirely new dataset of ALONE to highlight the toxicity prevailing among the youth. We have combined the two datasets of similar nature to establish our results.

III. PROPOSED METHODOLOGY

Fig. 1 shows the basic idea of the methodology proposed in our research work for the purpose of classifying text into either of the three categories: toxic, non-toxic or unclear. As an initial step, we have combined two different datasets to formulate our dataset for this research. Further analysis is divided into two separate steps, one involving the application of ML classifiers and the other of DL based algorithms. As a part of the first step, the data is pre-processed before being fed into machine learning algorithms of LR, SVM, DT and RF. Ensemble learning techniques of gradient boosting, bagging and stacking are also used. For the second step, BERT and fastText embeddings are used followed by the classification performed by various DL algorithms like MLP, LSTM and CNN. This work is concluded by comparing the results of these algorithms with one another.

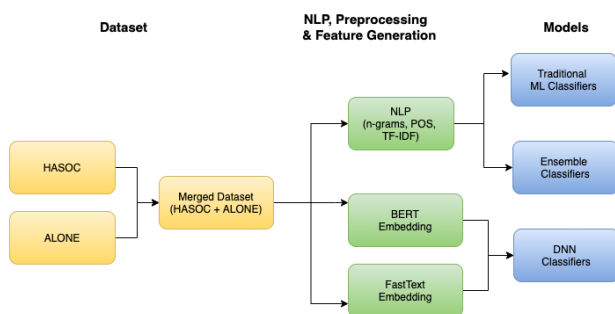


Fig. 1. Proposed Methodology

A. Preprocessing

Natural Language Processing is basically the interaction between humans and machines where machines detect and analyse large amount of data in the form of text, similar to the

way humans do. NLP is used to extract useful information from the text being studied. The NLTK library has been used to work with the WordNet corpus. Further steps of tokenization with an ngram range of (1,3), stopwords removal, stemming, lemmatization, Part-Of-Speech tagging for understanding of grammar have been performed during the data pre-processing. To obtain more accurate results, we have used TF-IDF approach over the traditional BOW approach for creating the count vectorizer.

B. Machine Learning

Logistic Regression (LR): This is one of the most common algorithms to be used when the dependent variable is categorical, especially in case of binary classification. The core method or the middle of logistic regression is the logistic function, which uses sigmoid function as main entity.

Support Vector Machine (SVM): This algorithm is extensively used in high dimensionality space for classification, regression or outlier detection. It classifies data points by developing a hyperplane in an N-dimensional space (represented by N features).

Decision Tree (DT): This is regarded as one of the most comprehensive algorithms in all machine learning algorithms. Decision trees help recognize the most important features of a data set and non-linear patterns can be easily captured by them. The splitting in a decision tree is based on factors like entropy and information gain.

Random Forest (RF): This is a supervised algorithm that can be used for both classification and regression. This algorithm creates a forest of trees with higher number of trees signifying higher accuracy results and reduced overfitting. To calculate the importance of each feature, random forest uses either MDI (mean decrease in impurity) or Gini Importance.

In this work, the count vectorizer generated from TF-IDF approach is fed into each of these algorithms and the results are then compared.

C. Ensemble Learning

Ensemble learning[11][12] methods are relatively recent discoveries in the area of machine learning and deep learning, where the main idea is to combine the predictions of multiple models(weak learners) to give rise to a new model(strong learner). Ensemble models often produce more accurate results than each of the individual ones. They can be used for classification of text.

Gradient Boosting: This algorithm learns the models sequentially, where each model is trained on the same dataset. It focuses on reducing the bias of the model and was developed to improve the efficiency of the algorithm in terms of memory resource requirements and computation time, while utilising the available resources for model training. The XGBoost (eXtreme Gradient Boosting) library has been used for implementing the algorithm.

Stacking: This is an algorithm that learns to combine the results of best performing algorithms that fit on a particular dataset. In this method of ensemble learning, the algorithms whose results are being combined are typically different and their results are learnt in parallel by the meta-model that predicts the final

output. It focuses on enhancing the model's predictive power. In this case, the results are not guaranteed to be an improvement over the results of individual models.

Bagging: This algorithm learns from more than one model based on similar algorithm, runs them in parallel to each other and combines their results using some deterministic averaging and voting process. Each of these models, chooses random sub-samples from the original dataset and as a result some samples of the dataset occur more than once and some of them do not occur at all. It focuses on reducing the variance of the model.

During this research, stacking and bagging has been performed on the results of four machine learning classifiers of LR, SVM, DT and RF.

D. Word Embeddings

The computers are meant to process information in the form of numbers and hence for them to draw semantic and syntactic understanding of the textual content, it needs to be converted into numbers. The Word embeddings [13] encode the words present in the text and act as building blocks for NLP. A word embedding clusters together semantically related (e.g. "car" & "road") or semantically similar words (e.g. "car" & "bus") present in an N-dimensional space. Since word embeddings require a large amount of training time to learn from an unlabelled huge corpus, pre-trained word embeddings have been used in this work. At the same time one of the major advantages of using these embeddings is that they do not require large annotated text for training. We have worked on two of these embeddings, fastText [14][15] and BERT [15][16] released by Facebook and by Google respectively.

fastText embedding: This is an extension of Word2Vec [17] which uses skip-gram model approach to represent each word in the form of an n-gram of characters. It provides an advantage of being able to understand suffixes, prefixes and shorter words in a more efficient and meaningful way.

BERT embedding: Unlike fastText embedding, this is a context-informed word embedding. It is based on attention models and the architecture of transformers. It is bidirectional in the sense that it trains itself while learning from the left as well as right side of each token. The key advantage that it provides over other embeddings is that it understands the context of the text and then predicts the embeddings for each word, i.e., a word can have different embeddings based on text-context.

E. Deep Learning

Following the generation of word embeddings using above mentioned techniques, these embeddings are fed into the deep neural network algorithms like MLP, CNN and LSTM.

Multi-Layer Perceptron (MLP): They are a form of classical neural networks which are extremely flexible and have proven to have a wide range of applications in case of tabular dataset, such as image, text or time-series. They consist of more than three layers of neurons. They are not of much use in situations requiring consideration of spatial information of the data and can cause redundancy with high-dimensionality input.

Convolutional Neural Networks (CNN): This is another class of neural networks with improved performance over MLP.

These were initially popular for work on images but [15] showcases their use in NLP as well. The layers here go deeper and are connected in a sparse manner, so it detects the patterns in text in a much more efficient way, especially in case of noise or presence of outliers.

LSTM: This is a type of RNN that can understand and learn the context of text along with the text dependency. They use memory cells for updating the hidden layers and hence prove to be extremely effective in learning long-range text dependencies sequentially.

IV. EXPERIMENTAL SETUP

A. Dataset

We performed our analysis on a combination of two datasets one of which was given in HASOC'20 competition and the other is the ALONE dataset. Alone (Adolescents On twitter) is a dataset created from youth's conversations on twitter. It classifies the tweets as toxic(T), non-toxic(N) and Unclear(U). It is made out of high precision from only a particular age group of youth. Meanwhile, the dataset provided in HASOC competition comprises of posts from Facebook and Twitter and it labels data as belonging to either of the two categories, offensive and hate content or not offensive and hate. According to annotators they classified hate, offensive and profane content under the category of hate and offensive speech. So, it was decided to label HOF tag as toxic and Not HOF as non-toxic.

To perform our analysis, both the above-mentioned datasets were merged into one major dataset. This increased the size of training our algorithm and moreover it solved the problem of imbalance labels in ALONE dataset, as shown in Table II.

TABLE II. LABEL DISTRIBUTION

S.No	Dataset	Toxic Count	Non-Toxic Count	Total Count
1.	ALONE (Adolescents On twitter)	118	547	665
2.	HASOC dataset	2261	3591	5852
3.	ALONE-HASOC-Mixed	2379	4138	6517

We performed the 10-fold cross validation procedure, with 90% of the data as training and the remaining as test data.

B. Pre-processing

After its formation, the dataset was cleaned and pre-processed to get rid of the many unnecessary characters which are of no use for classification. Firstly, we removed the stopwords, the names after @ symbol (that mention particular user id), the links provided in posts, the RT abbreviation (Re-Tweet) and convert all the characters to lowercase. These were the general steps that were performed before moving on to further steps. Before classifying the dataset by traditional models, we tokenized the data, introduced part-of-speech

(POS) tagging, performed stemming, used n-gram approach and then finally formed a TF-IDF vectorizer.

C. Experimental Configuration of ML Models

We performed **Logistic Regression** with l2 regularisation and balanced class weight with C=0.01. After that we performed **Decision Trees** with gini criteria in balanced class weight category with random state as 0. To obtain better score we used **Random Forest** with estimators as default i.e., 100. We also used **Support Vector Machine** classifier with linear kernel and regularisation factor as 0.01. We observed that these methods are not very efficient in tracking groups or sequences of words that are important in classifying toxic and non-toxic speech.

After that we moved on to **Ensemble Learning** models to try and increase the accuracies. We used **Boosting**, **Bagging** and **Stacking** techniques. According to the distribution of dataset, stacking and boosting performed better for us than bagging. We used **XGBoost** (Extreme Gradient Boosting) with splits=10 and repeats=3. Stacking gave us the best results among the three methodologies.

D. Embedding

We used two types of embedding techniques, **fastText** (non-context based) and **BERT** (context based). We used a fastText model which was pretrained on 1 Million vocabulary words from English webcrawl and Wikipedia. The BERT embedding model developed by Google was pretrained on uncased Wikipedia content (book_corpus_wiki_en_uncased). For it's fine tuning we used a batch size of 32 with 256 length of sequence and 3 epochs.

E. Experimental Configuration of DNN

We performed a bunch of algorithms here, which are **LSTM**, **CNN** and **MLP** among which CNN performed the best. LSTM with units varying between 50 and 128 followed by a dense layer of 128 and 256 dense units was used. The architecture followed for CNN is as follows. As a result of the previous step, the embedding performed on tweets generated an n-dimensional word vector. This vector was fed into the convolution layer. Max pooling was applied on the features generated from the convolution layer to capture the most important feature, i.e., the feature with maximum value. As a result 1024 most important features were learned from the convolution layer. These features were then sent into hidden layer with 256 perceptrons which together generate 256 higher level features. The newly generated higher level features are used as input to the final output layer with a single perceptron. The hidden layers and output layer used ReLU and sigmoid activation functions respectively.

We assumed LSTM to perform better as it is a widely used model in natural language processing tasks, it tracks the important set of words efficiently. But in our case CNN performed slightly better than LSTM, so we identified that it can detect small set of important words and takes them into consideration adjusting the noise in them. We have mentioned **precision**, **recall**, **F1-score** and **accuracy** of all the algorithms in Table III.

TABLE III. COMPARISON OF RESULTS

	LSTM	CNN	MLP	XGB	LR	DT	RF	SVM	Bagging	Stacking
Precision	0.80	0.83	0.63	0.68	0.67	0.64	0.70	0.45	0.61	0.70
Recall	0.79	0.82	0.64	0.70	0.69	0.64	0.72	0.68	0.65	0.72
F1-score	0.78	0.81	0.64	0.68	0.68	0.64	0.68	0.55	0.63	0.66
Accuracy	0.78	0.82	0.64	0.70	0.69	0.64	0.72	0.68	0.65	0.66

V. CONCLUSION & FUTURE SCOPE

In this paper we have combined two new datasets (ALONE and HASOC'20) to perform our analysis. Alone is a dataset entirely based on youths' toxic conversations on Twitter. We have performed data pre-processing followed by various machine learning and ensemble algorithms using TF-IDF, POS tagging and trigrams approach in which LR and XGBoost performed the best for us. In the second scenario we used word embedding techniques (fastText and BERT) and then fed in their results as inputs to DNN classifiers. We used various DNN classifiers and observed that a combination of BERT embedding with CNN gave us the best results. The ability of CNN to adapt itself to understand and efficiently identify appropriate patterns in case of small sequences of words and noise in dataset, explains better performance of CNN over LSTM.

In our future work, we would like to use transformer attention models in order to increase the score and accuracy. Data augmentation methods like SMOTE will help us in getting rid of and hence in classifying both toxic and non-toxic content with equal probability. We also aim to develop a generic pre-processing model that can work with datasets from multi platforms.

REFERENCES

- [1] T. Davidson, D. Warmesley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *arXiv*, no. 1cwsml, pp. 512–515, 2017.
- [2] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook," *CEUR Workshop Proc.*, vol. 1816, pp. 86–95, 2017.
- [3] P. L. Teh, C. Bin Cheng, and W. M. Chee, "Identifying and categorising profane words in hate speech," *ACM Int. Conf. Proceeding Ser.*, pp. 65–69, 2018.
- [4] A. Mittos, S. Zannettou, J. Blackburn, and E. De Cristofaro, "'And we will fight for our race!' a measurement study of genetic testing conversations on reddit and 4chan," *Proc. 14th Int. AAAI Conf. Web Soc. Media, ICWSM 2020*, no. 1cwsml, pp. 452–463, 2020.
- [5] A. Olteanu, C. Castillo, J. Boy, and K. R. Varshney, "The effect of extremist violence on hateful speech online," *arXiv*, no. 1cwsml, pp. 221–230, 2018.
- [6] P. Fortuna and S. Nunes, "A survey on automatic

- detection of hate speech in text,” *ACM Comput. Surv.*, vol. 51, no. 4, 2018.
- [7] Z. Waseem, J. Thorne, and J. Bingel, “Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection,” *Springer International Publishing*, 2018.
- [8] A. Gaydhani, V. Doma, S. Kendre, and L. Bhagwat, “Detecting hate speech and offensive language on twitter using machine learning: An N-gram and TFIDF based approach,” *arXiv*, 2018.
- [9] Z. Zhang, D. Robinson, and J. Tepper, “Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network,” vol. 10843 *LNCS. Springer International Publishing*, 2018.
- [10] T. Ranasinghe, M. Zampieri, and H. Hettiarachchi, “BRUMS at HASOC 2019: Deep learning models for multilingual hate speech and offensive language identification,” *CEUR Workshop Proc.*, vol. 2517, pp. 199–207, 2019.
- [11] O. Sagi and L. Rokach, “Ensemble learning: A survey,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 4, pp. 1–18, 2018.
- [12] S. Malmasi and M. Zampieri, “Challenges in discriminating profanity from hate speech,” *J. Exp. Theor. Artif. Intell.*, vol. 30, no. 2, pp. 187–202, 2018.
- [13] V. Indurthi, B. Syed, M. Shrivastava, N. Chakravartula, M. Gupta, and V. Varma, “FERMI at SemEval-2019 Task 5: Using Sentence embeddings to Identify Hate Speech Against Immigrants and Women in Twitter,” pp. 70–74, 2019.
- [14] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” *15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf.*, vol. 2, pp. 427–431, 2017.
- [15] A. G. D’Sa, I. Illina, and D. Fohr, “BERT and fastText Embeddings for Automatic Detection of Toxic Speech,” *Proc. 2020 Int. Multi-Conference Organ. Knowl. Adv. Technol. OCTA 2020*, 2020.
- [16] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.
- [17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc.*, pp. 1–12, 2013.
- [18] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, “Hate speech detection with comment embeddings,” *Proceedings of the 24th International Conference on World Wide Web*, pp. 29–30, 2015.
- [19] T. Wijesiriwardene, H. Inan, U. Kursuncu, M. Gaur, V. L. Shalin, K. Thirunarayan, A. Sheth, and I. B. Arpinar, “Alone: A dataset for toxic behavior among adolescents on twitter,” *International Conference on Social Informatics*, pp. 427–439, 2020.
- [20] R. Ottoni, E. Cunha, G. Magno, P. Bernardino, W. Meira Jr, and V. Almeida, “Analyzing right-wing youtube channels: Hate, violence and discrimination,” *Proceedings of the 10th ACM Conference on Web Science*, pp. 323–332, 2018.

Transfer Learning for Detection of COVID-19 Infection using Chest X-Ray Images

Nikhil Bhatia

Department of Information Technology,
Delhi Technological University,
New Delhi, India
nikhilbhatia346@gmail.com

Geetanjali Bhola

Assistant Professor,
Department of Information Technology,
Delhi Technological University,
New Delhi, India
geetanjali@dtu.ac.in

Abstract— Coronavirus is a contagious disease that affects individuals in a large scale. Coronavirus had a huge impact on the nation's economy and human lifestyle. The motivation behind this study was establishing a better diagnosis test for coronavirus infection. The RT-PCR test is used to diagnose the coronavirus frequently and returned a negative result for an infected individual. Furthermore, this test remains prohibitively expensive for most citizens, and not everyone could afford it due to financial hardship. An efficient imaging approach is developed for the evaluation of lung conditions, which has been done by examining the chest X-ray or chest CT of an infected person. Deep Learning is the well-suited sub domain of Artificial Intelligence [AI] technology, which offers helpful examination to consider more number of chest X-rays images that can basically have an effect on coronavirus screening. The goal of this research is to cluster the radiograph images present in the dataset into COVID-19, healthy and viral pneumonia by making use of the artificial neural networks. The training dataset was fine-tuned with eleven previously trained convolutional neural architectures. The assessment of the models on a test sample shows that AlexNet, DenseNet-121, GoogleNet and Squeezenet1.1 as the top performing models.

Keywords— Coronavirus, normal, viral pneumonia, chest x-ray images, deep learning, artificial intelligence, deep neural networks.

I. INTRODUCTION

COVID-19 is induced by a modern form of coronavirus and is a contagious infection. Common signs of infection are fever, cough, respiratory symptoms and breathing difficulties [1]. In Wuhan, China, the introduction of this disease into the human population was first recorded at the end of 2019. A clinical review of coronavirus shows that a person can catch COVID-19 when he/she comes into contact with an infected person followed by respiratory organ contamination. Also known as Reverse Transcription Enzyme Chain Reaction, RT-PCR could be a check habitually accustomed to verify the signs of the sickness in associate degree infected person. Accuracy of this test depends on which stage the disease is and the standard of the specimen collected from the infected patient [2]. More significantly, the output of the RT-PCR test takes about 48 hours, which is also an expensive test for some people. Many developing countries lacked adequate test kits to test every citizen. The RT-PCR often declares a person with virus as negative, then that person interacts with others, resulting in transmitting the virus to others. For these reasons a quicker and comparatively cheap technology for identifying COVID-19 is needed [3]. Therefore, artificial intelligence (AI) applications can help the hospitals and nations to quickly examine the CT findings or chest x-beams

and diagnose COVID-19 positive patients. Many diseases have been successfully diagnosed using chest CT images and x-beams, and these can also be used for the detection of coronavirus among people.

Coronavirus is a respiratory disease as it can proceed through the respiratory tract and cause respiratory disorder commonly known as pneumonia into a person's lungs. This results in person not getting enough oxygen and carbon dioxide from being cleared out from the bloodstream. This can occur to any spectrum of the age group [4]. People with comorbidities i.e., with the presence of one or more additional conditions after co-occurring with a primary condition cannot take in adequate oxygen or expel sufficient carbon dioxide which leads to life-threatening disorders. It also becomes life-threatening for people who smoke cigarettes, are exposed to chemical fumes at factories, asthma, etc.

Artificial Intelligence should be used in coronavirus testing because money related expenses of the lab units utilized for conclusion, particularly for underdeveloped nations, are a huge issue when battling the disease. Utilizing X- beam pictures for the robotized identification of COVID-19 can be useful specifically for nations and clinics that can't buy a lab unit for tests. For health systems, it is very important to have technical tools that can help them effectively collect patient data and enter test information. AI tools can save doctors and nurses a total of up to 50 hours of data entry time per day, and this number is projected to grow as more experiments are conducted in the society. It can help hospitals and clinics to provide the best care for patients. AI platforms can use a variety of predictive or forecasting models to help prepare for increase in the number of coronavirus patients, track the number of patients in local health center and the ratio of confirmed COVID-19 cases, death cases and recovery cases. AI strategies can help to obtain results quickly and highly accurately [5].

Transfer learning technique is used for diagnosing the coronavirus disease by making use of chest X-rays images. The process of using a previously trained architecture on a previous dataset for a different concept is transfer learning. The convolutional neural architecture learns about the input image to predict the class of the image. The features extracted from the input image can also be same for the other class of input images. Training a convolutional neural network from scratch takes a lot of effort and time, but transfer learning makes it easy. There are two possible approaches to solve a new problem using transfer learning. First technique is to fix the weights of some of the pre-trained architecture layers and train the remaining layers on

the new problem. Second technique is to train a new CNN architecture and use some of the features learned from the pre-trained architecture's layers into the new model. Some of the features in both approaches come from the pre-trained architecture, while the others come from the new dataset. The rest of the model learns to fit the new problem. In Fig. 1, the convolutional neural architecture (namely AlexNet, GoogleNet, VGG-16, VGG-11, ResNet-18, ResNet-34, DenseNet-121, Densenet-169, ShuffleNet, SqueezeNet 1.1, SqueezeNet 1.0) are trained on class of images (problem A) to extract features. Then these pre-trained architectures use the extracted features or knowledge gained on problem A for prediction of new class of images (problem B). This is useful rather than training a new convolutional neural network (CNN) from the ground up for problem B.

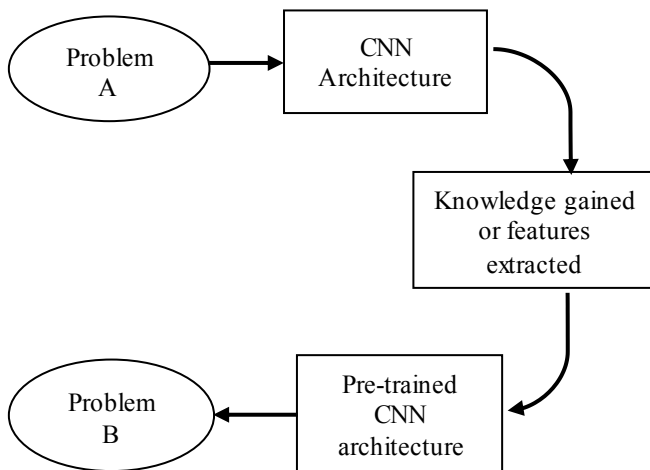


Fig. 1. Steps in Transfer Learning algorithm

Transfer learning is very popular and is widely used because it can train neural networks with comparatively little or inconsistent data [6]. The deep neural networks used in this study have previously been trained to acknowledge thousands of groups on large set of images, as in the ImageNet database, which consists of different pictures of pencils, animals, plants, buildings, fabrics, etc. composition. Deep Learning is the AI's best methodology, which helps to understand a variety of X-ray pictures of chest that could have a fundamental impact on disease detection. There are almost 3000 chest x-beam scans in the dataset that are of Coronavirus, Viral Pneumonia and Healthy. The aim of this experimentation is to design and compare various convolutional neural networks that can classify chest X-ray images according to the three categories with reasonable level of accuracy. On the basis of the chest X-ray data, various deep learning-based architectures were trained to predict whether the person had coronavirus, viral pneumonia, or strong lungs.

II. DETECTING COVID-19 USING CHEST X-RAY IMAGES

Detection of coronavirus disease was done using various image classification architectures. The input X-rays were resized and normalized before training, to make them appropriate for the architecture's input. The neural networks that had been previously trained on the ImageNet database were retrained using the COVID-19 training dataset. The COVID-19 testing dataset was then fed into these pre-trained

architectures, and the X-ray pictures were grouped into three categories: COVID-19, Normal, and Viral Pneumonia. The block diagram for coronavirus detection in chest X-ray images using pre-trained architecture is shown in Fig 2.

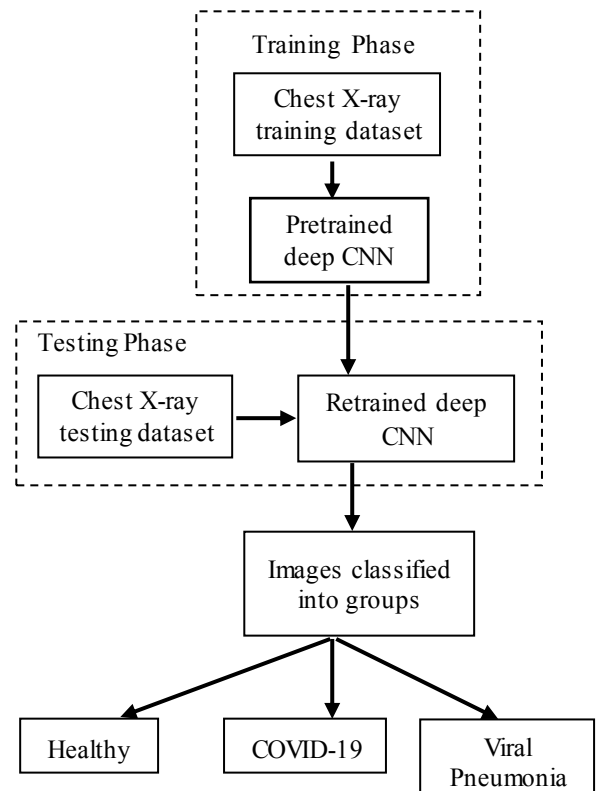


Fig. 2. Block Diagram for COVID-19 prediction in chest X-ray images using a pre-trained architecture.

III. RELATED WORK

Earlier studies have made use of chest X-ray scans and artificial intelligence to predict coronavirus. Pranav Rajpurkar in [7] created CheXNet, an architecture of 121-layers and trained it on a dataset called ChestX-ray14, which comprises of the front view of chest X-ray pictures with fourteen different lung illnesses. They also substituted the final fully connected layer with one that only has one output and also applied sigmoid nonlinearity to get expected probabilities. Tulin Ozturk in [8] used DarkCovidNet architecture to classify coronavirus from chest X-ray images. They considered 1127 chest X-ray images. For binary classes (COVID-19 vs. Healthy), the network had an accuracy of 98.08 percent, and for three-class situations, it had an accuracy of 87.02 percent (COVID-19 vs. Normal vs. Pneumonia). Ioannis D. Apostolopoulos in [9] considered and provided results for various deep neural networks such as VGG19, Xception, Inception, Inception ResNet V2 and MobileNet V2 using two datasets in their research. The first dataset contains 1428 chest X-rays. The other one contains 1442 X-ray images. Their best models were VGG19 and MobileNet V2. VGG19 achieved 98.75 percent accuracy for binary-class situations (COVID-19 vs Pneumonia and COVID-19 vs Healthy), 93.48 percent accuracy for three-class situations (COVID-19 vs. Normal vs. Pneumonia). MobileNet V2 achieved 97.40 percent accuracy for binary-class situations, 92.85 percent accuracy for multi-class situations. Shervin Minaee in [10] created a chest X-ray

sample of 5000 images. They trained SqueezeNet, ResNet50, DenseNet-121 and ResNet18 to recognize coronavirus infection. Their best model had a sensitivity rate of 98 percent and 92.9 percent of specificity. The authors also generated heatmaps of COVID-19 infected lungs using a method. Linda Wang in [11] introduced COVID-Net, an architecture built specifically for detecting coronavirus in chest X-ray pictures. For training and testing the generated architecture, the authors used 13,975 chest X-ray pictures. In addition, they investigated how COVID-Net uses explainability methods to make predictions in order to gain insight into key factors related to coronavirus cases. The proposed model, COVID-Net achieved 93.3 percent accuracy. Boran Sekeroglu in [12] considered 1583 images of healthy cases, 4292 images of pneumonia infection, and 225 images of COVID-19 infection for their study. They used convolutional neural networks in 38 test experiments, 5 neural networks in 10 experiments, and pre-trained networks in 14 experiments. Their models achieved 93.84 percent mean sensitivity rate, 99.18 percent mean specificity rate, 98.50 percent mean accuracy and 96.51 percent mean receiver operating characteristic score. Mohamed Elgendi in [3] compared the efficiency of 17 neural networks for predicting COVID-19 pneumonia. The authors used two datasets for the training and evaluation of the neural networks. The first dataset contains 85 COVID-19 infected chest X-rays, 2,772 images bacterial, and 1,493 viral pneumonia infected chest X-rays. The second dataset contains only 85 COVID-19 infected X-rays. Their results showed that DarkNet-19 was most efficient in diagnosing the infection. DarkNet-19 achieved 94.28 percent accuracy. Matteo Polsinelli in [13] developed a light architecture based on SqueezeNet architecture, to successfully recognize COVID-19 CT images in other CT images. They evaluated the output of their neural network on two arrangements of datasets. The first training dataset arrangement contains 191 Non COVID-19 CT scans and 191 COVID-19 CT scans. The other training dataset arrangement has 191 Non COVID-19 CT scans and 251 COVID-19 CT scans. On the test datasets, the proposed revised SqueezeNet architecture had an accuracy of 83.00 percent, a sensitivity of 85.00 percent, a specificity of 81.00 percent, a precision of 81.73 percent, and an F1Score of 0.8333. Majid Nour in [14] developed a CNN architecture containing five convolutional layers and trained the architecture from scratch. The features extracted by the neural network from the chest X-rays were leveraged into various algorithms such as decision trees, k-nearest neighbor and support vector machine. Then the authors used Bayesian optimization algorithm for the hyperparameter tuning. SVM classifier was the most accurate and reliable with 98.97 percent accuracy, 89.39 percent sensitivity, 99.75 percent specificity, 96.72 percent F-score. Lawrence O. Hall in [15] considered a small dataset of 135 chest X-ray pictures and 320 of other pneumonia cases. In a 10-fold cross validation, they retrained ResNet-50, a deep CNN architecture, on 102 other pneumonia chest X-rays and 102 COVID-19 infected chest X-rays. They were able to reach an accuracy of 89.2 percent. Then they trained a combination of VGG16, ResNet-50 and their own small neural network on a balanced dataset. An overall accuracy of 91.24 percent and AUC of 0.94 was achieved. Asmaa Abbas in [16] utilized Decompose, Transfer, and Compose (DeTraC), a deep architecture for the detection of coronavirus in chest X-ray pictures. The dataset contains 11 SARS, 105 COVID-19 and 80 healthy chest X-ray pictures. ResNet18 pre-trained

architecture was used in DeTraC's transfer learning section. DeTraC- ResNet-18 achieved accuracy of 95.12 percent, sensitivity of 97.91 percent, and specificity of 91.87 percent. Ezz El-Din Hemdan in [17] developed COVIDX-Net, which contains 7 different frameworks of deep convolutional neural network, namely VGG-19 and Google MobileNet. The performance of the network was tested using 50 chest X-ray images. The neural network achieved f1-scores of 0.91 and 0.89 for COVID-19 and normal cases. InceptionV3 network performed the worst with f1-scores of 0.00 and 0.67 for COVID-19 instances and normal instances. Halgurd S. Maghdid in [18] considered a dataset of both CT scan and chest X-ray images for their research work. They proposed a simple convolutional neural network with one convolutional layer and sixteen filters each of which using a 5x5 filter scale, rectified linear unit, batch normalization and a few other connected layers. They also used a previously trained architecture, AlexNet on the CT images and chest X-rays. The updated convolutional neural network was 94.1 percent accurate and 98 percent accurate via pre-trained architecture in diagnosing COVID-19. Arman Haghani in [19] used a collection of 780 coronavirus infected chest X-ray images. The authors developed a powerful architecture COVID-CXNet using a popular transfer learning network, CheXNet. They also created a lung segmentation module to improve the model localization of lung abnormalities. The highest accuracy achieved was by Base Model v2 with a score of 98.68 percent and COVID-CXNet v1 with a score of 99.04 percent. Asif Iqbal Khan in [20] used a pre-trained architecture, Xception as a base model for developing CoroNet. For a four-class experiment (COVID-19 / Bacterial pneumonia/Healthy/Viral pneumonia), CoroNet achieved an accuracy of 89.6 percent and 95 percent accuracy for three class experiment (COVID-19 vs Healthy vs Pneumonia). Amit Kumar Jaiswal in [21] used Mask-RCNN as a base model for developing their architecture. Mask-RCNN is a form of artificial neural network that was created to solve the problem of instance segmentation. Dr. R. Dhaya in [22] presented an enhancement technique to upgrade image processing in order to reduce the issues in user immersion. To enhance the image quality, the proposed technique employs the Retinex algorithm and Otsu's method is used to improve the processing speed. The image sets 1,2,3,4,5 yield -55, -43, -23, -49, -45 and 50, 40, 45, 49, 55 for conventional and suggested algorithm, respectively. Dr. Abraham Chandy in [23] proposed an algorithm which uses a resource allocation method to reduce the expense, time, memory, and power consumption of a massive data processing system. Machine learning methods like random forest is used in the proposed system to forecast work load and assign resources using the genetic algorithm. The obtained outcomes showed the proposed method's capability using prediction accuracy, device level features and resource usage.

Unlike all the related works, this research focus on doing a comparative study of all the deep learning based pre-trained neural networks and finding the most accurate and reliable architecture to predict COVID-19 infection by providing strong and promising evidences.

IV. DATASET

A total of 219 pictures with coronavirus infection, 1341 pictures of healthy cases and 1344 pictures with viral pneumonia were considered from Kaggle for the prediction of coronavirus. All pictures are modified to 224 x 224 pixels.

RandomHorizontalFlip was used after resizing to transform images in the training dataset. It returned some images as original and some images as flipped. Then the resized images in the dataset were transformed to PyTorch tensor. Then these tensor images were normalized with standard deviation = [0.229, 0.224, 0.225] and mean = [0.485, 0.456, 0.406]. There are 2,814 images in the training collection (1311 normal images, 1314 viral pneumonia images, and 189 images with coronavirus infection) and the test collection was prepared such that it includes 90 images chosen at random from the dataset (30 healthy images, 30 images with coronavirus infection, and 30 images with viral pneumonia).

Number of images in a batch = 6
Number of training batches = 470
Number of test batches = 15

Few samples from the chest X-ray dataset are shown below.

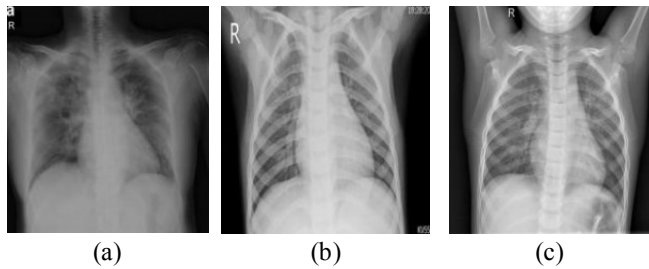


Fig. 3. Samples from the dataset. (a) COVID-19 infected chest X-ray; (b) Viral Pneumonia infected chest X-ray (c) Healthy chest X-ray.

V. METHODOLOGY

Since the number of chest X-ray instances considered in this research are small and insufficient to train a machine learning architecture from scratch, This research makes use of Deep Transfer Learning technique. In this paper, the proposed method uses pre-trained models, namely AlexNet [24], GoogleNet [25], VGG-16, VGG-11 [26], ResNet-18, ResNet-34 [27], DenseNet-121, Densenet-169 [28], ShuffleNet [29], SqueezeNet 1.1, SqueezeNet 1.0 [30]. The pre-trained version of these networks, which has learned on the images present in the ImageNet dataset, is simple to load. These neural networks have extracted features from broad range of images.

COVID-19 X-ray pictures were differentiated from healthy and viral pneumonia classes using these models. These neural networks were modified to output only three categories (COVID-19, Healthy, Viral Pneumonia) instead of thousand classes. All the CNN architectures were fine-tuned on the training sample. The models were trained using the cross entropy loss function, that attempts to reduce the gap between the predicted output and the actual probability. Cross entropy loss is specified as:

$$CE = - \sum_x^N p_x \log q_x \quad (1)$$

The actual and expected probabilities are represented by p_x and q_x , respectively. The training is done for 10 epochs and the model was evaluated at every 20th training step, with a batch size of 6. At every training step, Loss.backward() is used for back propagation. It computes the loss gradient for

all the loss parameters with requires_grad = True and stores the result in parameter.grad variable for every parameter. After calculating all the values of gradients for tensors in the network, optimizer.step() function is used. On the basis of parameter.grad, optimizer.step() adjusts all parameters. The optimizer.zero_grad() function is then called, which sets the gradients from the previous batch to zero. The loss.item() function is used to measure training and validation loss, which is the loss of the whole batch divided by the batch size.

Classification rate, specificity, recall, precision are just a few of the metrics that can help in evaluating a convolutional neural network's performance. Specificity, sensitivity and precision are three appropriate indicators for reporting a network's work as the current test data set is highly uneven.

$$\text{Sensitivity} = \frac{\text{True positive (TP)}}{\text{Total positive COVID 19 Images}} \quad (2)$$

$$\text{Specificity} = \frac{\text{True negative (TN)}}{\text{Total negative COVID 19 Images}} \quad (3)$$

$$\text{Precision} = \frac{\text{True positive (TP)}}{\text{True positive (TP)} + \text{False positive (FP)}} \quad (4)$$

In equation 2, True positive is the total predictions where the architecture predicts the positive class correctly. In equation 3, True negative is the total predictions where the architecture predicts the negative class correctly. In equation 4, False positive is the total predictions where the architecture predicts negative class as positive.

VI. EXPERIMENTAL RESULTS

The training and evaluation of these convolutional neural networks is carried out in Jupyter environment and Google colab using python programming language. Google colab has a powerful GPU processor that makes training large models like VGG very simple and fast. Machine learning libraries like PyTorch and torchvision are used in this experiment. Many common model architectures can be found in torchvision module. For training and testing the architectures, Intel Core i5 7th generation computer with a RAM of 8gb is used. The models were trained using the ADAM optimizer to optimize the loss function with a learning rate of 3e-5.

• AlexNet

	0	1	2
Truth 0	28	0	0
1	0	30	0
2	0	2	30
	0	1	2
	predicted		

Fig. 4. Confusion matrix of AlexNet

TABLE I. SPECIFICITY, SENSITIVITY AND PRECISION OF ALL THE THREE CLASSES FOR ALEXNET

		Specificity	Sensitivity	Precision
0	COVID-19	100%	100%	100%
1	Viral Pneumonia	96.6%	100%	93.75%
2	Normal	100%	93.7%	100%

- DenseNet-121**

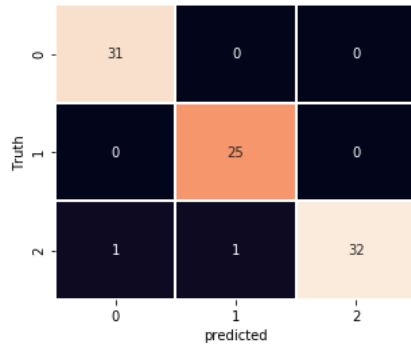


Fig. 5. Confusion matrix of DenseNet-121

TABLE II. SPECIFICITY, SENSITIVITY AND PRECISION OF ALL THE THREE CLASSES FOR DENSENET-121

		Specificity	Sensitivity	Precision
0	COVID-19	98.2%	100%	96.87%
1	Viral Pneumonia	98.4%	100%	96.15%
2	Normal	100%	94.1%	100%

- DenseNet-169**

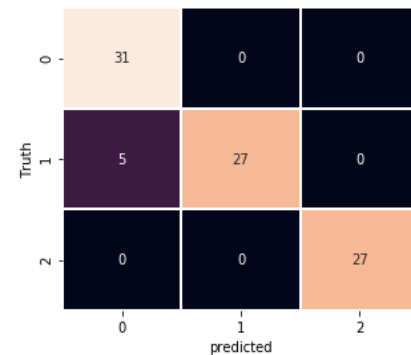


Fig. 6. Confusion matrix of DenseNet-169

TABLE III. SPECIFICITY, SENSITIVITY AND PRECISION OF ALL THE THREE CLASSES FOR DENSENET-169

		Specificity	Sensitivity	Precision
0	COVID-19	91.5%	100%	86.11%
1	Viral Pneumonia	100%	84.3%	100%
2	Normal	100%	100%	100%

- GoogleNet**

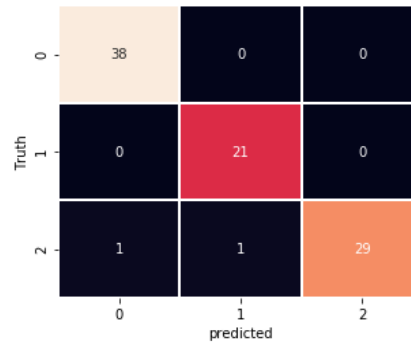


Fig. 7. Confusion matrix of GoogleNet

TABLE IV. SPECIFICITY, SENSITIVITY AND PRECISION OF ALL THE THREE CLASSES FOR GOOGLNET

		Specificity	Sensitivity	Precision
0	COVID-19	98%	100%	97.43%
1	Viral Pneumonia	98.5%	100%	95.45%
2	Normal	100%	93.5%	100%

- ResNet-34**

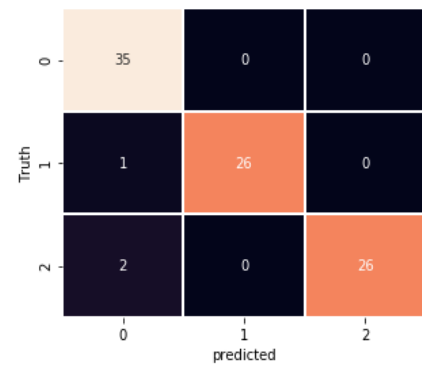


Fig. 8. Confusion matrix of ResNet-34

TABLE V. SPECIFICITY, SENSITIVITY AND PRECISION OF ALL THE THREE CLASSES FOR RESNET-34

		Specificity	Sensitivity	Precision
0	COVID-19	94.5%	100%	92.10%
1	Viral Pneumonia	100%	96.2%	100%
2	Normal	100%	92.8%	100%

- ResNet-18**

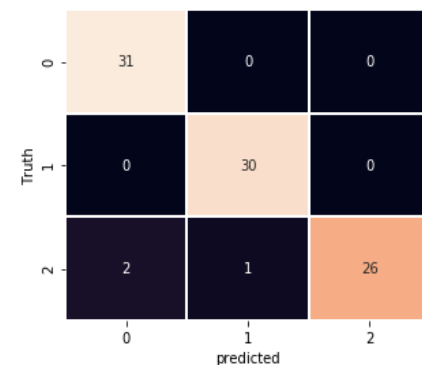


Fig. 9. Confusion matrix of ResNet-18

TABLE VI. SPECIFICITY, SENSITIVITY AND PRECISION OF ALL THE THREE CLASSES FOR RESNET-18

		Specificity	Sensitivity	Precision
0	COVID-19	96.5%	100%	93.93%
1	Viral Pneumonia	98.2%	100%	96.77%
2	Normal	100%	89.6%	100%

• ShuffleNet

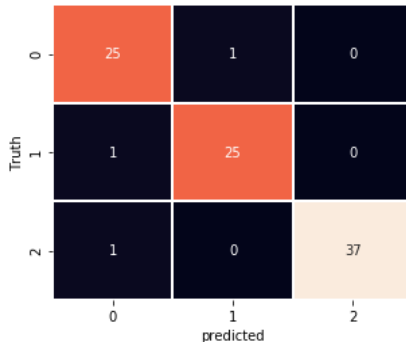


Fig. 10. Confusion matrix of ShuffleNet

TABLE VII. SPECIFICITY, SENSITIVITY AND PRECISION OF ALL THE THREE CLASSES FOR SHUFFLENET

		Specificity	Sensitivity	Precision
0	COVID-19	96.8%	96.1%	92.59%
1	Viral Pneumonia	98.4%	96.1%	96.15%
2	Normal	100%	97.3%	100%

• VGG-11

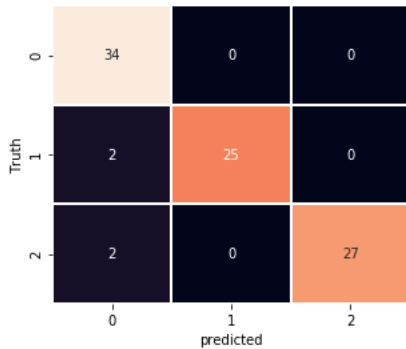


Fig. 11. Confusion matrix of VGG-11

TABLE VIII. SPECIFICITY, SENSITIVITY AND PRECISION OF ALL THE THREE CLASSES FOR VGG-11

		Specificity	Sensitivity	Precision
0	COVID-19	92.8%	100%	89.47%
1	Viral Pneumonia	100%	92.5%	100%
2	Normal	100%	93.1%	100%

• VGG-16

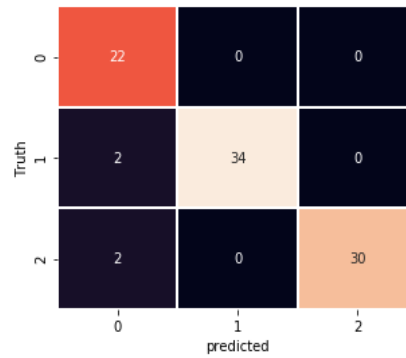


Fig. 12. Confusion matrix of VGG-16

TABLE IX. SPECIFICITY, SENSITIVITY AND PRECISION OF ALL THE THREE CLASSES FOR VGG-16

		Specificity	Sensitivity	Precision
0	COVID-19	94.1%	100%	84.61%
1	Viral Pneumonia	100%	94.4%	100%
2	Normal	100%	93.7%	100%

• Squeezenet1.0

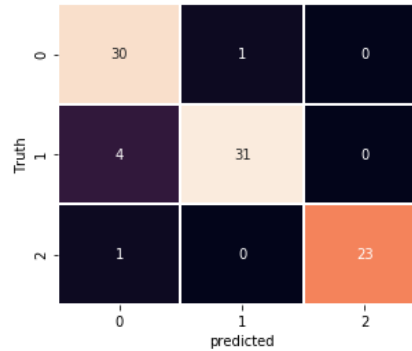


Fig. 13. Confusion matrix of Squeezenet1.0

TABLE X. SPECIFICITY, SENSITIVITY AND PRECISION OF ALL THE THREE CLASSES FOR SQUEEZENET1.0

		Specificity	Sensitivity	Precision
0	COVID-19	91.5%	96.7%	85.71%
1	Viral Pneumonia	98.1%	88.5%	96.87%
2	Normal	100%	95.8%	100%

• Squeezenet1.1

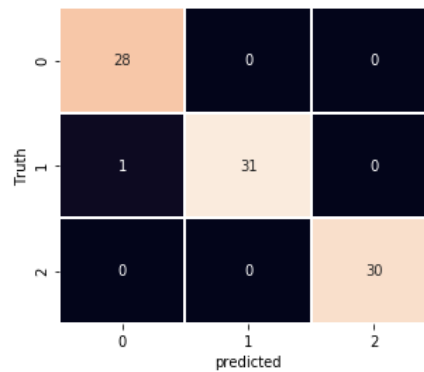


Fig. 14. Confusion matrix of Squeezenet1.1

TABLE XI. SPECIFICITY, SENSITIVITY AND PRECISION OF ALL THE THREE CLASSES FOR SQUEEZENET1.1

		Specificity	Sensitivity	Precision
0	COVID-19	98.3%	100%	96.55%
1	Viral Pneumonia	100%	96.8%	100%
2	Normal	100%	100%	100%

TABLE XII. ACCURACY OF DIFFERENT PRETRAINED DEEP NEURAL NETWORKS

AlexNet	97.78%
DenseNet-121	97.78%
DenseNet-169	94.44%
GoogleNet	97.78%
ResNet-34	96.67%
ResNet-18	96.67%
ShuffleNet	96.67%
VGG-11	95.56%
VGG-16	95.56%
Squeezenet1.0	93.33%
Squeezenet1.1	98.89%

The best models are indicated in bold in Table XII. The best performing models on the dataset are AlexNet, DenseNet-121, GoogleNet and Squeezenet1.1. The lowest accuracy was of Squeezenet1.0. Even though AlexNet has the most input features, SqueezeNet1.1 was able to achieve greater accuracy than AlexNet because SqueezeNet1.1 has 50x less parameters and a smaller model size than AlexNet [30]. Also, SqueezeNet1.1 was able to achieve more accuracy than SqueezeNet1.0 on the same dataset because it is a better model than SqueezeNet1.0. It computes 0.72 GFLOPS per image, while SqueezeNet1.0 computes 1.72 GFLOPS per image which 2.4 times less and has a few less parameters than SqueezeNet 1.0.

VII. CONCLUSION

The pandemic had a negative impact on the economy of many countries and everyday lives of individuals round the world. Since the year December 2019, the number of deaths due to the virus has been continued to increase around the world. AI has helped human population to win many battles, and is still aiding mankind in the arduous struggle against coronavirus. Although AI has advanced a lot in the past decades, still it is trying very hard to keep up against the illness. This is true, because the data related to COVID-19 is limited and AI techniques usually require abundance of data to grasp something or acquire understanding about the problem. However, the amount of AI research related to coronavirus will grow remarkably as soon as more COVID-19 data is available. The future work of establishing, hosting, and benchmarking COVID-19-related data sets is essential, as this will help speed up the discovery of discoveries that are useful to address the disease.

Among all the research published, the application of deep transfer learning technique in the prediction of coronavirus by making use of chest X-ray images appears to dominate. The current best quality level research facility tests are tedious and exorbitant, adding deferrals to the testing cycle. Chest X-rays is generally accessible and moderate for examining patients with little to none symptoms or a doubt about coronavirus. Expansion of artificial intelligence assisted radiography can help in upgrading the output and

before time determination of the infection; this is particularly obvious during a widespread, and in regions with a deficiency of radiologists. Deep learning models can cause better performance in identification of the disease when adequate data is available until then safety precautions are needed when calculating the work of the CNN architectures. The outcomes introduced here are basic because of the shortage of images used in the testing stage.

In this study, the performance of several pretrained architectures were reviewed for detecting the radiographic features in X-rays. After analyzing the pre-trained architectures, Squeezenet1.1, AlexNet, DenseNet-121, GoogleNet were better models than the others in identifying the coronavirus in the chest X-rays.

The current COVID-19 dataset is small, making it difficult to train a neural network from scratch. To improve the accuracy of these neural networks, more experimentation on the models with a larger dataset is required. As a result, attempts to gather more data on COVID-19 are still in progress.

VIII. REFERENCES

- [1] COVID-19 questions and answers [WHO EMRO | Questions and answers | COVID-19 | Health topics](#)
- [2] Emery SL, Erdman DD, Bowen MD, et al. Real-time reverse transcription-polymerase chain reaction assay for SARS-associated coronavirus. *Emerg Infect Dis.* 2004;10(2):311-316. doi:10.3201/eid1002.030759.
- [3] Elgendi M, Nasir MU, Tang Q, Fletcher RR, Howard N, Menon C, Ward R, Parker W and Nicolaou S (2020) The performance of Deep Neural Networks in differentiating Chest X-Rays of COVID-19 patients from other Bacterial and Viral Pneumonias. *Front. Med.* 7:550. doi: 10.3389/fmed.2020.00550.
- [4] Facts about Pneumonia [Learn About Pneumonia | American Lung Association](#)
- [5] Artificial Intelligence simplifies COVID-19 testing: [Artificial Intelligence Simplifies COVID-19 Testing, Workflows \(healthitanalytics.com\)](#)
- [6] Transfer Learning: [What Is Transfer Learning? A Simple Guide | Built In](#)
- [7] Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz CP, Patel BN, Yeom KW, Shpanskaya K, Blankenberg FG, Seekins J, Amrhein TJ, Mong DA, Halabi SS, Zucker EJ, Ng AY, Lungren MP. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* 2018 Nov 20;15(11):e1002686. doi: 10.1371/journal.pmed.1002686. PMID: 30457988; PMCID: PMC6245676.
- [8] Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-

- ray images. *Comput Biol Med.* 2020;121:103792. doi:10.1016/j.compbio.2020.103792.
- [9] Apostolopoulos, I.D., Mpesiana, T.A. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. *Phys Eng Sci Med* 43, 635–640 (2020). <https://doi.org/10.1007/s13246-020-00865-4>.
- [10] Minaee S, Kafieh R, Sonka M, Yazdani S, Soufi GJ. Deep-covid: Predicting covid-19 from chest X-ray images using deep transfer learning. arXiv preprint. 2020; arXiv:2004.09363.
- [11] Wang L and Wong A., COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images, 2020, arXiv, eess.IV, 2003.09871.
- [12] Sekeroglu B, Ozsahin I. Detection of COVID-19 from Chest X-Ray Images Using Convolutional Neural Networks. *SLAS TECHNOLOGY: Translating Life Sciences Innovation.* 2020;25(6):553-565. doi:10.1177/2472630320958376.
- [13] Polsinelli M, Cinque L, Placidi G, A Light CNN for detecting COVID-19 from CT scans of the chest; arXiv:2004.12837.
- [14] Majid Nour, Zafer Cömert, Kemal Polat, A Novel Medical Diagnosis model for COVID-19 infection detection based on Deep Features and Bayesian Optimization, *Applied Soft Computing*, Volume 97, Part A, 2020, 106580, ISSN 1568-4946, <https://doi.org/10.1016/j.asoc.2020.106580>.
- [15] Hall L. O., Paul R, Goldgof D. B., Goldgof G. M, Finding Covid-19 from Chest X-rays using Deep Learning on a Small Dataset; arXiv:2004.02060.
- [16] Abbas A, Abdelsamea M. M. and Gaber M. M., Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network, 2020, arXiv :2003.13815.
- [17] Ezz El-Din Hemdan and Marwa A. Shouman and Mohamed Esmail Karar, COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images, 2020, arXiv: 2003.11055.
- [18] Maghdid H. S., Asaad A. T., Ghafoor K. Z., Sadiq A. S., Khan M. K., Diagnosing COVID-19 Pneumonia from X-Ray and CT Images using Deep Learning and Transfer Learning Algorithms; arXiv:2004.00038.
- [19] Haghanifar A, Majdabadi M. M., Choi Y., Deivalakshmi S., Ko S., COVID-CXNet: Detecting COVID-19 in Frontal Chest X-ray Images using Deep Learning; arXiv:2006.13807.
- [20] Khan A. I., Shah J. L., Mohammad, Bhat M, CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images; arXiv:2004.04931.
- [21] Amit Kumar Jaiswal, Prayag Tiwari, Sachin Kumar, Deepak Gupta, Ashish Khanna, Joel J.P.C. Rodrigues, Identifying pneumonia in chest X-rays: A deep learning approach, *Measurement*, Volume 145, 2019, Pages 511-518, ISSN 0263-2241.
- [22] Dhaya, R. "Improved Image Processing Techniques for User Immersion Problem Alleviation in Virtual Reality Environments." *Journal of Innovative Image Processing (JIIP)* 2, no. 02 (2020): 77-84.
- [23] Chandy, Abraham. "Smart resource usage prediction using cloud computing for massive data processing systems." *J Inf Technol* 1, no. 02 (2019): 108-118.
- [24] Krizhevsky A, Sutskever I, Hinton G.E., Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- [25] Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D, Going Deeper with Convolutions; arXiv:1409.4842.
- [26] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition; arXiv:1409.1556.
- [27] He K., Zhang X., Ren S., and Sun, J. (2016). Deep residual learning for image recognition; arXiv:1512.03385.
- [28] Huang G., Liu Z., Van Der Maaten L., and Weinberger K. Q. (2017), Densely connected convolutional networks; arXiv:1608.06993.
- [29] Ma N., Zhang X., Zheng H and Sun J., ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design; arXiv:1807.11164.
- [30] Iandola, F. N., Han S., Moskewicz M. W., Ashraf K, Dally W. J., and Keutzer K, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size; arXiv:1602.07360.

Two-Stage Stochastic Programming Model for Optimal Scheduling of RES-Based Virtual Power Plants in Electricity Markets

Meegada Indeevar Reddy
Department of Electrical Engineering,
Delhi Technological University
Delhi, India
Indeevarreddy_mt2k19@dtu.ac.in

Radheshyam Saha
Department of Electrical Engineering,
Delhi Technological University
Delhi, India
rshahacno@yahoo.com

Sudarshan K. Valluru
Department of Electrical Engineering,
Delhi Technological University
Delhi, India
sudarshan_valluru@dce.ac.in

Abstract— To promote investment in the electricity sector, the deregulated electricity market regime has created an enabling environment to accelerate the all-around development of power generation, transmission and distribution systems. RE-based power generation is proliferating in the power sectors worldwide. Participation of large numbers of market players, and massive penetration of RE-generation have created enough complexities and has made fundamental changes in the deregulated electricity market conditions. Small scale RE generating units have limited participation in the electricity markets due to the uncertainties. These units integrate with other fossil fuel plants and forms as Virtual Power Plants (VPPs). Increasing participation of RE based VPPs in the competitive electricity market, has brought out further complexity in market operation primarily in terms of its generation scheduling, economic profitability, etc. In this paper a two-stage stochastic programming approach for optimal scheduling of VPPs in the electricity markets is presented, along with modeling of uncertainties in the electricity market price, available level of stochastic renewable generation and the request for reverse deployment. These uncertainties are modeled using scenario bounds and are formulated using stochastic programming approach. Simulation results are carried out on 4-h planning horizon.

Keywords—*Electricity markets, Renewable Energy Sources (RES), Virtual Power Plants, Stochastic programming, Day Ahead Markets (DAM).*

I. INTRODUCTION

Power system utilities in the world are disintegrated and restructured, resulting in the diminishing of monopoly existed in the erstwhile vertically integrated markets. To promote investment in the power sector, deregulated market regime has created an enabling environment to accelerate development in generation, transmission and distribution systems. These result in large participation of market players, stakeholders, independent power producers, electricity traders, and pro-active roles of regulators. Massive penetration of renewable energy-based electricity generation is proliferating in every country around the world. It has made fundamental changes in the deregulated electricity market conditions. These changes affect the financial health of incumbent fossil fuel generators having inherently high marginal costs of generation per unit.

As per rough estimates, burning of widely used fuels like coal, bio-fuels pollute air over fifty times more carbon per unit of energy than wind, water, or solar power. Due to environmental conservation and increasing efforts to reduce global greenhouse gas emissions, efforts are being made for continuous policy reforms in the power sectors are on the

cards across the globe, and thrusts are given towards sustainable sources of RE generation, replacing the predominant nonrenewable electricity generations. Such transition is widely prevalent in U.S and European Union, which have set the milestone of electricity requirement with 100 percent renewable energy in near future [10-12].

Virtual power plants are integrated with several type of energy resources to aggregate total energy production from distributed energy resources (DERs) such as small hydro plants, roof top solar system, wind farms etc. VPP includes from small scale to medium scale renewable generating units, flexible loads, diesel generation sets etc. These utilities can form into a cluster of energy sources along with other fossil fuel power generating units. Increasing participation of VPPs which aim at an integrated approach of a cluster of small distributed RE based generating entities, and participating in the competitive electricity market as a single entity, has brought out further complexity in market operation primarily in terms of its generation scheduling, economic profitability, etc. VPP acts as an intermediary between distributed energy resources and the whole sale electricity markets and trade energy on behalf of DERs owners who themselves are unable to participate in that market. The concept of VPPs allows small scale RE generating units to get in to electricity markets.

The real time load demand is dynamic and so the power generation for balancing the load. Further, due to intermittency, and variability of RE sources and imperfect forecasting, there is randomness in RE generation (viz. wind farms, solar PV plants, etc), and it leads to the complexity of RE integration with the electricity grids. RE power being largely non-dispatchable, the generation scheduling of CPPs in combination with RE generators is a tough challenge being encountered by system operators in DAM and real time markets.

Understanding the uncertainties in the process of RE generation and are considered as stochastic process. To mitigate these uncertainties during the generation of RE sources, it is endeavored to integrate RE generating utilities with other generating units such as CPPs, storage units and flexible demands i.e., VPPs. In this paper, a two-stage stochastic programming is proposed to model these uncertainties present in the process of integrating these VPPs containing RE sources and finds an optimal solution for scheduling generating units in electricity market clearing process [4,7].

A brief overview of the present electricity market scenario is studied and market mechanisms coupled with

bidding-based buy and sell of electricity in DAM and real time markets are well explained in [1-16]. The market clearing process in the electricity markets under various uncertainties is designed in [1]. The zonal market model with renewable integration is presented in [2] [3], [4]. Bidding strategy of virtual power plants (VPPs) for participating in energy and reserve markets is investigated in [5]. Introduction to the mathematical stochastic programming applied to electrical engineering is presented in [6]. Risk assessment in electricity markets and reserve market under uncertainties is carried out in [7]. Network constrained robust unit commitment model is explained in [8]. L. Tianqi et.al [9] analysed optimal scheduling of VPPs considering cost of battery loss. The statistical scenario of RE is illustrated in [10-13]. Impacts on power markets due to RE generation adequacy are presented in [14]. The challenges being faced by the electricity markets with the intermittence RE sources is explained in [15,16]. Optimal bidding based on Nash equilibrium strategy for VPP participation in the energy markets is proposed in [17]. However, the proposed approach adds to the existing research on RE based VPPs electricity markets.

This paper is organized into five sections. Section-1 provides RE-based market operation and pro-active role of Virtual power plants as introduction. Section-2 presents the impacts of renewable generation in deregulated markets. Stochastic programming methodology, modelling uncertainties and problem formulation for optimal scheduling process in electricity markets is explained in Section-3. Simulation results are illustrated in section-4 and section-5 ends with the conclusion followed by references.

II. IMPACT OF RE GENERATING UNITS ON DEREGULATED ELECTRICITY MARKETS

RE generation has made fundamental changes in the market conditions. Historically, there are steep reductions in costs of RES generation over the time and per unit cost is economical and compared to the cost of fossil fuel-based generation. RE penetration into the markets has resulted in reduction of wholesale electricity prices. It is indicated in [13], the levelized cost of electricity (LCOE) between 2009-2017 for PVs fell from \$304 per MWh to just \$86, a reduction of 72%. Onshore wind's LCOE dropped from \$93 to \$67 per MWh, a reduction of 27%. These factors and the feature of insignificant greenhouse gas emissions are instrumental to make a paradigm shift to create market of renewable. However, it has created a large impact on the economic profitability of conventional generators and considerably affects the financial health of the stakeholders of fossil fuel generators. It has also transpired that massive penetration of RE is and will be leading to even negative electricity prices, i.e., conventional generators are required to pay to produce electricity [3].

Uncertainty of RE generation is another critical factor which affects the scheduling and operation of the grid [15]. VPPs enable to provide a possible solution to mitigate such issues with its group of generating units (both conventional and RE sources), storage units, biomass plants and flexible demands. These VPPs can optimize their energy sources utilizing conventional plants during its low RE production and participate in the markets during its high RE production through storage units. It also reveals the importance of conventional sources in supporting the system. These

traditional generators can act as capacity plants in the capacity markets.

In the context of economic profitability of stakeholders [10-13], optimal utilization of resources to meet end user requirements, and for mitigating imbalance of load-generation dynamics in RE dominated electricity market, stochastic and optimal scheduling of VPPs is made as a part of market research.

III. STOCHASTIC OPTIMAL SCHEDULING PROGRAM

This section analyses the optimal scheduling problem of the electricity markets where virtual power plants (VPP) sells or buys the energy with the objective of profit maximization. On other hand, reserve markets provide the flexibility to increase or decrease the total energy production of VPPs upon the request of the system operator.

The Day-Ahead and reserve electricity markets are considered in this section to analyse the market scheduling decisions one day in advance. While making this scheduling decision the VPPs faces a number of uncertainties [1,9]. Following are the uncertainties faced in the market scheduling process:

- The market prices include the day-ahead market prices and the reserve markets prices (for both capacity and energy).
- Stochastic nature of the available renewable generating unit's production level.
- The requests to deploy reserves sources by the system operator.

The proposed uncertainties are modeled for obtaining optimal market scheduling decisions. As the proposed approach is probabilistic and not deterministic in nature, inappropriate modeling will result in loss or profit to the VPPs and even results in an infeasible operation of utilities/generation and demand assets [18].

TABLE I. NOMENCLATURE

Notation	Definition
Ω^C	Set of Conventional Power Plants
Ω^D	Set of Demands
Ω^R	set of renewable energy generating units
Ω^S	Set of storage units
Ω^T	Scheduling Time periods
Ω^θ	Set of discrete scenarios
$C_c^{C,F}$	Online cost of conventional generating unit c [\$ /h]
$C_c^{C,V}$	Variable cost of conventional generating unit c [\$ /h]
P_{ct}^C	Power generation of the conventional power plants in time period t [MW]
P_{dt}^D	demand d power consumption level in the time period t [MW]
P_{rt}^R	RE generating unit r production level during the time period t
$P_{rt}^{R,A}$	Available RE generating limit of unit r in the time period t
$P_{st}^{S,D}$	Power discharging level of storage unit s in the time period t
$P_{st}^{S,C}$	Charging level of storage unit s in the time period t [MW]
e_{st}^S	Energy stored by the storage unit s in the time period t [MWh]
P_t^{R+}	Power capacity traded in up-reserve market in time period t
P_t^{R-}	Power capacity traded in down-reserve market in time period t
P_t^E	Amount of Power traded in the market during the time period t

A. Modelling Uncertainties

The uncertainties mentioned above are modelled using a set of predefined discrete scenario realizations indicated by $\vartheta \in \Omega^V$. Each scenario of ϑ is defined by the parameters $\mu_{\vartheta t}^E, \tilde{\mu}_{\vartheta t}^{R+}, \mu_{\vartheta t}^{R-}, \tilde{\mu}_{\vartheta t}^{R-}, K_{\vartheta t}^{R+}, K_{\vartheta t}^{R-}$, and $P_{rt\vartheta}^{R,A}$ that indicates energy market price, market price acquired for power capacity in the down-reverse market, the up and down-reverse deployment request, and the available generating levels of stochastic RE generating units respectively. Each scenario ϑ is defined with probability of occurrence π_{ϑ} . The sum of overall probabilities of the scenarios is equal to 1, i.e., $\sum_{\vartheta \in \Omega^V} \pi_{\vartheta} = 1$.

B. Problem Formulation

The optimal decision-making problem under these scenarios is modelled as a two-stage stochastic programming model and is interpreted as follows:

$$\max_{\varphi^V} \sum_{\vartheta \in \Omega^V} \pi_{\vartheta} \left\{ \sum_{t \in \Omega^T} \left((\mu_{\vartheta t}^E P_t^E \Delta t) + (\tilde{\mu}_{\vartheta t}^{R+} + K_{\vartheta t}^{R+} \mu_{\vartheta t}^{R+} \Delta t) P_t^{R+} + (\tilde{\mu}_{\vartheta t}^{R-} - K_{\vartheta t}^{R-} \mu_{\vartheta t}^{R-} \Delta t) P_t^{R-} - \sum_{c \in \Omega^C} (C_c^{C,F} u_{ct}^C + C_c^{C,V} P_{ct\vartheta}^C \Delta t) \right) \right\} \quad (1)$$

Subject to:

$$\underline{P}_t^E \leq P_t^E \leq \bar{P}_t^E \quad (2)$$

$$\underline{P}_t^{R+} \leq P_t^{R+} \leq \bar{P}_t^{R+} \quad (3)$$

$$\underline{P}_t^{R-} \leq P_t^{R-} \leq \bar{P}_t^{R-} \quad (4)$$

$$P_t^E + K_{\vartheta t}^{R+} P_t^{R+} - K_{\vartheta t}^{R-} P_t^{R-} = \sum_{c \in \Omega^C} P_{ct\vartheta}^C + \sum_{r \in \Omega^R} P_{rt\vartheta}^R + \sum_{s \in \Omega^S} (P_{st\vartheta}^{S,D} - P_{st\vartheta}^{S,C}) - \sum_{d \in \Omega^D} P_{dt\vartheta}^D \quad (5)$$

$$\underline{P}_{dt}^D \leq P_{dt\vartheta}^D \leq \bar{P}_{dt}^D \quad ; \forall d \in \Omega^D \quad (6)$$

$$\sum_{t \in \Omega^T} P_{dt\vartheta}^D \Delta t \geq \underline{E}_d^D \quad ; \forall d \in \Omega^D \quad (7)$$

$$\underline{P}_{ct}^C u_{ct}^C \leq P_{ct\vartheta}^C \leq \bar{P}_{ct}^C u_{ct}^C \quad ; \forall c \in \Omega^C \quad (8)$$

$$0 \leq P_{rt\vartheta}^R \leq P_{rt\vartheta}^{R,A} \quad ; \forall r \in \Omega^R \quad (9)$$

$$\underline{P}_{st}^{S,C} \leq P_{st\vartheta}^{S,C} \leq \bar{P}_{st}^{S,C} \quad ; \forall s \in \Omega^S \quad (10)$$

$$\underline{P}_{st}^{S,D} \leq P_{st\vartheta}^{S,D} \leq \bar{P}_{st}^{S,D} \quad ; \forall s \in \Omega^S \quad (11)$$

$$e_{st\vartheta}^S = e_{s(t-1)\vartheta}^S + P_{st\vartheta}^{S,C} \Delta t \eta_s^{S,C} - \frac{P_{st\vartheta}^{S,D} \Delta t}{\eta_s^{S,D}} \quad ; \forall s \in \Omega^S \quad (12)$$

$$\underline{E}_{st}^S \leq e_{st\vartheta}^S \leq \bar{E}_{st}^S \quad ; \forall s \in \Omega^S \quad (13)$$

Where set $\varphi^V = \{P_t^E, P_t^{R+}, P_t^{R-}, \forall t \in \Omega^T; u_{ct}^C, \forall c \in \Omega^C; P_{ct\vartheta}^C, \forall c \in \Omega^C; P_{rt\vartheta}^R, \forall r \in \Omega^R; P_{st\vartheta}^{S,C}, \forall s \in \Omega^S; P_{st\vartheta}^{S,D}, \forall s \in \Omega^S;$

$\Omega^S; e_{st\vartheta}^S, \forall s \in \Omega^S\}$ are the optimization variables in the above problem. π_{ϑ} indicates, weight of each scenario ϑ [7,8,9]. VPPs objective is described by the Eq. (1) throughout the planning horizon and consists of the following terms:

- The term $\mu_{\vartheta t}^E P_t^E \Delta t, \forall t \in \Omega^T$ represents the revenues acquired by the VPPs for their participation in the DA markets. Here the variable P_t^E may be +ve (if VPPs sell power in the DA market) and -ve (if the VPPs buy power in the DA markets).
- The term $(\tilde{\mu}_{\vartheta t}^{R+} + K_{\vartheta t}^{R+} \mu_{\vartheta t}^{R+} \Delta t) P_t^{R+}, \forall t \in \Omega^T$ represents the revenue obtained by the VPP for participating in the Up-reserve markets. These revenues are again divided into $(\tilde{\mu}_{\vartheta t}^{R+} P_t^{R+})$ capacity payments and $(K_{\vartheta t}^{R+} \mu_{\vartheta t}^{R+} \Delta t P_t^{R+})$ energy payments.
- The term $(\tilde{\mu}_{\vartheta t}^{R-} - K_{\vartheta t}^{R-} \mu_{\vartheta t}^{R-} \Delta t) P_t^{R-}, \forall t \in \Omega^T$ represents the revenue obtained by the VPP for participating in the Down-reserve markets. These revenues are again classified into $(\tilde{\mu}_{\vartheta t}^{R-} P_t^{R-})$ capacity payments and $(K_{\vartheta t}^{R-} \mu_{\vartheta t}^{R-} \Delta t P_t^{R-})$ energy payments.
- Variable cost incurred by the CPPs is represented by the term $(C_c^{C,F} u_{ct}^C + C_c^{C,V} P_{ct\vartheta}^C \Delta t); \forall c \in \Omega^C, \forall t \in \Omega^T$

Where Eqs. (2), (3) and (4) are the constraints, representing upper and lower bounds on the amount of power traded in the DA, up, and down-reserve markets respectively. Eq. (5) represents the power balancing constraint. Eqs. (6) and (7) puts power consumption limits on the demands. $u_{ct}^C \in \{0,1\}; \forall c \in \Omega^C, \forall t \in \Omega^T$ denotes binary variable, it represents the on/off status of CPP. Constraints in the Eqs. (8), and (9) limits the power produced by the CPPs and stochastic RE generation level respectively. Eqs. (9) and (10) represents the constraints on the charging and discharging level of storage units, while the Eq. (12) represents the energy production level in storage units and Eq. (13) represents the limiting constraint on the energy level of the storage units. The above problem is a Mixed Integer Linear Programming (MILP) problem solved using CLPEX solver.

IV. SIMULATION RESULTS

The proposed two-stage stochastic model for optimal scheduling is tested on 4-hour planning horizon, and the required data is collected from [18]. The simulation results are implemented in GAMS software using CLPEX solver on a PC with an Intel i7 3.6GHZ CPU and 8-GB RAM.

The maximum power traded (sold/buy) in the energy market is limited to 100MW. The up and down reverse market capacity is limited at 50 MW. Energy Market prices along with up and down reverse market prices for the power capacity are presented in Table-2. Generation limits of the CPPs and their economic data along with the flexible demand data is referred from IEEE-5 bus system. The forecasted wind power production level is provided in Table-3. Reverse deployment request is considered to be 80% of the power capacity scheduled in down reserve market during the time period-2, similarly for up reserve market 50% and 100% of scheduling capacity are requested during the time period-1 and 3 respectively. No reserve deployment is requested during the time period-4. This data

is assumed and considered based on the system operators request for the reverse deployments.

In the above two-stage stochastic programming model, the uncertainty in the RE (wind) generating levels along with the uncertainty present in the reserve deployment request are modeled by using two equiprobable scenarios in each stage. Thus 4 scenarios (two of each) are considered. For the sake of simplicity these scenarios are independent of each other.

TABLE II. ENERGY AND REVERSE MARKETS PRICE DATA

Time Period	Price [\$/MWh]				
	Energy Markets	Up reserve market		Down reserve market	
		Energy	Capacity	Energy	Capacity
1	12	14	4	14	4
2	14	15	10	38	10
3	22	30	8	26	8
4	32	20	6	25	6

TABLE III. TOATAL WIND FORECASTING LEVEL FOR DIFFERENT TIME PERIODS AND WEIGHTS[PU]

Time Period	Wind Power Generating level [MW]			Scenario Weights
	R_1	R_2	R_3	
1	70	100	120	0.25
2	100	83	140	0.25
3	95	75	115	0.25
4	55	80	100	0.25

The proposed model presented in the section-III is runed by the system operator to determine optimal scheduling for each generating unit in each time period. Considering the data presented in the tables-II and III, the optimal power scheduled and market prices in Day-ahead and reserve markets are presented and explained in Figs. 1 to 7.

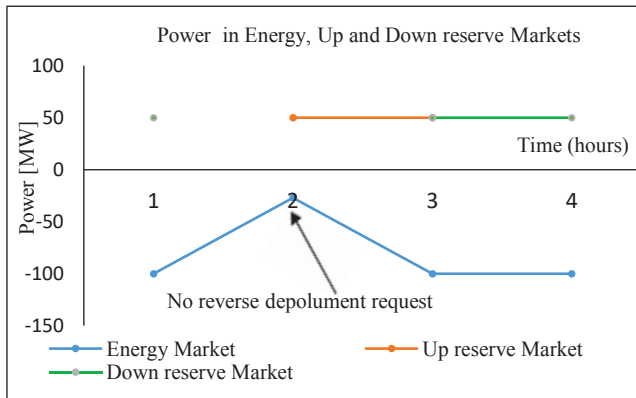


Fig. 1. Plants Power traded in energy, up-reserve and down reserve markets

VPPs participate in the energy markets and submit their bids based upon their demand levels in the specific time period, these bids are submitted in terms of price and quantity to the system operator. Fig. 1 shows, the optimal amount of power traded in each time period in the energy, up and down reserve markets. In these markets the VPPs decides to buy the energy, expect for the time period-2 when its demand level is being low and during this period energy is supplied by their own renewable generating units. In all other cases the VPPs tries to buy the energy in the energy

market. In the case of up reserve market, the power is traded in time periods 2 and 3. While in the case of down reserve market power is traded during the periods 1, 3 and 4. No power is traded during the time period-2. This is because of the maximum demand levels and low prices for the reserve deployment as explained in Table-II.

Based up on the amount of power traded in the energy and reserve markets, conventional power plants are scheduled. During the low demand level i.e., during the time period-2, these plants are turned off. Fig. 2 shows, scheduling of CPP. The power plants with highest economical prices are not scheduled for the entire market operation as shown in Fig 5. The renewable generation is maximum at each time period as shown in the Fig. 4. Power consumption levels in each time period is same expect for the time period T-2. In order to supply their demand level during the maximum demand periods, VPPs enter in to the power markets. Therefore, based on the market prices the VPP decides to buy maximum power from RE sources in the markets.

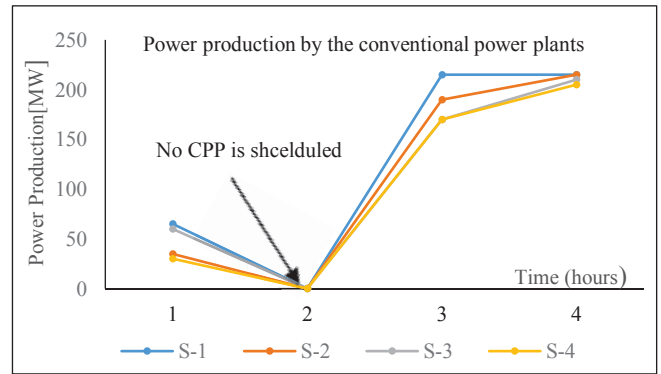


Fig. 2. Power consumption by the conventional power planys in the market

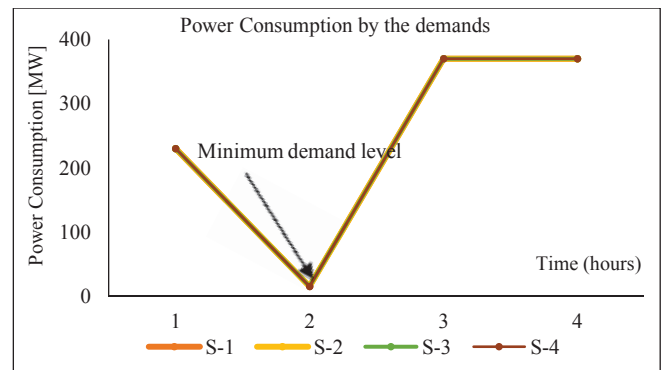


Fig. 3. Power consumption by the demand in the market

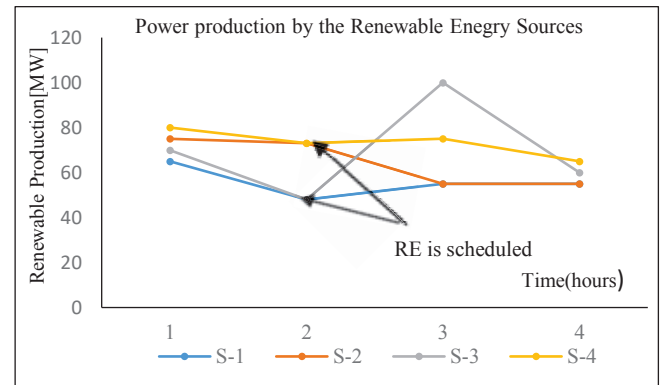


Fig. 4. Forecasting Power generating level of Wind power source

From the above optimal scheduling process, it is clear, that the stochastic RE sources are made to dispatch in all time periods and based on the power demand level and market prices the system operator request for up and down reverse deployments during a sepefic time period.

The conventional power plants are scheduled only when the required demand is more than the RE and reverse deployment capacity. Fig.5 shows scheduled dispatch of CPPs. During every hour generator (G-4) is not dispatched. Similarly, during second period no CPP is scheduled. This puts economic burden on the conventional generators. The prices incurred during the power production is less than the revenues obtained. Hence, it is required to provide policy incentives and to take standard tariff policy mechanism for conventional generation. Updating to the current technology, increasing ramp up and ramp down rates of the generators may make their way possible to compete with the RE sources.

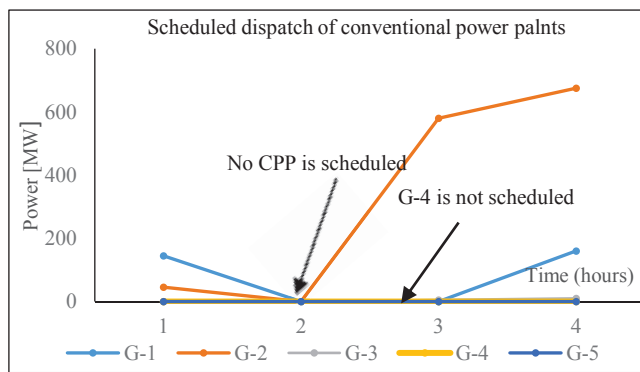


Fig. 5. Scheduled power dispatch of conventional power plants

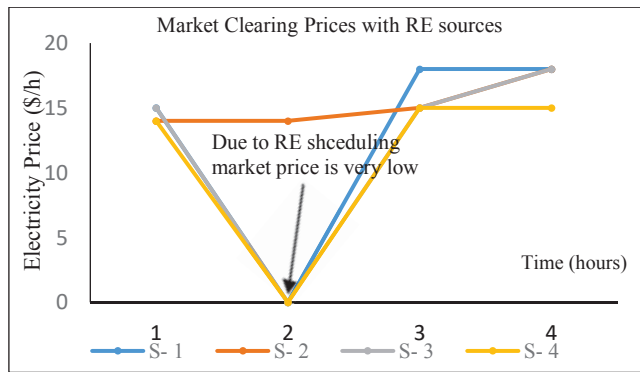


Fig. 6. Market clearing prices in \$/h with RE sources for different time periods.

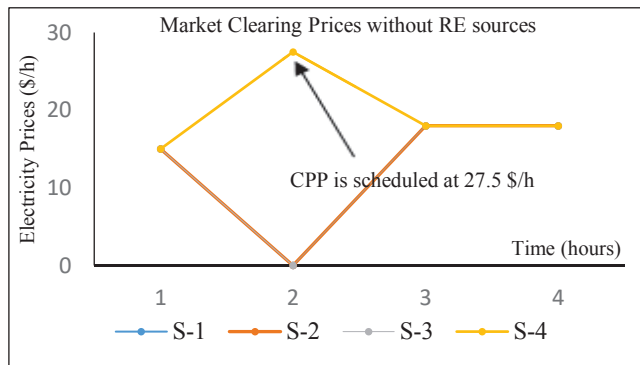


Fig. 7. Market clearing prices in \$/h without RE sources for different time periods.

Fig.6 and Fig.7 represents market clearing price (\$/h) variations with and without RE sources. It is observed that MCP's without RE sources is always greater than the MCP'S with RE. Therefore, it is clear that the conventional generators are forced to generate power for lesser prices. This price variations will result in economic losses. Therefore, it is required to provide cost-based policy incentives for the conventional plants. Some of the countries are following fed in tariff policy, purchase obligations and contract for difference mechanism to create a balance pricing mechanism.

The above simulation result has established that during low demand periods VPPs optimizes its resources by using RE (wind) generation only. During maximum demand periods CPPs are scheduled, and reserve deployment requests are made accordingly to the system operator request. With interest participation of VPPs market price levied on the consumers is reduced but burden on the conventional generators increases. The simulation graphs shown in Fig.6 and 7 has clearly indicated the price variations with and without RE sources. This clearly indicate that the VPPs in the electricity market are acting as price makers and sometimes as price takers.

The above two stage stochastic problem is executed using deterministic approach, in such case the optimal scheduling of VPP is found infeasible. This is due to the error while providing reverse deployment request. This highlights the importance of an accurate modelling of the uncertainties in the problem. Economic impact on CPPs due to RE sources can also be interpreted from the above results.

V. CONCLUSION

Power and energy balancing mechanisms are evolutionary in market operation from the cost economics angle. It depends on RE policies, power sector reform strategies, price discovery mechanisms, generation scheduling economics, load management techniques, role of VPPs, etc. Like every generator looking for its profitability and services to the system operation. VPPs with its resources look for maximization of their profits. The proposed two-stage stochastic modelling for optimal scheduling of VPPs in electricity markets has established the merits of its generation scheduling to mitigate certain uncertainties as explained in Para-A, section-III of this paper. The simulation work provides impressive results for accurate optimal scheduling of the generating units of VPPs. This method may be extended to large scale market operations and big power system networks by dividing the system into several subsystems. In future, the presented method may be extended to model forecasting uncertainties along with modelling of large solar generating units.

REFERENCES

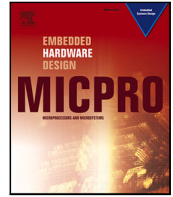
- [1] Conejo, A.J., Carrión, M., Morales, J.M.: Decision Making Under Uncertainty in Electricity Markets. Springer, New York, US (2010).
- [2] I. Aravena and A. Papavasiliou, "Renewable Energy Integration in Zonal Markets," *IEEE Trans on Power Systems*, vol. 32, no. 2, pp. 1334-1349, 2017.
- [3] R.Bayindir, S.Demirbas, E.Irmak, et al., "Effects of renewable energy sources on the power system", 2016 IEEE conference (PEMC), pp.1-6, Sep-2016.
- [4] B. Jie,et.al, "An analysis of market mechanism and bidding strategy for power balancing market mixed by conventional and RE,"

- International Conf on the European Energy Market*, EEM, 2017, pp.1–6.
- [5] Mashhour, E., Moghaddas-Tafreshi, S.M.: Bidding strategy of virtual power plant for participating in energy and spinning reserve markets–Part-I: *Problem form IEEE Trans. Power Syst.* 26(2), 949–956 (2011).
- [6] Birge, J.R., Louveaux, F., “Introduction to Stochastic Programming”, 2nd edn. *Springer*, New York, US (2011)
- [7] S. R. Dabbagh and M. K. Sheikh-El-Eslami, "Risk Assessment of Virtual Power Plants Offering in Energy and Reserve Markets," in *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3572–3582, Sept. 2016, doi: 10.1109/TPWRS.2015.2493182.
- [8] Jiang, T., Zhang, M., Li, G., Guan, Y. “Two stage network constrained robust unit commitment problem”, *Eur J Oper Res* 234(3), 751–7–7
- [9] L. Tianqi et al., "Analysis of Optimal Scheduling Model for Virtual Power Plants Considering the Cost of Battery Loss," 2019 *2nd International Conference on Information Systems and Computer Aided Education (ICISCAE)*, Dalian, China, 2019, pp. 218–222, doi: 10.1109/ICISCAE48440.2019.221621.
- [10] D. Spencer, “B.P. Statistical Review of World Energy Statistical Review of World,” *The Editor B.P. Statistical Review of World Energy*, pp. 1–69, 2019.
- [11] IEA, “Market Report Series: Renewables 2019 Analysis and Forecasts to 2024,” 2019.
- [12] World Bank, “The World Bank: state and trends of the carbon market,” 2010.
- [13] Renewable Energy Agency International IRENA, “Renewable Energy Market Analysis: Latin America,” 2019.
- [14] Stefan Jaehnert and Gerard Doorman, “Analyzing the generation adequacy in power markets with renewable energy sources”, 11th *International Conference on the European Energy Market*, pp.1–6, 2014.
- [15] P. L. Joskow, “Challenges for Wholesale Generation at Scale: Intermittent Renewable Electricity Markets with The U.S. Experience,” 2009
- [16] Morales J.M, Conejo A.J, Madsen. H, Pinson.P, Zungo.M, “Integrating Renewables in Electricity Markets-Operational Issues”, *springer*, New york, USA, 2014
- [17] H. Nezamabadi, P. Nezamabadi, M. Setayeshnazar and G. B. Gharehpetian, "Participation of virtual power plants in energy market with optimal bidding based on Nash-SFE Equilibrium Strategy and considering interruptible load," *The 3rd Conference on Thermal Power Plants*, Tehran, 2011, pp. 1–6.
- [18] Luis Baringo and, Morteza Rahimiyan, “Virtual Power Plants and Electricity Markets”, e-Book, *Springer Nature Switzerland AG* 2020, <https://doi.org/10.1007/978-3-030-47602-1>



Contents lists available at ScienceDirect

Microprocessors and Microsystems

journal homepage: www.elsevier.com/locate/micpro

VLSI implementation of transcendental function hyperbolic tangent for deep neural network accelerators

Gunjan Rajput^a, Gopal Raut^a, Mahesh Chandra^b, Santosh Kumar Vishvakarma^{a,*}^a Department of Electrical Engineering, Indian Institute of Technology Indore, India^b NXP semiconductors, India

ARTICLE INFO

Keywords:

Activation function
Artificial neural network
Hyperbolic tangent (tanh)
Digital implementation
Combinational logic

ABSTRACT

Extensive use of neural network applications prompted researchers to customize a design to speed up their computation based on ASIC implementation. The choice of activation function (AF) in a neural network is an essential requirement. Accurate design architecture of an AF in a digital network faces various challenges as these AF require more hardware resources because of its non-linear nature. This paper proposed an efficient approximation scheme for hyperbolic tangent (tanh) function which purely based on combinational design architecture. The approximation is based on mathematical analysis by considering maximum allowable error in a neural network. The results prove that the proposed combinational design of an AF is efficient in terms of area, power and delay with negligible accuracy loss on MNIST and CIFAR-10 benchmark datasets. Post synthesis results show that the proposed design area is reduced by 66% and delay is reduced by nearly 16% compared to state-of-the-art.

1. Introduction

Artificial Neural Networks (ANN) are pertinent in different applications such as image processing, speech recognition, and language processing [1]. ANN has implemented using the software predominantly. The software method has an advantage, as there is no need for designers to know the inner model of ANN elements, but can easily take care of the application part. Presently the work is going on in ANN application by using CPUs and GPUs only, which are ill-suited for the applications where low power and optimum latency are obligatory. To accelerate neural network applications and reduce their power consumption with less latency is a prior requirement.

The main building blocks required for designing a neural network are adder, multiplier, and activation function (AF). Implementation of MAC unit is easy as it requires multiplier and adder tree. Whereas implantation of AF complicated due to its non-linear features. Moreover, the implementation of the Tanh function needs both positive and negative exponential function [2]. An activation functions are non-linear such as sigmoid, Elliot, Tanh, ReLU, soft-max and many more [3,4]. It consists of a division and positive/negative exponential calculations [5].

The tanh and sigmoid activation functions are more efficient for better training due to their non-linear behavior compared to earlier AFs such as step, linear, etc. Moreover, various other AFs are proposed, such as ReLU, ELU, SWISH, Parametric ReLU, etc. Here, we have focused

on the implementation of the Tanh function. In contrast, the method can elaborate for all activation functions. The sigmoid output ranges from 0 to 1, and the hyperbolic tangent range from -1 to 1 , but both form s-curve shape as hyperbolic tangent and sigmoid function given in [6]. Tanh is much better for learning than the sigmoid function [7]. Tanh and sigmoid function include an exponential and division term which is very challenging to realize using digital design architecture. An approximation method is generally taken into consideration to eradicate these problems.

Various approximation techniques have been used for the implementation of the activation function based on a Lookup table (LUTs), function's series expansion, Coordinate Rotation Digital Computer (CORDIC) algorithm, Stochastic computing, Piece-wise linear function (PWL) [8–11]. However, directly storing a functional value of a non-linear AF in LUT's is costly since it requires more parameters. Most of the accelerators have not been implemented by Instruction set architecture, but they can create modules separately, preventing designers from reducing hardware costs. Thus, always thinking about the multiplication and adder block, special attention should be given to other components such as the AF block. There are hidden layers of neurons in a neural network, and each neuron has its AF. Therefore, highly efficient AF in terms of power, area, and delay with adequate accuracy are required.

* Corresponding author.

E-mail address: skvishvakarma@iiti.ac.in (S.K. Vishvakarma).<https://doi.org/10.1016/j.micpro.2021.104270>

Received 10 April 2020; Received in revised form 12 February 2021; Accepted 16 April 2021

Available online 11 May 2021

0141-9331/© 2021 Elsevier B.V. All rights reserved.

Table 1
Computational equations for configurable AF design exploration.

Preliminary Work	Activation function implementation	
	Sigmoid	Tanh
[15,16]	$1/(1 + e^{-x})$	$1 - 2 \text{ Sigmoid}(-2x)$
[17]	$1/(1 + e^{-x})$	$(e^{2x} - 1)/(e^{2x} + 1)$
[16]	$[1 + \tanh(x/2)]/2$	$(e^x - e^{-x})/(e^x + e^{-x})$

1.1. Motivation

The non-linear AF such as Sigmoid and Tanh provides a smooth transition between excitation and Inhibition that leads to better neuron response [12]. The mathematical equation that is used for implementation in state-of-the-art is summarized in Table 1. However, they employed an approach to computing AF that lead to low throughput. It requires both positive and negative exponential functions for the final desired output for AF hardware implementation. All those methods and design technique is hardware costly for the hardware implementation. The non-linear transformation tanh function calculation requires both positive and negative exponential function, which is expensive for the LUT-based approach. Moreover, those function requires multiplier for computing the e^{2x} and divider for further extension for the equations that increase the area overhead. Further, those approaches are not feasible for higher precision implementation [13,14]. To overcome these limitations, we have designed a tanh function using combinational logic with the help of OR and AND plane. By using combinational logic, that can provide low latency with less area.

1.2. Contribution

This article explores the design-space trade-offs of neural networks with a digital design of AF implementation. The work has focused on the tanh AF at the 180 nm technology node. The key contributions are

- We have implemented tanh non-linear transformation functions with the help of truncation of Taylor's series and combinational logic circuits.
- Performance and inference accuracy validation are done using benchmark LeNet deep network with MNIST and CIFAR-10 dataset.
- We analyze and discuss the circuit's physical parameters like area, power, and throughput and evaluate it with the 180 nm technology node. Further, it compare with the state-of-the-art designs.

1.3. Organization

The rest of the paper is organized as follows: Background and Related Work discussed in Section 2. In Section 3 explains the digital design and mathematical analysis of a proposed method for AF. Section 4 presents Results and discussion with experimental analysis of a proposed function. Section 5 give the experimental validation of the work design followed by a conclusion in Section 6.

2. Background and related work

The main challenge of DNN is resource utilization and power consumption for resource-limited devices. Whereas TOT-Net architecture has achieved a higher level of accuracy and less computational load [18]. In this novel, using TOT Net reduced the cost of multiply operators. Different types of activation function and their learning performance and optimization approach have investigated in [19]. Various techniques have been used for the implementation of an activation function. It mainly categorizes into two parts. First, a LUT-based approach and second piece-wise approximation method [20,

21]. The error contribution due to activation function with different precision is expressed in [4]. Moreover, the nonlinear activation function like sigmoid cannot be approximated efficiently using only combinational logic. However, using purely combinational logic has the benefits of providing low latency with small area overhead compared to conventional ROM-based approaches.

This paper has explained the most commonly used activation function (Tanh) concisely using a LUT-based approximation approach. Precisely, implementing an activation function is a bottle-neck between area, power, and accuracy. A minimal approximation in design architecture requires fewer hardware resources and has less latency than the approximation in design. In Fig. 1 we have plotted the exponential curve (e^x) using various methods, which are explained in the following subsection.

2.1. Storing activation function value in LUT [20]

In this method, the function value is divided into several ranges by approximating that range value and store the functional value directly in LUT. This method can be convenient with a highly precise approximate function. But by increasing precision, hardware requirement and complexity in design will also get increases. So, it is a barrier between the high precision and area, power, delay, etc. This method is LUT based approach in which value is directly stored in the LUT.

2.2. Storing parameter value of activation function in LUT [22,23]

In this method, parameters of the function are stored in the LUT. For example, a and b are the intercept on the respective x and y-axis in a straight line equation. Since these parameters are constant so by varying coordinates (x and y) in the Eq. (1), it can easily get a line.

$$\frac{x}{a} + \frac{y}{b} = 1 \quad (1)$$

We can save the parameters of the function in LUT. Here, a and b can be added in LUT to implement an AF in this method. This method is convenient with the above method. But the drawback of this method is that there is a use of adder and multiplier units to calculate constant function value. If the function is complex, then more parameters have to be added and stored in LUT, especially in higher precision representation.

2.3. Series expansion [24]

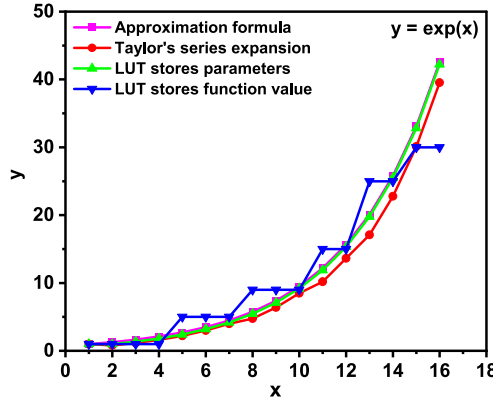
Series expansion through Taylor's series, McLaurin series, Bernstein polynomial and many more are used to implement non-linear AFs. The most popular Taylor series expansion representation shown in Eq. (2) for $f(x)$ is

$$f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \dots = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!}x^k \quad (2)$$

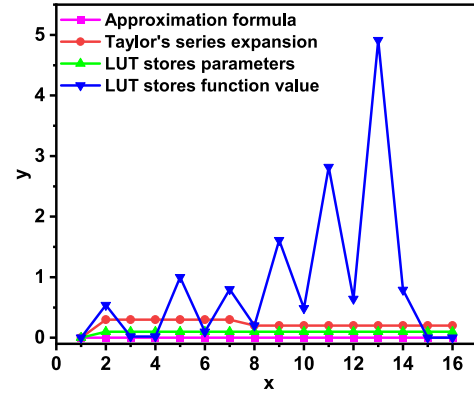
The mathematical modeling of this equation requires multipliers and adders, whereas multipliers are power-hungry blocks. However, if higher precision is not the primary requirement, higher-order values can truncate for non-linear transformation function implementation. This method comes under the category of series expansion method.

2.4. Piece-wise linear, non-linear function transformation [25–28]

In the piece-wise linear method, the non-linear input range is divided into regions, and respective values are stored in the LUT. In the case of sigmoid and Tanh function, there is a linear region also. The value of that linear region can be directly stored in LUT. The remaining non-linear part can be approximated, and the value can be stored in the LUT as shown in . This method is much efficient but has less precision, and it also requires more area and latency. In this method, the pre-calculated ROM value is stored in LUT.



(a) Approximation of exponential function $y=\exp(x)$ using different design approach



(b) Error of exponential curve using various methods for $y=\exp(x)$

Fig. 1. Performance and accuracy comparison of exponential function calculation using different logic design approach.

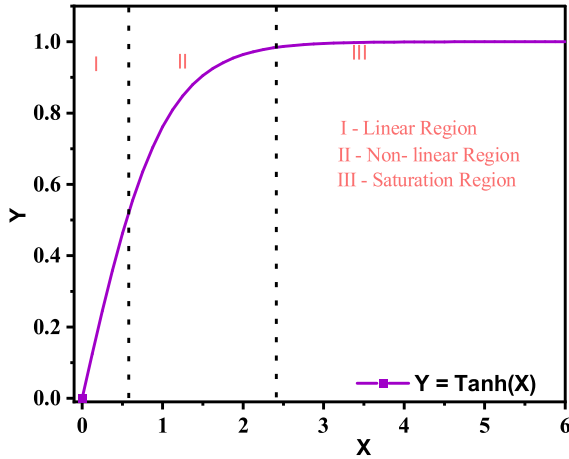


Fig. 2. Hyperbolic tangent transformation curve and different regions of curve based on its slope.

2.5. Coordinate rotational digital computer [10,13,16]

A coordinated rotational digital computer (CORDIC) is a simple and highly effective power and resource utilization method. The method is based on the elementary operation of trigonometric equations. It uses shift, addition, subtraction operations for the computation of the non-linear AF. Although the CORDIC algorithm efficient for the area, it requires more clock cycles. This algorithm has efficient with high accuracy but low latency. The general equation used for the realization of the AF is shown below

$$\alpha_{i+1} = \alpha_i + m\psi_i\beta_i\chi^{S_{m,i}} \quad (3a)$$

$$\beta_{i+1} = \beta_i + \psi_i\alpha_i\chi^{S_{m,i}} \quad (3b)$$

$$\gamma_{i+1} = \gamma_i + \psi_i\delta_{m,i} \quad (3c)$$

where ψ shows the direction of a micro rotation, where direction can be clockwise or anticlockwise. m represents the type of a coordinate system if $m = 1$, then the system is circular; if $m = 0$, then the system is linear, and for $m = -1$ system is hyperbolic. $S_{m,i}$ is an integer that is non-decreasing. $\delta_{m,i}$ is the rotation angle. In this method, while doing operations like addition, shift, etc., the output value is stored in LUT.

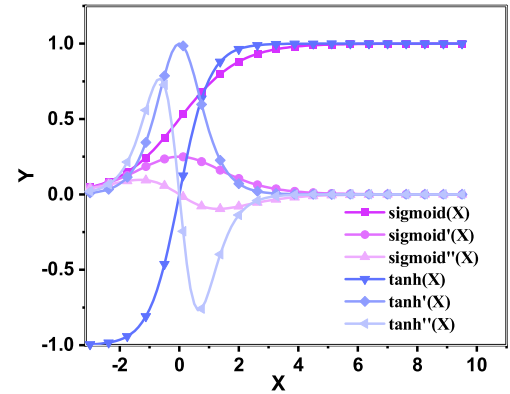


Fig. 3. Tanh and sigmoid curve with their derivatives.

2.6. Approximating activation function [29,30]

To approximating AF, the mathematical function will be approximated such as

$$\exp(x) \approx E_x(x) = 2^{1.44x} \quad (4a)$$

$$\text{sigmoid}(x) \approx \frac{1}{1 + 2^{-1.5x}} \quad (4b)$$

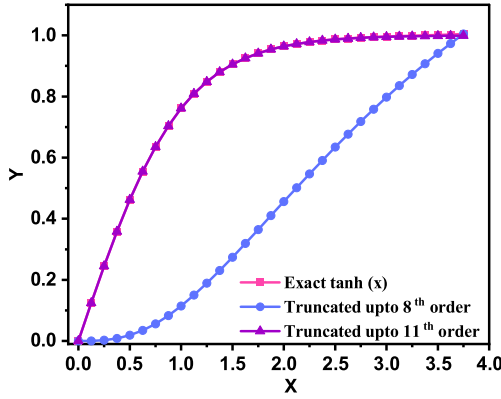
Similarly, the tanh function can also be implemented in the same manner as sigmoid. This method requires fewer cycles as a comparison to that of the CORDIC, but still, latency is high. This method requires four cycles to implement this Sigmoid function. There are other methods, such as the range addressable look-up (RALUT) method, Hybrid methods consisting of LUT and series. This paper has implemented an AF based on the combinational circuit by truncated series expansion.

2.7. Vanishing gradient problem

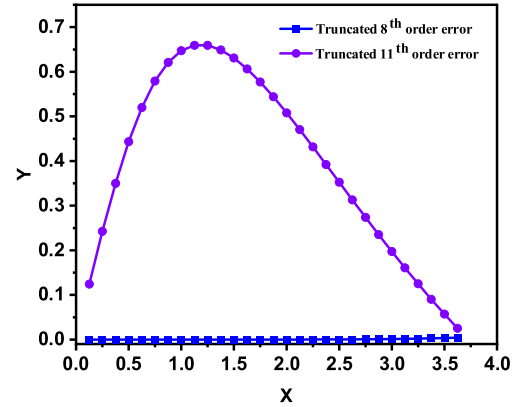
As sigmoid and tanh functions are having non-linear behavior, the digital implementation is complicated at all exploration points. The problem is a significant impact on weight updates. The weight updates rule is therefore explained in the below equation.

$$\text{weight}^{(L)} = \text{weight}^{(L)} - lr \times \frac{\delta(C)}{\delta \text{weight}^{(L)}} \quad (5)$$

Derivative term in the weight updating equation is responsible for the vanishing gradient problem. In Eq. (5), lr is the learning rate, C is constant. In vanishing gradient problem derivative term shows that



(a) Tanh curve and its truncated curve



(b) Truncated error

Fig. 4. Tanh function analysis and truncated error for different order of series expansion.

the number of weights and biased values updates with a very less amount. In back-propagation, at every time, small weight gradients will be updated. The sigmoid and its derivative is shown below represented as:

$$\sigma(x) = \frac{1}{1 + e^{(-x)}} \quad \& \quad \sigma'(x) = \frac{e^{(-x)}}{1 + e^{(-x)^2}} \quad (6)$$

Here, a large x value at the input of the sigmoid function resulting in 0, almost, i.e., when the input value $w \times a + b$ then the output is almost equal to 0 further, there is no updation of the weights. In sigmoid max value reaches of derivative reaches to 0.25 as shown in Fig. 3. Therefore, in every layer, gradients are vanishingly small, and after this, there is no updation of the weights. Therefore it results in the network being very far from the optimal value. Moreover, at another hand, the Tanh function derivative ranges up to 1 as shown in Fig. 3. In tanh, while learning, w and b are larger than that of the sigmoid, where w is the weight and b is the bias values. The mathematical proof is shown below in Eq. (7):

$$\max \sigma'(x) < \max \text{Tanh}'(x) \quad (7a)$$

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)) \leq 0.25 \quad (7b)$$

$$0 < \sigma(x) < 1 \quad \text{maximize concave quadratic} \quad (7c)$$

$$\text{Tanh}'(x) = \text{sech}^2(x) = \frac{2}{\exp(x) + \exp(-x)} \quad (7d)$$

Above equations prove that the probability of vanishing gradient is more in sigmoid as compared with that of the Tanh.

3. Mathematical analysis and proposed digital design approach for activation function

The design of neural network accelerators requires trigonometric function calculations such as tanh and sigmoid. The proposed digital design architecture enables such computation using minimum resource utilization with maximum accuracy. The tanh function is divided into three regions as shown in . Region I and III value can be directly stored in the lookup table, and the remaining II region value can be approximated according to the precision required.

3.1. Mathematical modeling and analysis for an activation function

The trigonometric hyperbolic function is shown in Eq. (8). Whereas, the trigonometric hyperbolic function expansion in terms of exponential function is shown in Eq. (8c). Similarly, sigmoid function representation shown in Eq. (8b).

$$\sinh(x) = (e^x - e^{-x})/2 \quad (8a)$$

$$\cosh(x) = (e^x + e^{-x})/2 \quad (8b)$$

$$f_1(x) = \text{Sigmoid}(x) = 1/(1 + e^{-x}) \quad (8c)$$

$$f_2(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = 2f_1(2x) - 1 \quad (8d)$$

We compared sigmoid and tanh functions performance; both the functions comparatively show the same characteristics. Both functions resemble the same, and both can be implemented $y = x$, but tanh converges faster than sigmoid function having the same quantized values. The implementation of a sigmoid function in neural networks requires bias value, which can affect the optimization values. Tanh is used when the value of the neuron is restricted between $[-1, 1]$, then the AF output is more likely to come between $[-1, 1]$ as shown in Fig. 3. In contrast, it is different in the case of the sigmoid function. As shown in the paper by LeCun et al. clearly, a strong gradient is required, and one should avoid bias in the gradients [31]. Hence, tanh has benefits over sigmoid function.

Hence, We explore the design technique for tanh and have implemented by using combinational logic. We have used series expansion for the implementation of an AF and truncated the series up to 11th order to get optimum accuracy. It is observed that the accuracy is higher at 11th order than compared with 8th shown in Fig. 4(a) and Fig. 4(b). Tanh function is a transformation of an exponential function. Taylor's series expansion of $\exp(x) = p(x)$ is shown in Eq. (9a). Whereas, for negative exponential expansion $\exp(-x) = q(x)$ is shown in Eq. (9b).

$$p(x) = \sum_{k=0}^{\infty} \frac{p^{(k)}(0)}{k!} x^k = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots \quad (9a)$$

$$q(x) = \sum_{k=0}^{\infty} \frac{q^{(k)}(0)}{k!} x^k = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 - x - \frac{x^2}{2!} - \dots \quad (9b)$$

The behavior of this AF has 3 basic properties such as:

$$\lim_{y \rightarrow \infty} \tanh(y) = 1 \quad (10a)$$

$$\lim_{y \rightarrow 0} \tanh(y) = y \quad (10b)$$

$$\lim_{y \rightarrow -\infty} \tanh(y) = -1 \quad (10c)$$

After substituting the values and making approximations in Eq. (9) up to 11th orders, we choose those points where our function values changes. The Tanh function is an odd function that is symmetric concerning 0. For implementing this non-linear AF, we will select a positive half of the function. Using three fundamental properties of this tanh function, we can optimize our tanh equation based on these properties. In this, we perform a quantization process. Tanh curve is divided into three segments.

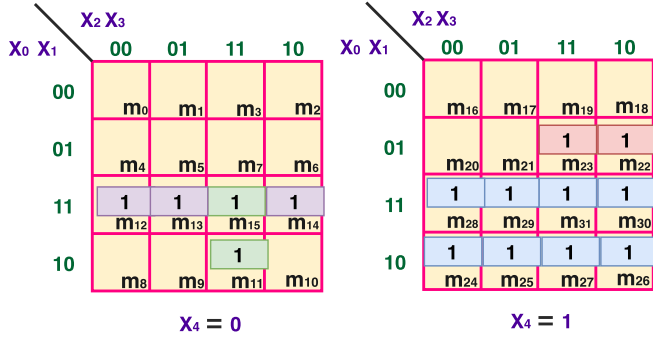


Fig. 5. k-map realization for proposed tanh function with a 5-bit input and 7-bit output. In this figure d_3 output has been shown.

Moreover, if the segments are more, we get more precision. So, it is a bottleneck situation between efficiency and hardware requirement. Suppose there are 2^N quantization levels with N number of input bits, i.e., Number of Fragmentation (NoF). N will depend upon the precision required. The NoF and size of the one frame, i.e., Q , is shown in below Eq. (11b).

$$NoF = 2^N \quad (11a)$$

$$Q = \frac{Fullscale}{2^N} \quad (11b)$$

3.2. Implementation of a tanh AF

The Tanh AF implementation techniques proposed in the state of the art as shown in Table 4. By selecting an appropriate sampling rate, we find out the corresponding value of each sample using Eqs. (9a) and (9b) in tanh function up to 11th order. Then convert this decimal value sample into a binary value with certain fix bit-width. In this implementation, we choose 7-bit for the encoding of the AF using 5-bits. Then generated the k-map of the truncated tanh equation as shown in Fig. 5. By analyzing a k-map one can easily get the min-terms or max-terms for the implementation of a function. For example, the output d_3 is shown in Eq. (12d). Similarly, we have written the remaining outputs from the Tables 2 and 3.

The detail Implementation of a tanh function is shown in algorithm 1. After getting all these outputs, we verify the RTL code of this approximated tanh function by not considering the pair condition. It may increase variables that result in sharing variables simultaneously, which causes race conditions and is responsible for creating hazards. Implementation is done by using 180 nm, comparing the results with that of the exact tanh function, and LUT-based approximation in terms of area, power, delay, and accuracy. We have to take respective algorithms and implement our technology node to verify our proposed AF.

$$d_0 = p_0 \quad (12a)$$

$$d_1 = p_1 + p_2 + p_3 + p_4 \quad (12b)$$

$$d_2 = p_5 + p_{41} + p_7 + p_8 + p_9 + p_{10} \quad (12c)$$

$$d_3 = p_5 + p_9 + p_{11} + p_{12} + p_{13} + p_{14} + p_{15} + p_{16} + p_{17} \quad (12d)$$

$$d_4 = p_{18} + p_{19} + p_{20} + p_{21} + p_{22} + p_{23} + p_{24} + p_{25} + p_{26} \quad (12e)$$

$$d_5 = p_{27} + p_{28} + p_{29} + p_{30} + p_{31} + p_{32} \quad (12f)$$

$$d_6 = p_{33} + p_{34} + p_{35} + p_{36} + p_{37} + p_{38} + p_{39} + p_{40} \quad (12g)$$

The combinational logic range of the above approximated simulated function is shown in Eq. (13).

$$f(x) = \tanh(x) = \begin{matrix} x \leq -3.5 = -1 \\ -3.5 \leq x \leq 3.5 = f(x) \\ 3.5 \leq x = 1 \end{matrix} \quad (13)$$

Table 2
Tanh implementation AND plane representation.

Input(AND plane)		$x_4 x_3 x_2 x_1 x_0$	
P ₀	x_4	P ₂₃	x_0, x_1, x_3, x_4
P ₁	x_0, x_1, \bar{x}_4	P ₂₄	$\bar{x}_0, x_1, \bar{x}_3, x_4$
P ₂	x_0, x_2, x_3, \bar{x}_4	P ₂₅	$\bar{x}_1, \bar{x}_2, \bar{x}_3, x_4$
P ₃	x_0, x_4	P ₂₆	$x_0, \bar{x}_1, x_2, \bar{x}_3, x_4$
P ₄	\bar{x}_0, x_2, x_4	P ₂₇	x_2, x_3, \bar{x}_4
P ₅	x_0, x_1, x_2, \bar{x}_4	P ₂₈	x_1, x_3, \bar{x}_4
P ₆	$\bar{x}_3, x_2, \bar{x}_2, \bar{x}_4$	P ₂₉	$\bar{x}_0, x_1, \bar{x}_3$
P ₇	$\bar{x}_1, x_0, \bar{x}_4, \bar{x}_2$	P ₃₀	x_1, x_3, x_4
P ₈	$\bar{x}_0, x_1, \bar{x}_2, x_4$	P ₃₁	x_0, x_1, \bar{x}_3, x_4
P ₉	x_0, x_1, x_4	P ₃₂	$x_0, \bar{x}_1, \bar{x}_2, x_4$
P ₁₀	\bar{x}_1, x_2, x_3, x_4	P ₃₃	$x_0, \bar{x}_1, \bar{x}_2, \bar{x}_3, x_4$
P ₁₁	x_1, x_2, x_3, \bar{x}_4	P ₃₄	$\bar{x}_0, x_1, \bar{x}_2, x_3, \bar{x}_4$
P ₁₂	$x_0, x_1, \bar{x}_2, \bar{x}_4$	P ₃₅	$x_0, \bar{x}_1, x_2, \bar{x}_4$
P ₁₃	$x_0, \bar{x}_2, x_3, \bar{x}_4$	P ₃₆	$\bar{x}_0, \bar{x}_1, x_2, \bar{x}_3, \bar{x}_4$
P ₁₄	x_0, \bar{x}_3	P ₃₇	x_0, \bar{x}_1, x_3
P ₁₅	x_1, \bar{x}_2, x_3, x_4	P ₃₈	$\bar{x}_0, \bar{x}_2, x_3, x_4$
P ₁₆	x_0, \bar{x}_1, x_3, x_4	P ₃₉	$\bar{x}_0, \bar{x}_1, x_2, x_4$
P ₁₇	$\bar{x}_0, \bar{x}_1, x_2, \bar{x}_3$	P ₄₀	$\bar{x}_0, x_2, \bar{x}_3, x_4$
P ₁₈	$\bar{x}_0, x_1, \bar{x}_2, \bar{x}_4$	P ₄₁	$x_2, \bar{x}_3, \bar{x}_4$
P ₁₉	$\bar{x}_0, x_1, x_3, \bar{x}_4$		
P ₂₀	$x_1, \bar{x}_2, x_3, \bar{x}_4$		
P ₂₁	$x_0, x_2, \bar{x}_3, \bar{x}_4$		
P ₂₂	x_0, x_1, \bar{x}_2, x_4		

Table 3
Tanh Implementation OR plane Implementation.

Output(OR plane)	$d_6 d_5 d_4 d_3 d_2 d_1 d_0$
d_0	P ₀
d_1	P ₁ , P ₂ , P ₃ , P ₄
d_2	P ₅ , P ₄₁ , P ₇ , P ₈ , P ₉ , P ₁₀
d_3	P ₅ , P ₁₁ , P ₁₂ , P ₁₃ , P ₁₄ , P ₉ , P ₁₅ , P ₁₆ , P ₁₇
d_4	P ₁₈ , P ₁₉ , P ₂₀ , P ₂₁ , P ₂₂ , P ₂₃ , P ₂₄ , P ₂₅ , P ₂₆
d_5	P ₂₇ , P ₂₈ , P ₂₉ , P ₃₀ , P ₃₁ , P ₃₂
d_6	P ₃₃ , P ₃₄ , P ₃₅ , P ₃₆ , P ₃₇ , P ₃₈ , P ₃₉ , P ₄₀

Algorithm 1 : Implementation of a Tanh Activation Function

- 1: With the help of the series expansion, expand tanh
- 2: Depend on the precision required truncate the series. In this paper we implemented up to 11th order.
- 3: Now decide the sampling rate, by $\frac{1}{2^P}$ each sample size and $\forall P \in \mathbb{N}$
- 4: Required according to the accuracy P should be selected
- 5: Find out the corresponding function value of each sample.
- 6: Encode the decimal value into bits. Select a appropriate bit-width for this implementation.
- 7: Now this boolean function can be expressed into canonical form.
- 8: Realization in SOP and POS form and design a circuit with the help of OR and AND plane.

4. Results and discussion

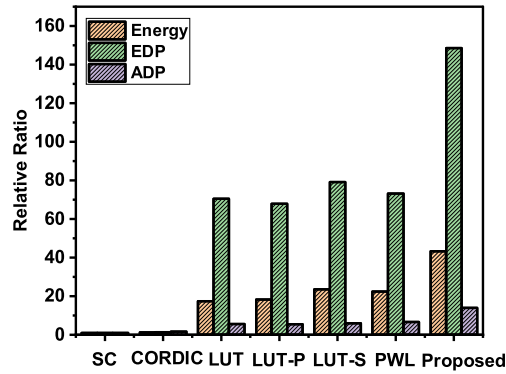
The validation of the proposed method is verified by using the LeNet neural architecture model using the Keras library, and performance parameters are extracted for experimental analysis at a 180 nm technology node

Table 4
Reference activation functions and their design techniques and features.

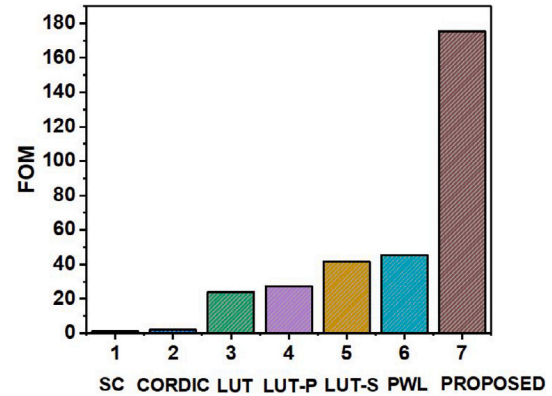
AF model	Technique & Features	Acronyms
1	By storing activation function value directly in LUT or ROM based implementation [20]	LUT
2	Storing parameters value used in the model of activation function [23]	LUT-P
3	Series expansion method based on LUT approach [24]	LUT-S
4	Piece-wise linear function and then its implementation [32]	PWL
5	CORDIC based Configurable activation function [13,33,34]	CORDIC
6	Implementation using stochastic computing [35]	SC
7	Proposed method by truncated the series and then implement with the help of a combinational circuit.	Proposed

Table 5
Implementation of AF using various algorithms with the help of 180 nm technology node at 0.1 GHz.

Quantization level	Area (μm^2)	Leakage power (nW)	Delay (ns)	Energy (nJ)	EDP (ns \times nJ)	ADP (ns \times μm^2)
LUT[20]	3014.85	30.95	2.42	74.89	181.23	7295.93
LUT-P [23]	2789.12	26.82	2.65	71.07	188.23	7391.17
LUT-S [24]	2353.21	18.92	2.92	55.24	161.30	6871.37
PWL [32]	2052.31	19.27	3.01	58.00	174.58	6177.45
CORDIC [33]	2825	119.40	8.87	1050.72	9319.88	25057.75
SC [35]	4150.81	132.42	9.82	1300.36	12769.57	40760.95
Proposed	1024.04	10.51	2.86	30.06	85.97	2928.75



(a)



(b)

Fig. 6. Performance analysis and comparison: (a) Comparison of various designs implementations in terms of energy, EDP, ADP. Ratio were computing by taking smallest value as 1. (b) Figure of merit for different algorithms for implementation of a AF.

4.1. Resources utilization

The experimental evaluation is carried out using RTL of the proposed design, and state-of-the-art architectures are synthesized using design vision-Synopsys. Results are produced by *Design Compiler-Synopsys* at 180 nm technology node. The LUT technique is based on piece-wise linear implementation, whereas stochastic computing and CORDIC-based methods are approximate with respect to iterative computation [33]. The approximate technique has some degree of error. Moreover, with increasing the bit-precision and iterations of computation, the degree of error can be decreased. Hence, the physical parameters are given at 180 nm for 8-bit precision memory element (LUT) based architecture. In stochastic computing-based architecture has a minor degree of error for 8-bit and higher precision, and in CORDIC architecture, computational accuracy is increased with higher precision (16-bit and higher bit) representation. Hence, we selected the 8-bit and 16-bit precision for stochastic computing and CORDIC-based architecture. The parameters are extracted and compared as shown in Table 5. Moreover, based on those bit precision, we have calculated the test accuracy and baseline error for the CIFER-10 dataset as shown in Table 11. The proposed method occupies less area than that of the

other reference methods as shown in the Table 5. The area of the proposed design is reduced by 66.03% as compared with LUT based method, and 63.28%, 56.48%, and 50.01% are compared with LUT-P, LUT-S, PWL methods, respectively. Moreover, among all methods, the SC and CORDIC have less accuracy and more baseline error areas than other techniques.

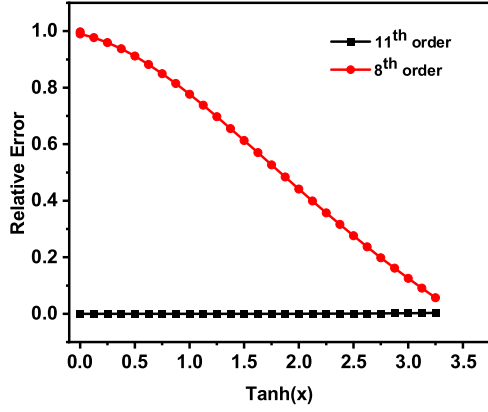
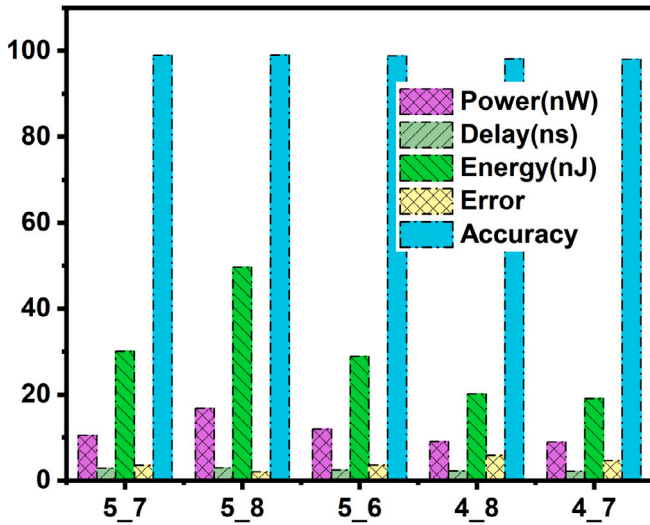
4.2. Power and delay analysis

In the proposed method, power is reduced approximately up to $3\times$ as compared with the conventional LUT-based approach. Power is reduced by 66.04%, but the delay is slightly increased by 18.18% as compared with the LUT-based approach. If we consider the piece-wise linear (PWL) approach, then power and delay are reduced by 45.46% and 4.98% respectively. The area-delay-product (ADP) and the energy-delay-product (EDP) is shown in Fig. 6(a). It shows that proposed method has lower energy, EDP, and ADP as compared with that of the proposed methods as shown in Table 5.

Table 6

Implementation of AF using different quantization Level effects on Tanh with the help of 180 nm technology node at 0.1 Ghz.

Quantization level	Area (μm^2)	Leakage power (nW)	Delay (ns)	Energy (nJ)	EDP (ns \times nJ)	ADP (ns $\times \mu\text{m}^2$)
5_7	1024.04	10.51	2.86	30.06	85.97	2928.75
5_8	1289.12	16.82	2.95	49.62	141.91	3802.90
5_6	1001.21	11.92	2.42	28.85	69.82	2422.93
4_8	995.18	9.11	2.21	20.13	44.49	2109.78
4_7	982.12	8.99	2.12	19.06	40.41	2082.09

**Fig. 7.** Absolute relative error of the proposed AF with respect to the exact tanh function.**Fig. 8.** Comparison between different Quantization level.

4.3. Figure of Merit (FOM)

Here the proposed figure of merit is expressed as Eq. (14). Where A_{norm} is the normalized area, P_{norm} is normalized power and D_{norm} is the normalized delay.

$$\text{Figure of Merit (FOM)} = \frac{1}{A_{norm} \times P_{norm} \times D_{norm}} \quad (14)$$

The FOM of AF for various methods including proposed technique is shown in Fig. 6(b). From the results observed that the proposed design has lower Energy, EDP, and ADP and higher FOM compared with the other implementations; the proposed circuit consumes less power and high performance with a lower area overhead. Since. The proposed circuit is well suited for deep neural network applications.

Table 7

Implementation of AF using various algorithms with the help of 180 nm technology node at 0.1 GHz with quantization level 5_7.

AF model	Area (μm^2)	Leakage power (nW)	Delay (ns)
ReLU	895.21	9.12	1.82
SWISH	1031.21	11.21	3.12
ELISH	1028.12	10.98	2.99
ELU	995.29	10.55	2.88
SELU	994.21	10.41	2.78
Tanh	1024.04	10.51	2.86

Table 8

Summary of datasets (MNIST and CIFAR-10).

Datasets	Training set	Test image set	Output	Image pixel
MNIST	60K	10K	10	28×28
CIFAR-10	60K	10K	10	32×32

4.4. Error analysis

This section analyzes the maximum average error and relative error for the proposed AF algorithm. In this paper, tanh has symmetry property; we have shown these errors for the positive half of the graph. As shown in Fig. 4(b), the absolute error is maximum at 8th order as compared with 11th order. The absolute error of 11th order is calculated using Eq. (15a), where x_a is the true value and x_b is the approximate value. This is why choosing the 11th order tanh function as an AF.

$$\text{Absolute Error} = |x_a - x_b| \quad (15a)$$

$$\text{Average Absolute Error} = \text{Avg} |x_a - x_b| \quad (15b)$$

Average error of 11th order is calculated by using Eq. (15b). The average value comes out to be 0.048% which is very small compared to that of the 8th order having a value of 0.38. The Relative error is calculated using an Eq. (16a). It is observed that the relative error is more for lesser order as a comparison with that of the higher-order is shown in Fig. 7. The average relative error is estimated by using Eq. (16b) for 11th order is 0.52 % and for 8th order 57.28 %. Therefore, seeing these errors, we have truncated the series by 11th order.

$$\text{Relative Error} = \frac{|x_a - x_b|}{|x_a|} \quad (16a)$$

$$\text{Average relative Error} = \text{Avg} \frac{|x_a - x_b|}{|x_a|} \quad (16b)$$

4.5. Different quantization level effects on tanh

The Tanh is implemented using various quantization levels with the help of series expansion and then convert into the combinational design. In this paper, for the comparison, we have chosen 5-bit input and 7-bit output (5_7), 5-bit input 8-bit outputs (5_8), 5-bit input 6-bit outputs (5_6), 4-bit input 8-bit outputs (4_8), 4-bit input 7-bit outputs (4_7). In Table 6 there is a comparison with different quantization levels. It shows that 5_7 has more area but better performance parameters than the lower quantization level. The accuracy and other physical performance parameters are shown in Fig. 8.

Table 9

Experimental results using LeNet architecture after customizing AF for MNIST dataset.

AF deign	Baseline error (%)	Test accuracy (%)	Compute time (s)
LUT	1.98	98.62	348
LUT-P	2.02	98.94	352
LUT-S	3.92	98.82	421
PWL	4.25	98.71	361
CORDIC	5.15	97.98	429
SC	6.12	97.81	435
Proposed	3.48	98.92	249

Table 10

Experimental results at different quantization level using LeNet architecture after customizing AF for MNIST dataset.

Activation function	Baseline error (%)	Test accuracy (%)
5.7	3.48	98.92
5.8	2.01	98.99
5.6	3.52	98.81
4.8	5.82	98.12
4.7	4.62	97.98

4.6. Implementations of other non-linear activation functions

This section has implemented other famous AFs using the above-described implementation using combinational logic and series expansion. Table 7 shows the implemented results of various AFs such as ReLU, SWISH, ELISH, ELU, SELU. For the implementation of these AFs, we have taken 5.7 as a quantization level. We can implement any non-linear function by the method described in the Algorithm. 1. by selecting the appropriate quantization level. This method applies to all the AFs.

$$SWISH = x/(1 + e^{-x}) \quad (17a)$$

$$ReLU = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases} \quad (17b)$$

$$ELU = \begin{cases} \alpha(\exp(x) - 1) & \text{for } x \leq 0 \\ x & \text{for } x > 0 \end{cases} \quad (17c)$$

$$SELU = \lambda \begin{cases} \alpha(\exp(x) - 1) & \text{for } x \leq 0 \\ x & \text{for } x > 0 \end{cases} \quad (17d)$$

$$ELISH = \frac{(e^x - 1)}{(1 + e^{-x})} \text{ or } \frac{x}{(1 + e^{-x})} \quad (17e)$$

5. Experimental analysis and validation

We have performed experiments based on the MNIST and CIFAR-10 dataset on the LeNet architecture model for benchmark analysis. Summary of datasets is shown in Table 8.

5.1. Experiment 1: MNIST

The MNIST dataset is the benchmark dataset for image classification [36,37]. It consists of a 60 thousand training set with a test set of

10 thousand, which has 28×28 grayscale representation which has a range between 0 to 9. This MNIST data is trained with the help of the LeNet architecture model using the Keras module. Training over epochs using the proposed architecture of *Tanh* AF for MNIST dataset on LeNet architecture is shown in Fig. 9. For 11th order the Baseline error is 3.48%. Whereas, for the exact tanh AF, the baseline error is 1.98%, which is less than the proposed one. However, we can trade-off by seeing the overall performance matrix in terms of hardware implementation such as area, power, and Delay.

The experimental analysis of a customized AF in the LeNet model is shown in the Table 9. We have chosen the same batch size and epochs for all the AF designs. Although the baseline error of the proposed one is higher in comparison to the LUT and LUT-P method and test loss is also higher. The Proposed method has an optimum accuracy with lower computation time. In this paper, we have chosen inference time for the computation. However, if we talk about hardware designs and their implementation proposed algorithm is better than other comparisons in terms of energy and accuracy trade-off. Experimental analysis of various quantization levels shown in Table 10 on LeNet architecture using the MNIST dataset.

5.2. Experiment 2: CIFAR-10

The CIFAR-10 dataset is used as a benchmark for image classification. It consists of a training set of 60 thousand and a test set size of 10 thousand. In this CIFAR-10, each instance has size 32×32 colored images of birds, automobiles, dogs, frogs, etc. For validation of our customized AF. We choose the CIFAR-10 dataset on the LeNet model as shown in Table 11. We have fix batch size and epoch 25 and 100, respectively, for all the methods. The Table 11 shows that the baseline error of the proposed one is a little bit higher but has good computation time and accuracy. We see these results proposed that implementing an AF is good enough in terms of all the performance parameters. Experimental analysis of various quantization levels as shown in Table 12 on LeNet architecture using CIFAR-10 dataset.

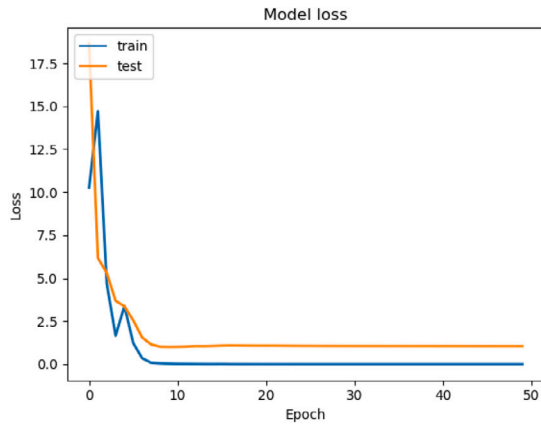
6. Conclusion

A new approximation method for the implementation of a tanh was proposed in this paper using purely combinational logic design. We showed the implementation and its comparative studies with various other approximation techniques. Based on the quantization level, the

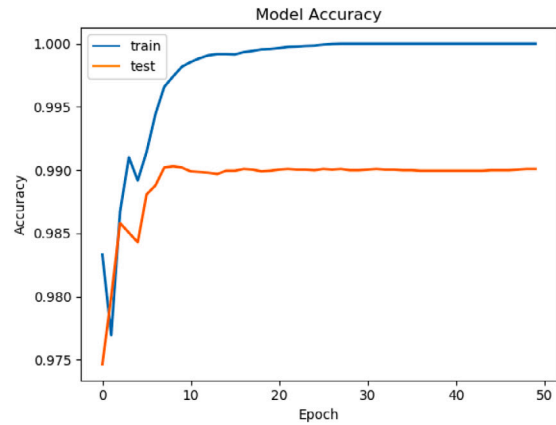
Table 11

Experimental results using LeNet architecture after customizing AF for CIFAR-10 dataset.

AF deign	Baseline error (%)	Test accuracy (%)	Compute time (s)
LUT	6.02	69.02	352
LUT-P	6.03	67.99	361
LUT-S	7.42	68.72	412
PWL	6.12	69.63	392
CORDIC	7.25	66.12	418
SC	7.92	66.02	421
Proposed	6.92	68.99	273



(a) Training and test loss over epoch



(b) Training and test accuracies over epoch

Fig. 9. MNIST dataset on LeNet architecture using customizing AF.

Table 12

Experimental results at different quantization level using LeNet architecture after customizing AF for CIFAR-10 dataset.

Activation function	Baseline error (%)	Test accuracy (%)
5.7	6.92	68.99
5.8	7.12	70.85
5.6	7.59	67.92
4.8	9.82	65.24
4.7	9.12	66.23

proposed model has little effect on accuracy. The hardware implementation of the proposed AF is realized using 180 nm for further evaluation in terms of area, power, and delay. *FOM* of the proposed design is $3.5\times$ as compared to the prior arts. Experimental results on the LeNet model for MNIST and CIFAR10 dataset also show that the proposed design has also optimum accuracy.

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 would like to thank the Council of Scientific and Industrial Research (CSIR), India New Delhi, Government of India under SRF scheme for providing financial support and Special Manpower Development Program Chip to System Design, Department of Electronics and Information Technology (DeitY) under the Ministry of Communication and Information Technology, Government of India for providing necessary Research Facilities.

References

- [1] Kenji Suzuki (Ed.), *Artificial Neural Networks: Architectures and Applications*, BoD-Books on Demand, 2013.
- [2] Z. Hajduk, Hardware implementation of hyperbolic tangent and sigmoid activation functions, *Bull. Pol. Acad. Sci. Tech. Sci.* 66 (5) (2018).
- [3] Naleih M. Botros, M. Abdul-Aziz, Hardware implementation of an artificial neural network using field programmable gate arrays (FPGA's), *IEEE Trans. Ind. Electron.* 41 (6) (1994) 665–667.
- [4] Chigozie Nwankpa, et al., Activation functions: Comparison of trends in practice and research for deep learning, 2018, arXiv preprint arXiv:1811.03378.
- [5] Stamatis Vassiliadis, Ming Zhang, José G. Delgado-Frias, Elementary function generators for neural-network emulators, *IEEE Trans. Neural Netw.* 11 (6) (2000) 1438–1449.
- [6] Pramod Kumar Meher, An optimized lookup-table for the evaluation of sigmoid function for artificial neural networks, in: 2010 18th IEEE/IFIP International Conference on VLSI and System-on-Chip, IEEE, 2010.
- [7] Barry L. Kalman, Stan C. Kwasny, Why tanh: choosing a sigmoidal function, in: [Proceedings 1992] IJCNN International Joint Conference on Neural Networks, Vol. 4, IEEE, 1992.
- [8] K. Basterretxea, Jose Manuel Trela, I. Del Campo, Approximation of sigmoid function and the derivative for hardware implementation of artificial neurons, *IEE Proc.-Circuits Dev. Syst.* 151 (1) (2004) 18–24.
- [9] Ashkan Hosseinzadeh Namin, et al., Efficient hardware implementation of the hyperbolic tangent sigmoid function, in: 2009 IEEE International Symposium on Circuits and Systems, IEEE, 2009.
- [10] Vipin Tiwari, Nilay Khare, Hardware implementation of neural network with sigmoidal activation functions using CORDIC, *Microprocess. Microsyst.* 39 (6) (2015) 373–381.
- [11] Ji Li, et al., Hardware-driven nonlinear activation for stochastic computing based deep convolutional neural networks, in: 2017 International Joint Conference on Neural Networks, IJCNN, IEEE, 2017.
- [12] J. Kadmon, H. Sompolinsky, Transition to chaos in random neuronal networks, *Phys. Rev. X* 5 (4) (2015) 041030.
- [13] Gopal Raut, et al., A CORDIC based configurable activation function for ANN applications, in: 2020 IEEE Computer Society Annual Symposium on VLSI, ISVLSI, IEEE, 2020.
- [14] Djork-Arné Clevert, Thomas Unterthiner, Sepp Hochreiter, Fast and accurate deep network learning by exponential linear units (elus), 2015, arXiv preprint arXiv:1511.07289.
- [15] Guido Baccelli, et al., NACU: a non-linear arithmetic unit for neural networks, in: 2020 57th ACM/IEEE Design Automation Conference, DAC, IEEE, 2020.
- [16] Gopal Raut, et al., Efficient low-precision CORDIC algorithm for hardware implementation of artificial neural network, in: International Symposium on VLSI Design and Test, Springer, Singapore, 2019.
- [17] Vipin Tiwari, Ashish. Mishra, Neural network-based hardware classifier using CORDIC algorithm, *Modern Phys. Lett. B* 34 (15) (2020) 2050161.
- [18] Najmeh Nazari, et al., TOT-net: An endeavor toward optimizing ternary neural networks, in: 2019 22nd Euromicro Conference on Digital System Design, DSD, IEEE, 2019.
- [19] Mohammad Loni, et al., DeepMaker: A multi-objective optimization framework for deep neural networks in embedded systems, *Microprocess. Microsyst.* 73 (2020) 102989.ii.
- [20] Karl Leboeuf, et al., High speed VLSI implementation of the hyperbolic tangent sigmoid function, in: 2008 Third International Conference on Convergence and Hybrid Information Technology, Vol. 1, IEEE, 2008.
- [21] Mustafa Glsu, Mehmet Sezer, A taylor polynomial approach for solving differential-difference equations, *J. Comput. Appl. Math.* 186 (2) (2006) 349–364.
- [22] Tao Yang, et al., Design space exploration of neural network activation function circuits, *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* 38 (10) (2018) 1974–1978.
- [23] Joshua Yung Lih Low, Ching Chuen Jong, A SCory-efficient tables-and-additions method for accurate computation of elementary functions, *IEEE Trans. Comput.* 62 (5) (2012) 858–872.
- [24] Barry Lee, Neil Burgess, Some results on Taylor-series function approximation on FPGA, in: The Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, Vol. 2, IEEE, 2003.

- [25] Babak Zamanlooy, Mitra Mirhassani, Efficient VLSI implementation of neural networks with hyperbolic tangent activation function, *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* 22 (1) (2013) 39–48.
- [26] D.J. Myers, R.A. Hutchinson, Efficient implementation of piece-wise linear activation function for digital VLSI neural networks, *Electron. Lett.* 25 (1989) 1662.
- [27] Ehsan Rasekh, Iman Rasekh, Mohammad Eshghi, PWL Approximation of Hyperbolic Tangent and the First Derivative for VLSI Implementation, *CCECE 2010*, IEEE, 2010.
- [28] Hussein M.H. Al-Rikabi, et al., Generic model implementation of deep neural network activation functions using GWO-optimized SCPWL model on FPGA, *Microprocess. Microsyst.* (2020) 103141.
- [29] Shaghayegh Gomar, Mitra Mirhassani, Majid Ahmadi, Precise digital implementations of hyperbolic tanh and sigmoid function, in: 2016 50th Asilomar Conference on Signals, Systems and Computers, IEEE, 2016.
- [30] Karl Leboeuf, et al., High speed VLSI implementation of the hyperbolic tangent sigmoid function, in: 2008 Third International Conference on Convergence and Hybrid Information Technology, Vol. 1, IEEE, 2008.
- [31] Yann A. LeCun, et al., Efficient backprop, in: *Neural Networks: Tricks of the Trade*, Springer, Berlin, Heidelberg, 2012, pp. 9–48.
- [32] Che-Wei Lin, Jeen-Shing Wang, A digital circuit design of hyperbolic tangentsigmoid function for neural networks, in: 2008 IEEE International Symposium on Circuits and Systems, IEEE, 2008.
- [33] G. Raut, S. Rai, S.K. Vishvakarma, A. Kumar, RECON: Resource-efficient CORDIC-based neuron architecture, *IEEE Open J. Circuits Syst.* 2 (2021) 170–181, <http://dx.doi.org/10.1109/OJCAS.2020.3042743>.
- [34] Martine Wedlake, Harry L. Kwok, A CORDIC implementation of a digital artificial neuron, in: 1997 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, PACRIM.
- [35] Van-Tinh Nguyen, et al., An efficient hardware implementation of activation functions using stochastic computing for deep neural networks, in: 2018 IEEE 12th International Symposium on Embedded Multicore/Many-Core Systems-on-Chip, MCSoc, IEEE, 2018.
- [36] Han Xiao, Kashif Rasul, Roland Vollgraf, Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017, arXiv preprint [arXiv: 1708.07747](https://arxiv.org/abs/1708.07747).
- [37] L. Bottou LeCun, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* 86 (11) (1998) 2278–2324.



Gunjan Rajput received B.Tech. degree in Electronics & Telecommunication Engineering and M.tech. degree in VLSI and Embedded Engineering from Delhi Technological University (DCE), Delhi, India in 2012 and 2015 respectively. Currently she is with Nanoscale Devices, VLSI Circuit & System Lab and pursuing Ph.D. degree in Electrical Engineering Department, Indian Institute of Technology Indore, India. Her current research interest includes in Machine learning, Edge computing, hardware accelerator design, and Reliability.



Gopal Raut received the B.Engg. in electronic engineering and M.Tech. in VLSI Design from G H Raisoni College of Engineering Nagpur, India, in 2015. He is currently pursuing the Ph.D. degree with the Electrical Engineering Department, Indian Institute of Technology Indore, India. His research focus is compute-efficient and configurable VLSI circuit design for low power IoT and Edge AI applications.



Santosh Kumar Vishvakarma is currently an Associate Professor with the Department of Electrical Engineering, Indian Institute of Technology Indore, India, where he is leading Nanoscale Devices and VLSI Circuit and System Design Lab. He got his Ph.D. degree from Indian Institute of Technology Roorkee, India, in 2010. From 2009 to 2010, he was with the University Graduate Center, Kjeller, Norway, as an Post-Doctoral Fellow under European Union Project “COMON.”. His current research interests include nanoscale devices, reliable SRAM memory designs, and configurable circuits design for IoT application.



WASTE MANAGEMENT BY “WASTE TO ENERGY” INITIATIVES IN INDIA

M. Kumar^{1*}
S. Kumar²
 S.K. Singh³

¹Assistant Professor, Dept. of Physics, SSNC, University of Delhi, Delhi, India.

Email: physics.ssn@gmail.com Tel: 919212711976

²Research Scholar, Dept. of Environmental Engineering, Delhi Technological University, Delhi, India.

Email: sandeepkumar200393@gmail.com Tel: 918860351400

³Professor and Head, Dept. of Environmental Engineering, Delhi Technological University, Delhi, India.

Email: sksinghdce@gmail.com Tel: 919891599903



(+ Corresponding author)

ABSTRACT

Article History

Received: 12 January 2021

Revised: 22 February 2021

Accepted: 26 March 2021

Published: 28 April 2021

Keywords

Waste management
Waste to energy
Plasma technology
Incineration.

This century is known for exponentially growing population and development but with huge waste production. The waste produced requires land, labour and capital to for treatment and disposal of such huge amount of waste. In India, people throw or consider it as waste after single use so Indian waste can be good resource for recovery of various products. The waste produced is difficult to manage using conventional methods and is ever increasing, blocking essential land that has become an expensive commodity in today's world. This work explores the current practices of the various waste management initiatives and a critical assessment of traditional waste to energy procedure adopted in India. It gives an overview of the various waste management systems in India. Suggestions for improving the health of society, waste management processes, process performance, environmental assessment parameters to plasma gasification, an alternate waste to energy has been also discussed. Recommendation has been made for the micro-waste plant to solve the waste challenges.

Contribution/Originality: The main contribution of the paper is to assess waste management in India and certain waste emerging innovations –Waste to Energy, which are technically applicable and relevant. It also addresses how the use of advanced waste technologies like plasma can be a way of achieving a circular economy as well as less environmental impact.

1. INTRODUCTION

Energy is one of the foods for technical or economic development of human beings. Rapid increase in population has resulted in huge demands for energy to for material production. Such thrust for energy and to recover more energy requires technological exploitation of energy resources (Ramos & Rouboa, 2020; Young, 2010). The materials byproducts are related to waste, an inevitable by-product of industrial production. The exponentially growing population has increased the waste production to many fold. Although waste is shown to be a non-essential qualitative component of industrial production, the quantitative scope of waste can vary according to the degree of (in)efficiency with which these processes are operated within certain limits. Over the years, the invention of new products, innovations and facilities has altered the quantity and quality of waste. Waste characteristics do not only depend on income, culture and geography but also on a society's economy and situations like disasters that affect that economy (Ionescu et al., 2013; Kumar, Khare, & Alappat, 2002; Kumar, 2014; Kumar,

Kumar, & Singh, 2020; Kumar & Samadder, 2017; Leena, Sunderesan, & Renu, 2014; Rawat, Kaalva, Rathore, Gokak, & Bhargava, 2016; Vats & Singh, 2014; Vats & Singh, 2014).

Table-1. MSW waste estimate with population.

Year	Population (in million)	Total waste generation (million MT/year)(@0.4kg/capita/day)	Total waste generation (million MT/year) (@0.6kg/capita/day)
2015	1310.15	191.2819	286.9229
2020	1381.59	201.7121	302.5682
2025	1450.52	211.7759	317.6639
2030	1503.64	219.5314	329.2972
2035	1553.723	226.8436	340.2653
2040	1592.69	232.5327	348.7991
2045	1620.61	236.6091	354.9136
2050	1639.17	239.3188	358.9782

Note: *population data from worldometer web.

Waste challenges in metropolitan centers include the growing challenge of acquiring expensive land for disposal, producing emissions from waste treatment and disposal, etc (Sharma & Shah, 2005; Vats & Singh, 2014). The disposal of waste has caused resource depletion and the huge cost involved in waste processing and transportation.

Established processes for the collection, transport and treatment of solid waste are mired in confusion in India. Uncontrolled waste disposal has created overflowing landfills on the outskirts of neighborhoods, which are not only very difficult to retrieve due to haphazard dumping practices, but can have significant environmental effects in terms of water contamination, land degradation and air pollution that lead to global warming. Environmental degradation is taking place and organizations that are responsible for environmental management are facing many problems and challenges. Uncontrolled waste disposal and unsustainable waste management not solely harm the atmosphere, but conjointly have an effect on human health (Central Pollution Control Board (CPCB), 2004; Jha, Singh, Singh, & Gupta, 2011; Kumar & Samadder, 2017). The new scheme relies on the storage and transport of mainly mixed, unsegregated waste.

The 5R solution - Recycling, Reduce, Reuse, Refuse, Recover, Residual Management with sustainable disposal of residual waste in science-based landfills is grossly ignored (Abhishek & Mukherjee, 2019; Alam & Ahmade, 2013; Anubhav, Abhishek, & Durgesh, 2012; Cleary, 2009; Kumar et al., 2017; Nandan, Yadav, Baksi, & Bose, 2017; Otitoju & Seng, 2014; Srinivas, 2007; Sudha, 2008; UN, 2000; World Energy Council Report, 2013; Young, 2010). This work explores the solid waste production status and its environmental and financial impact on Indian cities. This study also analyses the growing number of municipal solid waste (Kumar et al., 2016; Sudha, 2008; World Energy Council Report, 2013) the changing nature of municipal solid waste, from biodegradable waste, dry waste to the increasing volume of plastic in the waste (Cleary, 2009; Devi & Satyanarayana, 2001; Hargreaves, Adl, & Warman, 2008; Indo-UK Seminar Report, 2015; Jha et al., 2011; Kumar & Samadder, 2017).

This work also presents the sources of waste-to- energy / energy-from-waste conversion technology for the solid waste sector. Laws for sustainable solid waste disposal have already been set in motion, but a big obstacle is the need to plan and maintain the scheme and ensure implementation of the rules (Sharma & Shah, 2005; Vats & Singh, 2014).

In addition to providing some mitigation options to respond to the growing problem, current governments recommend publicly-engaged frameworks to ensure that the framework is financially sustainable. There are many cleaner technologies for dealing with waste but lack of knowledge and public awareness makes waste management a menace. Public participation is required to deal with the generated waste at source itself.

2. MUNICIPAL SOLID WASTE IN INDIAN CONTEXT

India, the second most populated country in the world and one of the fastest-growing economies, is experiencing unprecedented growth in its industrial sector and is undergoing rapid urbanization. The population of India is approximately 1.3 billion and experts believe that each day a single person is generating 450 grams of waste (Central Pollution Control Board (CPCB), 2000, 2004; Kumar et al., 2016).

Current rate of municipal waste projection as per population growth is shown in Table 1. The study predicts that MSW generation will reach 219- 330 million MT/year by 2030 and 240-358 million MT/year by 2050. Much variability of per capita waste generation is found in accordance with the size and class of the cities. As per CPCB report, in 2012, 1,27,486 tons per day of MSW is being produced from household activities and other commercial & institutional activities (Abhishek & Mukherjee, 2019; Kumar et al., 2016; World Energy Council Report, 2013).

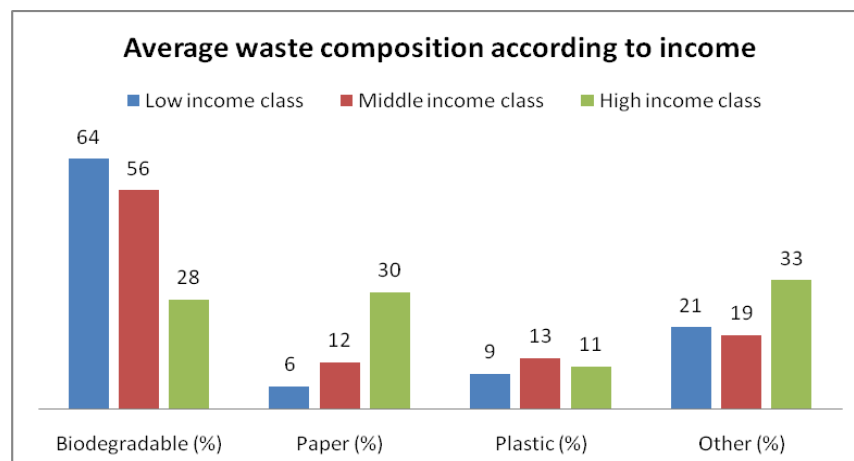


Figure-1. Waste composition with different income class.

Source: Kumar and Samadder (2017).

There is no difference in the types of waste generated in the physical characterization data of MSW in metropolitan cities of India for the last 2 decades, although there is an increase in the quantity of waste produced. Figure 1 show that the urban MSW in India can be classified as 40-50% biodegradables, 15-20% recyclables and 31% of inert wastes with moisture content of 47% and average calorific value of 7.3 MJ/k (Jha et al., 2011; Kumar et al., 2017; Kumar et al., 2020; Leena et al., 2014).

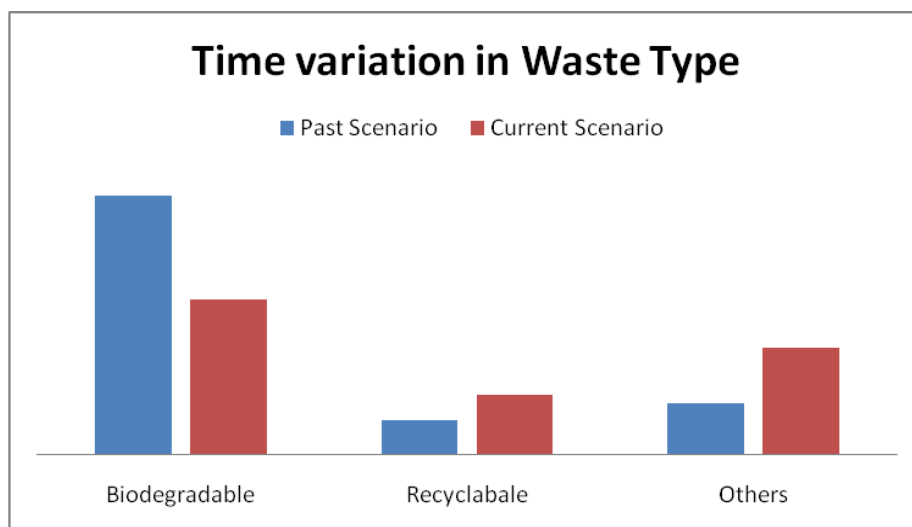


Figure-2. Changing waste scenario.

Source: Abhishek and Mukherjee (2019); Nandan et al. (2017); Paulraj, Bernard, Raju, and Abdulmajid (2019).

The 20th century has led to drastic change in the type of waste produced as shown in Figure 2. Earlier, there were large quantity (70-80%) of biodegradable waste but there is only 45-50% of biodegradable waste and 50% of other waste including plastic, paper, hazardous, biomedical etc (Abhishek & Mukherjee, 2019; Nandan et al., 2017; Paulraj et al., 2019). Higher consumerism, rapid population growth with unplanned urban development, and lifestyle changes have led to increased volumes in solid waste as well as more plastics and certain inorganic materials contents.

The generated waste engagement is not a new cup of tea, but a long pending and ignore field in view of low quantity. The rapid increase in population and industry has grown as shown in data table. The searches for new and new technology are in force, as it has started affecting the human beings. The method like composting, bio-methanation and combustion are few oldest methods for waste treatment in India, but limited to organic waste like food / plant material. They basically target organic waste biological decomposition with /without the presence of oxygen, e.g. bio-methanation, combustion. The biggest advantage of such technique is that it does not only reduce nature affecting gas like methane, but also generates – a powerful greenhouse gas. It can simultaneously generate electricity, cooking gas and inert residue which can be used as manure. One of the biggest limitations of these processes is the long and spacious process. Therefore, their rate of treatment fails to target the amount of waste generated, and men has started using the landsite near /outside man colonies for waste treatment and safety issues. Land filling has emerged as one of the cheapest and easiest methods of SWM; burns on low level areas are target areas of dumping solid waste thus leveling the ground for useful purpose. Neither manmade technologies nor nature is capable to treat this huge quantity of waste. The escaped harmful gases and products have started affecting the environment and mankind.

The traditional solid waste management processes, such as composting, bio-methanation and land filling, suffer from the disadvantage of environmental deterioration and space problem, as composting and bio-methanation requires large area for treatment and it takes long period of time. For land disposal of solid waste, India needs 1,240 hectares of extra valuable land every year to include untreated solid waste. As per report published by Ministry of Urban Development, government of India, 2014, Solid waste produced was 133000 MT/day, Total waste collected is 91000 MT/day, waste littered 42000 MT/day, from the collected MSW 26000MT/day is treated and 66000 lakh MT/day landfilled (crude dumping). The waste generated is 133000 MT/day and waste collected, treated, littered and land filled has been shown in Figure 3.

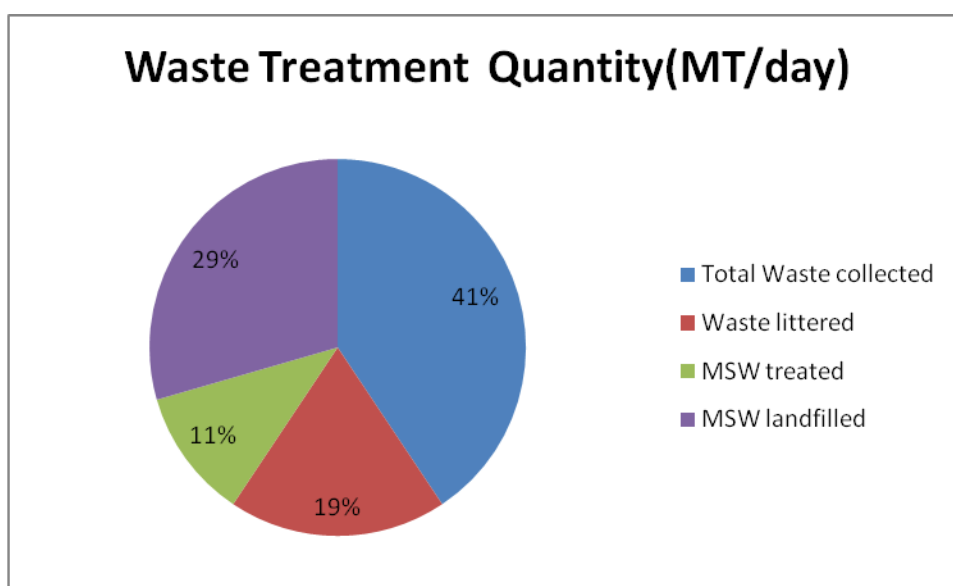


Figure-3. Current municipal waste management.

Source: Central Pollution Control Board (CPCB) (2000) and Central Pollution Control Board (CPCB) (2004) and Satpal (2020).

Figure 4 show a comparisons of various state waste segregation in % .

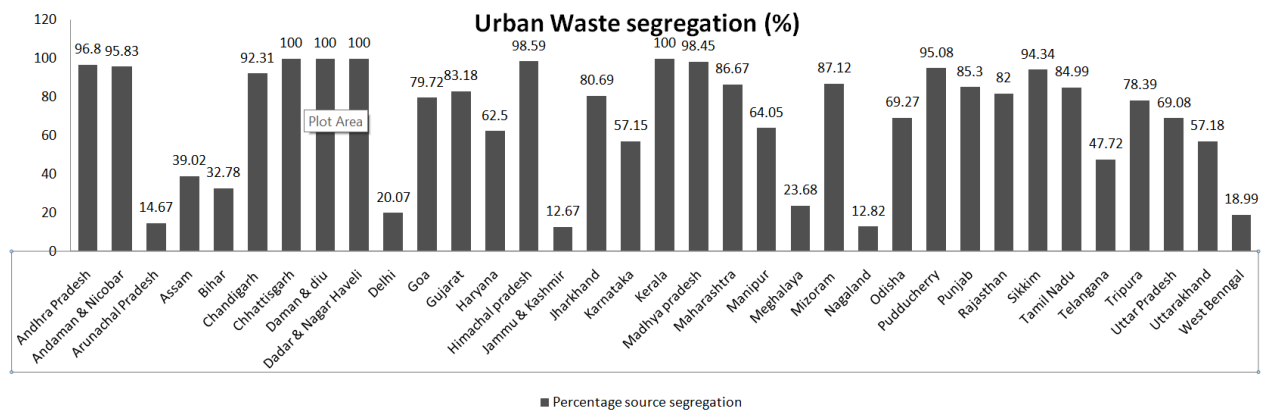


Figure-4. State wise percentage waste segregation.

Date Source: Central Pollution Control Board (CPCB) (2000) and Central Pollution Control Board (CPCB) (2004) and Satpal (2020); Sudha (2008).

The calorific value of Indian municipal waste ranges from 800 Kcal to 1100 Kcal and moisture content 40% to 50%. The traditional techniques are time taking processes so heap/ mountain of waste has formed which has affected the aesthetic beauty of the city; and in these processes, foul smell is produced and also contributed to many environmental problems, such as global warming, ozone depletion, human health hazards, ecosystem damages, abiotic resource depletion, etc. (Khandelwal, Dhar, Thalla, & Kumar, 2019). This further leads to a lack of public approval for new waste management sites. The existing landfill sites in mega cities like Delhi, Kolkata and Mumbai have dangerously exceeded their capacity already. Moreover, the traditional waste disposal technique by landfill is considered the most unfavorable route in the waste management hierarchy, as it wastes valuable land and gives rise to Green House Gases (GHG) emissions, primarily methane (Khandelwal et al., 2019).

3. MODERN TECHNOLOGIES FOR MUNICIPAL SOLID WASTE MANAGEMENT (MSWM)

The various studies have been conducted on traditional waste management methods based on cost of the technology, environmental impact assessment, life cycle etc. Top 5 cities in waste processing is shown in Figure 5 (Satpal, 2020). The performance of applied methods is based on their geographical location, and input waste type. All the traditional technologies have shortcomings of waste generation, time required and ash content produced so there is a need of technological advancement in this field which can overcome all these limitations.

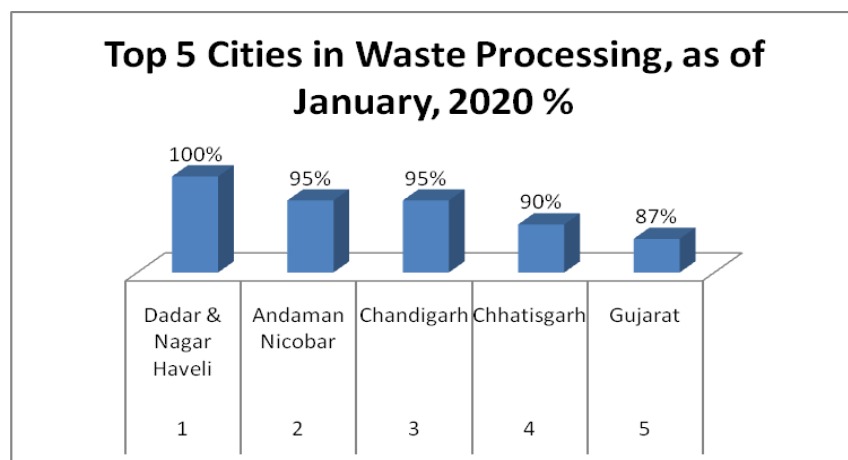


Figure-5. Top 5 cities in waste processing (JAN, 2020).

Source: Central Pollution Control Board (CPCB) (2000) and Satpal (2020).

The various methodologies for waste to energy (WtE) have evolved from combustion, gasification, incineration (Table 2). The different processes of thermo chemical treatment, such as composting, incineration, pyrolysis, etc., are an essential component of the management system of sustainable integrated municipal solid waste (MSW) (Kumar & Samadder, 2017; Otitoju & Seng, 2014; Sudha, 2008). Thermal treatment plants can in fact convert MSW into different energy forms, such as electricity and heat for both utilization in industrial facilities or district heating (Kumar, 2013; World Energy Council Report, 2013; Young, 2010). The advancement in technologies for solid waste management cannot limit generation of waste, the only possible solution can be to help nature to convert waste into natural components. From combustion, gasification, incineration to plasma technology (Abhishek & Mukherjee, 2019; Paulraj et al., 2019) various waste-to-energy (WtE) methodologies have emerged. All WtE itself needs additional material and energy resources and waste as a resource and contributes in absolute terms to a decrease in per capita waste generated (Devi & Satyanarayana, 2001). Table 2 show such WtE plants with energy generation in various Indian cities.

Table-2. Modern technologies for Municipal Solid Waste Management.

	Combustion	Land filling	Incineration	Gasification	Pyrolysis
Aim of the process	Waste to high T flue gases	Maximize waste decomposition,	Waste conversion to high temperature	Waste to high heating value flue gases	Max. thermal decomposition of solid
Flue Gases	CO ₂ , H ₂ , CO, H ₂ O and particulate matter.	CO ₂ and CH ₄	CO ₂ and H ₂ O	CO, H ₂ , CO ₂ , H ₂ O and CH ₄	CO, H ₂ , CH ₄ and other hydrocarbons
Operating condition reaction environment	Oxidant amount larger than required) in presence of air.	Oxidizing at the upper layer and reducing beneath the surface.	Oxidant amount larger in presence of air between 850°C and 1200°C	Lower oxidant in oxygen enriched air, steam Between 550°C and 900°C (in air gasification) and 1000-1600°C	Total absence of any oxidant between 500°C and 800°C
Pressure P	Atmospheric	Atmospheric	Generally atmospheric	Generally atmospheric	Slight over-pressure
Pollutants	CO ₂ , H ₂ , CO, H ₂ O and particulate matter.	CO ₂ , CH ₄ , SO _x , NO _x H ₂ , CO, H ₂ O & particulate matter.	SO ₂ , NO ₂ , HCl, PCDD/F, particulate	H ₂ S, HCl, COS, NH ₃ , HCN, tar, alkali, particulate	H ₂ S, HCl, NH ₃ , HCN, tar, particulate
Ash	Large amount of ash is produced.	No ash	Ash – ferrous, non-ferrous metals and inert materials for sustainable utilization.	Vitreous slag that can be utilized as backfilling material	Non- negligible carbon content
Gas cleaning	-	-	Can be made under emission limits	Possible to have clean synthetic gas to meet the standards of chemicals production processes or with high efficiency energy conversion devices	
Waste reduction (w/w)	60%	10-20%	70%	82%	84%
Ash production	Yes	No	Yes	Yes	Yes

Source: Abhishek and Mukherjee (2019); Paulraj et al. (2019); Kumar et al. (2020); Young (2010).

As per 2020 study, almost every state process waste for WtE and Figure 5 show top 5 waste processing cities. Owing to the increasing collection, recycling and reuse of waste are economically viable choices since a large portion of the waste management budget is used to collect and transport waste (Devi & Satyanarayana, 2001;

Khandelwal et al., 2019). Such utilization establishes waste as resources and contributes to the circular economy as a key. Thus, parameters such as the investment, the return period and the monetary revenue constitute challenges to overcome so as to implement this environmentally favorable technique.

Table-3. Operating WtE plant in India

Operational WtE plants in India			
Location	Developer	Capacity (TDP)	Electricity Generation (MW)
Delhi-Okhla	Jindal	1950	16
Delhi- Gazipur	IL&FS	1300	14
Delhi- Bawana	Ramky	2000	24
Hyderabad	Ramky	2400	20
Hyderabad	IL&FS	1000	11
Chennai	Essel	300	2.9
Jabalpur	Essel	600	0.9
Shimla	Elephant Energy	70	1.75

Source: Municipal bodies of different cities/ miscellaneous.

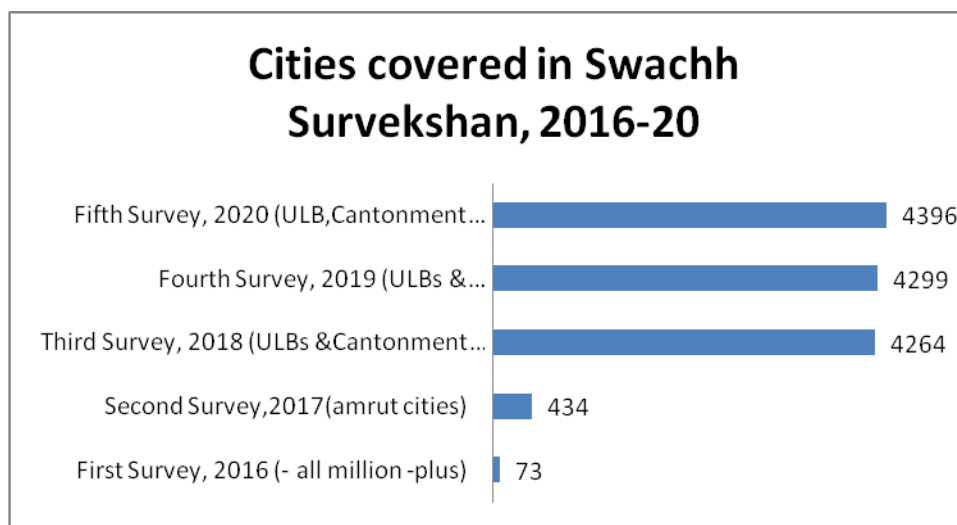


Figure-6. Cities covered in swachh survekshan.

Source: Kumar (2013); Rada, Istrate, and Ragazzi (2009), Central Pollution Control Board (CPCB) (2000) and Satpal (2020).

4. MSW SOCIAL APPROACH

The various waste planning approaches are based on waste generation quantities, local waste characteristics, local geographical conditions, land accessibility and other relevant criteria. The large volume of waste places special emphasis on community or stakeholder contribution and inter-departmental coordination at the local-authority level to ensure implementation success. The easiest way to reduce waste reaching landfill sites is by involving more stakeholders like NGOs, local people and other organizations. This can be achieved by spreading education and awareness among people about waste as resource, it will help in saving money (Kumar, 2013; Rada et al., 2009).

In this context, various Indian government Initiatives for waste management like Urban Infrastructure Development Scheme for Small & Medium Towns (UIDSSMT), "Recycled Plastics Manufacture and Usage Rules (1999), The Plastics Manufacture and Usage (Amendment) Rules (2003), Central Pollution Control Board (CPCB), Non-biodegradable Garbage (Control) Ordinance, 2006, Municipal Solid Wastes (Management and Handling) Rules, 2000, Swachh Bharat Mission(2014) , Solid Waste Management Rules (2016), are few of them. The government of India also understand the importance of public participation either at source level or management level so public participation has increased from 2016-2020 in Swachh Survekshan (Satpal, 2020) as shown in Figure 6. The sustainable waste management of solid waste without public participation is not possible.

5. PLASMA TECHNOLOGY AS AN ALTERNATIVE FOR INDIAN MSW MANAGEMENT

Plasma is the ionized state of matter and is created through the application of energy sourced from electric discharges of frequencies ranging from Direct Current (DC) to the optical range. It is formed whenever ordinary matter is heated over few thousand degree C, which results in electrically charged gases or fluids. The various developing countries have started using plasma as a most feasible solution to the impending and escalating waste management crisis from household waste to other hazardous wastes such as medical wastes. The plasma treatment is based on their high temperature, intense and non-ionising radiation nature. Thermal plasmas can be used to treat all kinds of waste streams, be it solid such as regular MSWs, liquid such as urine or poisonous gases. Due to the high temperature and high energy density generated by thermal plasma, a large throughput can be accommodated with a small scale reactor. The high flux densities generated by the plasma at the reactor boundaries lead to a rapid attainment of steady state conditions, effectively reducing the start-up and shutdown times. The plasma for MSW is effective in two forms. - Plasma pyrolysis and Plasma gasification (Kumar et al., 2020).

Plasma pyrolysis is the combination of thermal-chemical properties of plasma with the pyrolysis process. It completely decomposes waste material into simple molecules with use of extremely high temperatures of plasma-arc in an oxygen starved environment. This technology is particularly appropriate for treatment of solid waste and can also be employed for destruction of toxic molecules by thermal decomposition. Unlike incinerators, segregation of waste is not required in this process. Another advantage of plasma pyrolysis is the reduction in volume of waste, nearly 95%. The numerous advantages of plasma technology it is evident that in the near future, plasma pyrolysis reactors will be widely accepted for toxic waste treatment. The quantity of toxic emissions (dioxins and furans) is much below the accepted emission standards and does not require segregation of hazardous waste. In addition, the disease causing micro-organisms are completely killed and there is a possibility to recover energy.

In plasma gasification, waste is heated to temperatures anywhere from about 1000–15,000°C (1800–27,000°F), but typically in the middle of that range, melting the waste and then turning it into vapor. The end result is the production of synthetic gas (syngas), composed pre-dominantly of carbon monoxide and hydrogen, although certain percentage of carbon dioxide and hydrochloric acid are present, along with vitrified slag which contains molten form of all the inorganic components such as metal scrap present in the MSW feed along with any residual toxic components in inert form. The syngas can be piped away and burned to make energy (some of which can be used to fuel the plasma arc equipment), while the "vitrified" (glass-like) rocky solid can be used as aggregate (for road building and other construction). Plasma gasification used for MSW would require no sorting of materials, eliminate the need for landfills, remove long-haul trucking from our roads and be financially viable. Syn gas can also be converted into high-value products such as highly pure hydrogen, fuels, and other valuable chemical compounds.

Plasma is the sole source of heat in both technologies. No combustion takes place and the end result is the production of synthetic gas (syngas). In fact, the syngas may be contaminated with poisonous gases such as dioxins that must somehow be scrubbed out and disposed of, although some contaminated material may also be found in the rocky solid. Such revenue generation can make it financially viable (Kumar et al., 2020). Plasma gasification used for MSW would require no sorting of materials, eliminate the need for landfills, remove long-haul trucking from our roads and be financially viable.

6. PLASMA TECHNOLOGY ASSESSMENT

The plasma technology has many challenges such as high installation cost, moderated community readiness level, requirement of proper waste sorting less popular, limited process understanding. Because of congested and narrow roads, no single collection mode is effective, economical and efficient in India. Because of the heterogeneity of urban waste, the process of selecting the right waste disposal method is complex. Appropriate method of waste disposal can save money and avoids future problems. The cost of plasma technology is relatively high as compared to other technologies but the cost can be balanced with the revenue generation by selling of the products such as

electricity generation from syn gas. This technology is solving many problems such as ash disposal. All technologies require land for ultimate disposal of waste but plasma technology is returning all material in atomic form back to the nature/environment.

The rapid increase in population is also affecting the electricity demand of the country. The current available energy supply is much lower than the actual energy demand for consumption in many of the developing countries. At present, major source of energy throughout the world is fossil fuels that meet the demand of approximately 84% of the total electricity generation. With the use of plasma technology, the generated electricity can serve as a potential to overcome the energy demand and load on fossil fuels. The Plasma gasification technology is relatively new and people have limited awareness about the technology also people have various safety concerns about its extreme process conditions so it was rated at a moderate community readiness level. These observations may also be due to plasma gasification being a relatively new technology for waste-to-value processing and waste management, the current lack of standards and government regulations, a limited number of prototype units, and scepticism of environmental effects of the technology. From a practical point of view, it is necessary for plasma gasification to have higher levels of CRL and general public approval. Regardless of how sound its technical concept is, if the public is concerned about the technology then politicians, companies, or end-users will be less motivated towards the implementation of such a technology.

Public readiness can be improved by spreading public awareness about waste to value technology. Health equipments and kits can be developed for the operator to make it safe for health of the people working on the plant. Technology readiness levels assessment examine a technology based on requirement, concept and capabilities on a scale of 0 to 9 with 9 being the most mature technology. The plasma gasification technology is rated moderate to high as it partially achieved the first eight levels of TRL.

All technologies require large area for installation, operation and transportation of waste. To overcome this problem, small plants of plasma technology can be in-situ installed. The small plasma technology plant can be installed at a community level or it can be installed in hospitals etc. this will reduce the cost of transportation as well as reduce the traffic on roads from trucks transporting waste. The products can be utilized at the community level itself or it can be sold in market. All technologies have pros and cons related in their application. Considering all the fields and cost to benefit ratio, plasma technology can be said as suitable option for treatment of all type non-biodegradable waste in India.

7. DISCUSSION

The amount of MSW produced in India is rapidly increasing because of population increase and lifestyle change. The utilization of traditional methods for waste treatment isn't sufficient to handle such an outsized quantity of waste. Now Indian government has understood the gravity of the waste management problem, and shifting to science-based solution of waste to energy conversion. The installation of such waste plant at community level can solve our problem. The multiple approaches with the assistance of plasma waste technology through public participation will eliminate the necessity for landfills, remove long-haul trucking from our roads. Government understands the necessity of public involvement and initiated several policies, activities and initiatives like Swachhta Bharat, Zero plastic use in solving MSW problem. Implementation of such policies will help to extend the share recycling of waste and re-using, an economically attractive option. Plasma based WtE is comparatively new technology which has high technical value, high efficiency, high installation cost, but with low awareness level and safety. the top product syngas features a high revenue generation capacity which may make it financially viable and may be a proven solution for all MSW- burning issue in India.

Funding: This study received no specific financial support.

Competing Interests: The authors declare that they have no competing interests.

Acknowledgement: All authors contributed equally to the conception and design of the study.

REFERENCES

- Abhishek, & Mukherjee, S. (2019). *Review of waste to energy projects in developing countries*. Paper presented at the The International Conference on Energy and Sustainable Futures (ICESF).
- Alam, P., & Ahmade, K. (2013). Impact of solid waste on health and the environment. *International Journal of Sustainable Development and Green Economics (IJSDEG)*, 2(1), 165-168.
- Anubhav, O., Abhishek, C., & Durgesh, S. (2012). Solid waste management in developing countries through Plasma Arc gasification. *APCBEEES Procedia*, 1, 193-198.
- Central Pollution Control Board (CPCB). (2000). Management of municipal solid waste in Delhi. Retrieved from http://www.cpcb.nic.in/divisionsofheadoffice/pcp/MSW_Report.pdf. [Accessed 27 June, 2015].
- Central Pollution Control Board (CPCB). (2004). *Management of municipal solid waste*. New Delhi, India: Ministry of Environment and Forests.
- Cleary, J. (2009). Life cycle assessments of municipal solid waste management systems: A comparative analysis of selected peer-reviewed literature. *Environment International*, 35(8), 1256-1266. Available at: <https://doi.org/10.1016/j.envint.2009.07.009>.
- Devi, K., & Satyanarayana, V. (2001). *Financial resources and private sector participation in SWM in India*. New Delhi: Indo-US Financial Reform and Expansion (FIRE) Project.
- Hargreaves, J., Adl, M., & Warman, P. (2008). A review of the use of composted municipal solid waste in agriculture. *Agriculture, Ecosystems & Environment*, 123(1-3), 1-14. Available at: <https://doi.org/10.1016/j.agee.2007.07.004>.
- Indo-UK Seminar Report. (2015). Sustainable solid waste management for cities: Opportunities in SAARC countries. Retrieved from [http://www.neeri.res.in/Short%20Report_Indo-UK%20Seminar%20\(25-27th%20March%202015\).pdf](http://www.neeri.res.in/Short%20Report_Indo-UK%20Seminar%20(25-27th%20March%202015).pdf). [Accessed 27 June 2015].
- Ionescu, G., Rada, E. C., Ragazzi, M., Mărculescu, C., Badea, A., & Apostol, T. (2013). Integrated municipal solid waste scenario model using advanced pretreatment and waste to energy processes. *Energy Conversion and Management*, 76, 1083-1092. Available at: <https://doi.org/10.1016/j.enconman.2013.08.049>.
- Jha, A. K., Singh, S., Singh, G., & Gupta, P. K. (2011). Sustainable municipal solid waste management in low income group of cities: a review. *Tropical Ecology*, 52(1), 123-131.
- Khandelwal, H., Dhar, H., Thalla, A. K., & Kumar, S. (2019). Application of life cycle assessment in municipal solid waste management: A worldwide critical review. *Journal of Cleaner Production*, 209, 630-654. Available at: <https://doi.org/10.1016/j.jclepro.2018.10.233>.
- Kumar, S., Smith, S. R., Fowler, G., Velis, C., Kumar, S. J., Arya, S., & Cheeseman, C. (2017). Challenges and opportunities associated with waste management in India. *Royal Society Open Science*, 4(3), 160764.
- Kumar, D., Khare, M., & Alappat, B. J. (2002). *Threat to the groundwater from the municipal landfill sites in Delhi, India*. Paper presented at the Proceedings of 28th WEDC Conference, 18-22 November, Kolkata, India.
- Kumar, M. (2014). Taming waste via laws of Physics. *International Journal of Sustainable Energy and Environmental Research*, 3(3), 164-170.
- Kumar, M., Kumar, S., & Singh, S. K. (2020). Plasma technology as waste to energy: A review. *International Journal of Advanced Research*, 8(12), 2320-5407.
- Kumar, S., Dhar, H., Nair, V. V., Bhattacharyya, J., Vaidya, A., & Akolkar, A. (2016). Characterization of municipal solid waste in high-altitude sub-tropical regions. *Environmental Technology*, 37(20), 2627-2637.
- Kumar, A., & Samadder, S. R. (2017). A review on technological options of waste to energy for effective management of municipal solid waste. *Waste Management*, 69, 407-422.
- Kumar, M. (2013). Project report on solid waste management through enhancing people awareness for environmental sustainability in Delhi.
- Leena, S., Sunderesan, R., & Renu, S. (2014). Waste to energy generation from municipal solid waste in India. *International Journal of ChemTech Research*, 6(2), 1228-1232.

- Nandan, A., Yadav, B. P., Bakshi, S., & Bose, D. (2017). Recent scenario of solid waste management in India. *World Scientific News*, 66, 56-74.
- Otitoju, T. A., & Seng, L. (2014). Municipal solid waste management: Household waste segregation in Kuching South City, Sarawak, Malaysia. *American Journal of Engineering Research*, 3(6), 82.
- Paulraj, C. R. K. J., Bernard, M. A., Raju, J., & Abdulmajid, M. (2019). Sustainable waste management through waste to energy technologies in India-opportunities and environmental impacts. *International Journal of Renewable Energy Research (IJRER)*, 9(1), 309-342.
- Rada, E., Istrate, I., & Ragazzi, M. (2009). Trends in the management of residual municipal solid waste. *Environmental Technology*, 30(7), 651-661.
- Ramos, A., & Rouboa, A. (2020). Renewable energy from solid waste: Life cycle analysis and social welfare. *Environmental Impact Assessment Review*, 85, 106469. Available at: <https://doi.org/10.1016/j.eiar.2020.106469>.
- Rawat, J., Kaalva, S., Rathore, V., Gokak, D., & Bhargava, S. (2016). Environmental friendly ways to generate renewable energy from municipal solid waste. *Procedia Environmental Sciences*, 35, 483-490. Available at: <https://doi.org/10.1016/j.proenv.2016.07.032>.
- Satpal, S. (2020). Solid waste management in urban India: Imperatives for improvement. ORF Occasional Paper No. 283, November 2020, Observer Research Foundation.
- Sharma, S., & Shah, K. (2005). *Generation and disposal of solid waste in Hoshangabad*. Paper presented at the Proceedings of the Second International Congress of Chemistry and Environment, Indore, India.
- Srinivas, H. (2007). The 3R concept and waste minimization. GDRC Research Output - Concept Note Series E-093. Kobe, Japan: Global Development Research Center.
- Sudha, G. (2008). Municipal solid waste management in India: A critical review. *Journal of Environment, Science and Engineering*, 50(4), 319-328.
- UN. (2000). *State of the environment in Asia and the pacific*. (UN), United Nations. New York: United Nations Publication.
- Vats, M. C., & Singh, S. K. (2014). Status of E-waste in India-a review. *International Journal of Innovative Research in Science, Engineering and Technology*, 3(10), 16917-16931.
- Vats, M., & Singh, S. (2014). E-waste characteristic and its disposal. *International Journal of Ecological Science and Environmental Engineering*, 1(2), 49-61.
- World Energy Council Report. (2013). World energy resources: Waste to energy. Retrieved from https://www.worldenergy.org/wpcontent/uploads/2013/10/WER_2013_7b_Waste_to_Energy.pdf.
- Young, G. (2010). *Municipal solid waste to energy conversion processes, economic, technical, and renewable comparisons*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Views and opinions expressed in this article are the views and opinions of the author(s), International Journal of Sustainable Energy and Environmental Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.